

Revealing More Details: Image Super-Resolution for Real-World Applications

Aakerberg, Andreas

DOI (link to publication from Publisher):
[10.54337/aau561799641](https://doi.org/10.54337/aau561799641)

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Aakerberg, A. (2023). *Revealing More Details: Image Super-Resolution for Real-World Applications*. Aalborg Universitetsforlag. <https://doi.org/10.54337/aau561799641>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

**REVEALING MORE DETAILS:
IMAGE SUPER-RESOLUTION
FOR REAL-WORLD APPLICATIONS**

**BY
ANDREAS AAKERBERG**

DISSERTATION SUBMITTED 2023



AALBORG UNIVERSITY
DENMARK

Revealing More Details: Image Super-Resolution for Real-World Applications

Ph.D. Dissertation
Andreas Aakerberg

Dissertation submitted June 14, 2023

Dissertation submitted: June 14, 2023

PhD supervisor: Prof. Kamal Nasrollahi
Aalborg University and Milestone Systems

PhD committee: Associate Professor Jesper Rindom Jensen (chair)
Aalborg University, Denmark

Professor Alexandros Iosifidis
Aarhus University, Denmark

Associate Professor Tomer Michaeli
Technion – Israel Institute of Technology, Israel

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture, Design
and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-690-4

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

© Copyright: Andreas Aakerberg

Printed in Denmark by Stibo Complete, 2023

Curriculum Vitae

Andreas Aakerberg



Andreas Aakerberg commenced his professional journey as an automotive technician but soon realized his dissatisfaction with only being able to measure the in and output signals of engine management systems and communication interfaces without fully comprehending the inner workings. This led him to Aalborg University, Denmark where he received a BSc degree in Computer Engineering on the topic of Networks and Distributed Systems in 2015 and a MSc degree in Vision, Graphics and Interactive Systems in 2017. During his studies, he was employed as a student software developer at Intel Mobile Communications where he gained expertise in device verification and test automation. Additionally, he collaborated with HSA Systems, Aalborg, where he completed two computer vision projects during the final year of his master's studies, which eventually led to his employment as a computer vision and deep-learning specialist for three years until the initiation of his PhD study. The PhD was done at the Visual Analysis and Perception (VAP) Laboratory at Aalborg University, and during his PhD studies, Andreas collaborated with the Research Department at Milestone Systems A/S, Brøndby, Denmark, and the Image and Visual Representation Lab (IVRL) at the École polytechnique fédérale de Lausanne (EPFL), Switzerland. The collaboration involved a mix of online collaboration and physical visits. His main interests revolve around computer vision and machine learning, especially low-level vision problems such as image processing and computational photography, and their application to real-world scenarios, which is also reflected in his PhD thesis focusing on enhancement of real-world low-quality images. During his PhD studies, he has been actively involved in supervising graduate projects within the domains of computer vision and image processing.

Curriculum Vitae

Abstract

Digital images have become an integral part of our daily lives, but their quality may be compromised due to technical limitations. This PhD thesis, conducted from 2020 to 2023, focuses on investigating novel methods to enhance the visibility and quality of such images through Super-Resolution (SR).

The goal of SR is to restore high-resolution details from low-resolution images. State-of-the-art deep-learning-based SR models utilize learned priors, which can lead to unwanted artifacts when the input image resides outside the training distribution. Consequently, there is a growing interest in developing methods that generalize to real-world images. In this PhD thesis, we continued this line of research and explored even more challenging scenarios.

Specifically, we investigated SR of face images from surveillance footage and proposed a method to handle the artifacts present in such images, which obtained improved performance. Considering the scarcity of paired real-world low-resolution and high-resolution training image pairs, we explored the use of semantic segmentation guidance, which yielded highly improved results. Additionally, we developed a method to estimate per-pixel degradations and adapt the SR process accordingly, addressing the challenge posed by images with spatially variant degradations. To facilitate objective evaluation, we assembled the first dataset comprising such images and ground truths. Our method showcased superior performance when evaluated on this demanding dataset. Furthermore, we investigated SR of images degraded by both low-light and low-resolution, for which we curated another comprehensive dataset. Our benchmarking on this demonstrated significant advantages of joint processing over sequential processing. Leveraging these insights, we developed a dedicated method based on Transformers, leading to further performance improvements. Moreover, we investigated the usefulness of SR for other computer vision tasks and found that combining SR and semantic segmentation substantially enhance segmentation performance.

In conclusion, this thesis contributes significantly to the field of image SR, especially for demanding real-world applications, by deepening our understanding of challenges, proposing novel solutions and benchmark datasets, and enhancing performance.

Abstract

Resumé

Digitale billeder er blevet en integreret del af vores daglige liv, men kvaliteten kan være forringet grundet tekniske begrænsninger. Denne PhD-afhandling, udført fra 2020 til 2023, fokuserer på at undersøge nye metoder til at forbedre kvaliteten af sådanne billeder vha. Super-Resolution (SR). Formålet med SR er at genskabe højtopløste detaljer fra lavtopløselige billeder. Moderne deep-learning-baserede SR-modeller anvender indlærte priorer, der medføre uønskede artefakter, når indgangsbilledet ligger uden for træningsdistributionen. Der er derfor en stigende interesse i at udvikle metoder, der kan generalisere til virkelige billeder. I denne PhD-afhandling fortsatte vi denne forskningslinje og udforskede endnu mere udfordrende scenarier. Konkret undersøgte vi SR af ansigtsbilleder fra overvågningsoptagelser og foreslog en metode til at håndtere artefakterne i sådanne billeder, hvilket resulterede i forbedret præstation. Da der er få parvise lavtopløsnings- og højtopløsningsbilleder fra den virkelige verden til træning, undersøgte vi brugen af semantisk segmentering til guiding, hvilket gav markant forbedrede resultater. Derudover udviklede vi en metode til at estimere forringelser på pixelniveau og tilpasse SR-processen derefter, hvilket adresserer udfordringen med billeder med rumligt varierende forringelser. For at muliggøre objektiv evaluering har vi samlet det første datasæt bestående af sådanne billeder og deres referencer. Vores metode viste overlegen præstation ved evaluering med dette krævende datasæt. Derudover undersøgte vi SR af billeder, der er forringet af både svagt lys og lav opløsning, til hvilket vi sammensatte et omfattende datasæt. Vores benchmarking heraf viste betydelige fordele ved fælles behandling frem for sekventiel behandling. Ved at udnytte denne indsigt udviklede vi en dedikeret metode baseret på Transformers, hvilket førte til yderligere præstationsforbedringer. Derudover undersøgte vi anvendeligheden af SR til andre computer vision-opgaver og fandt, at kombinationen af SR og semantisk segmentering i høj grad forbedrer segmenteringspræstationen. Slutteligt, bidrager denne afhandling væsentligt til forskningsfeltet for SR, især for krævende virkelige anvendelser, ved at uddybe vores forståelse af udfordringer, foreslå nye løsninger og benchmark-datasæt og forbedre præstationen.

Resumé

Contents

Curriculum Vitae	iii
Abstract	v
Resumé	vii
List of Abbreviations	xvii
Thesis Details	xix
Preface	xxi
I Overview of the Work	1
1 Introduction	3
1 Thesis Structure	6
2 Introduction to Digital Image Acquisition	7
2.1 The Lens	8
2.2 The Digital Image Sensor	9
2.3 The Image Signal Processor	11
3 Introduction to Super-Resolution	12
3.1 State-of-the-Art	13
3.2 Current Challenges in Super-resolution Research	16
3.3 Image Quality Assessment	16
3.4 Ethics in Relation to Super-Resolution	18
References	20
2 Real-World Super-Resolution	29
1 Introduction	29
2 State-of-the-Art	31
2.1 Real-World Super-Resolution Methods	31
2.2 Datasets	33

Contents

3	Scientific Contributions	33
	References	40
3	Joint Low-Light Image Enhancement and Super-Resolution	47
1	Introduction	47
2	State-of-the-Art	49
3	Scientific Contributions	50
	References	54
4	Improving Downstream Vision Tasks With Super-Resolution	59
1	Introduction	59
2	State-of-the-Art	60
3	Scientific Contributions	61
	References	64
5	Conclusion	69
II	Papers	73
A	Real-World Super-Resolution of Face-Images From Surveillance Cam- eras	75
1	Introduction	77
2	Related Work	79
3	The Proposed Framework	81
3.1	Novel Image Degradation	81
3.2	Blur Kernel Estimation	82
3.3	Noise Estimation	82
3.4	Degradation with Compression artifacts	83
3.5	Backbone Model	83
3.6	Datasets	84
3.7	Evaluation Metrics	85
4	Experiments and Results	86
4.1	Comparison with State-of-the-Art	86
4.2	Ablation Study	90
4.3	Failure Cases	90
5	Conclusion	91
6	Acknowledgments	92
	References	92
B	Real-World Thermal Image Super-Resolution	99
1	Introduction	101
2	Related Work	102
2.1	RGB Image Super-Resolution	102

2.2	Thermal Image Super-Resolution	104
3	Dataset	105
4	Thermal RealSR	105
4.1	Realistic Degradation using KernelGAN and Noise In- jection	106
4.2	Super-Resolution Model	107
5	Experiments and Results	108
5.1	Evaluation Metrics	108
5.2	Comparison with the State of the Art	109
6	Conclusion	110
	References	112
C	Semantic Segmentation Guided Real-World Super-Resolution	115
1	Introduction	117
2	Related Work	119
2.1	Single image super-resolution	119
2.2	Guided super-resolution	120
2.3	Semantic segmentation	121
3	The Proposed Method	122
3.1	Guiding with semantic segmentation	123
3.2	Domain adaptation	123
3.3	Backbone networks	124
4	Implementation details	125
5	Experiments and results	126
5.1	Datasets	126
5.2	Quantitative Evaluation metrics	126
5.3	Qualitative results	128
5.4	Quantitative results	128
5.5	Ablation study	130
6	Conclusion	130
	References	131
D	PDA-RWSR: Pixel-Wise Degradation Adaptive Real-World Super-Resolution	135
1	Introduction	137
2	Related Work	139
2.1	Single Image Super-resolution	139
2.2	Classic Blind Super-Resolution	139
2.3	Real-World Super-resolution	139
3	Method	140
3.1	Spatially Variant Degradation Model	141
3.2	Pixel-Wise Degradation Estimation	142
3.3	Pixel-Wise Feature Modulation	143

4	Spatially Variant Super-Resolution (SVSR) Dataset	144
4.1	Data Collection	144
4.2	Data Pre-processing	145
4.3	Data Analysis	145
5	Experiments and Analysis	147
5.1	Experimental Setup	147
5.2	Comparison with State-of-The-Art Methods	148
5.3	Ablation Studies	150
6	Conclusion	151
	References	151
E	RELLISUR: A Real Low-Light Image Super-Resolution Dataset	157
1	Introduction	159
2	Related Work	162
2.1	Real-world super-resolution datasets	162
2.2	Low/normal-light datasets	163
3	RELLISUR Dataset	163
3.1	Collection method	164
3.2	Preprocessing	165
3.3	Analysis of dataset content	166
4	Experiments	168
4.1	Baseline methods for end-to-end learning	168
4.2	Implementation details	169
4.3	Results	169
5	Conclusion	171
	References	172
F	RELIEF: Joint Low-Light Image Enhancement and Super-Resolution with Transformers	179
1	Introduction	181
2	Background	184
2.1	Low-light image enhancement	184
2.2	Image Super-resolution	184
2.3	Vision Transformer	185
3	Method	185
3.1	Overall pipeline	185
3.2	ECSSwin Self-Attention Transformer Block	186
3.3	Locally-enhanced Feed-Forward Network	187
3.4	Locally-enhanced positional encoding	188
4	Experiments and Analysis	188
4.1	Datasets	188
4.2	Evaluation metrics	189
4.3	Implementation details	189

Contents

4.4	Comparison with existing methods	190
4.5	Results	191
4.6	Ablation Studies	193
5	Conclusion	195
	References	195
G	Single-Loss Multi-Task Learning for Improving Semantic Segmentation Using Super-Resolution	203
1	Introduction	205
2	Related Work	206
3	The Proposed Framework	207
4	Experiments and Results	208
4.1	Datasets	208
4.2	Implementation Details	208
4.3	Results	209
4.4	Ablation Study	210
5	Conclusion	212
6	Acknowledgements	212
	References	212
III	Patent applications	215
I		217
II		221

Contents

List of Abbreviations

AI	Artificial Intelligence.
APE	Absolute Positional Encoding.
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator.
CNN	Convolutional Neural Network.
CPE	Conditional Positional Encoding.
CSWin	Cross-Shaped Window.
DFEB	Degradation Feature Extraction Block.
DISTS	Deep Image Structure and Texture Similarity.
DNN	Deep Neural Network.
DOF	Depth-of-Field.
DSLR	Digital single-lens reflex.
DSN	Degredation Simulation Network.
ECSWin	Enhanced Cross-Shaped Window.
EV	Exposure value.
FFHQ	Flickr-Faces-HQ Dataset.
FR-IQA	Full-Reference Image Quality Assessment.
GAN	Generative Adversarial Network.
GELU	Gaussian Error Linear Unit.
GMACs	Giga Multiply-Accumulates per Second.
GT	Ground-Truth.
HR	High-Resolution.
i.i.d	independent and identically distributed.

List of Abbreviations

IQA	Image Quality Assessment.
ISP	Image Signal Processor.
LeFF	Locally-enhanced Feed-Forward.
LePE	Locally-Enhanced Positional Encoding.
LL	Low-Light.
LLE	Low-Light Enhancement.
LLLR	Low-Light Low-Resolution.
LPIS	Learned Perceptual Image Patch Similarity.
LR	Low-Resolution.
MAE	Mean Average Error.
mIoU	mean Intersection over Union.
MISR	Multiple Image Super-Resolution.
MLP	Multi-Layer Perceptron.
MOR	Mean Opinion Rank.
MOS	Mean Opinion Score.
MS-SSIM	Multi Scale Structural Similarity index.
MSE	Mean Squared Error.
MT-SSSR	Multi-Task Semantic Segmentation and Super-Resolution.
NIMA	Neural Image Assessment.
NIQE	Natural Image Quality Evaluator.
NIR	Near-Infrared.
NL	Normal-Light.
NLHR	Normal-Light High-Resolution.
NLPD	Normalized Laplacian Pyramid Distance.
NR-IQA	No-Reference Image Quality Assessment.
OHEM	Online Hard Example Mining.
PDA-RWSR	Pixel-Wise Degradation Adaptive Real-World Super-Resolution.
PIQE	Perception based Image Quality Evaluator.
PSF	Point Spread Function.
PSNR	Peak Signal-to-Noise Ratio.
RELIEF	Resolution and Light Enhancement Transformer.
RELLISUR	Real Low-Light Image Super-Resolution.
RGB	Red Green and Blue.
RMI	Region Mutual Information.

List of Abbreviations

RPE	Relative Positional Encoding.
RRDB	Residual-in-Residual Dense Block.
RTB	Restormer Transformer Block.
RWSR	Real-World Super-Resolution.
SFT	Spatial Feature Transformation.
SFTB	Spatial Feature Transformation Block.
SISR	Single-Image Super-Resolution.
SNR	Signal-to-Noise Ratio.
SoTA	State-of-The-Art.
SR	Super-Resolution.
SS	Semantic Segmentation.
SSG-RWSR	Semantic Segmentation Guided Real-World Super-Resolution.
SSIM	Structural Similarity index.
SVSR	Spatially Variant Super-Resolution.

List of Abbreviations

Thesis Details

Thesis Title: Revealing More Details: Image Super-Resolution for Real-World Applications
PhD Student: Andreas Aakerberg
Supervisor: Prof. Kamal Nasrollahi, Aalborg University

The thesis consists of the following publications:

- [A] **Andreas Aakerberg**, Kamal Nasrollahi, and Thomas B Moeslund, “Real-world super-resolution of face-images from surveillance cameras”. In: *IET Image Processing*, Volume 16, Issue 2, pp. 442-452, 2022.
- [B] Moaaz Allahham, **Andreas Aakerberg**, Kamal Nasrollahi, and Thomas B. Moeslund, “Real-World Thermal Image Super-Resolution”. In: *Advances in Visual Computing: 16th International Symposium (ISVC)*, Proceedings, Part I, pp. 3-14, 2021.
- [C] **Andreas Aakerberg**, Anders S. Johansen, Kamal Nasrollahi, and Thomas B Moeslund, “Semantic segmentation guided real-world super-resolution”. In: *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops*, pp. 449-458, 2022.
- [D] **Andreas Aakerberg**, Majed El Helou, Kamal Nasrollahi, Sabine Süssstrunk, and Thomas B Moeslund, “PDA-RWSR: Pixel-Wise Degradation Adaptive Real-World Super-Resolution”. Under Review: *International Conference on Computer Vision (ICCV)*, 2023.
- [E] **Andreas Aakerberg**, Kamal Nasrollahi, and Thomas B Moeslund, “REL-LISUR: A Real Low-Light Image Super-Resolution Dataset”. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS)*, 2021.
- [F] **Andreas Aakerberg**, Kamal Nasrollahi, and Thomas B Moeslund, “RELIEF: Joint Low-Light Image Enhancement and Super-Resolution with

Transformers”. In: *Scandinavian Conference on Image Analysis (SCIA)*, Lecture Notes in Computer Science, vol. 13885. Springer, pp. 157-173, 2023.

- [G] **Andreas Aakerberg**, Anders S. Johansen, Kamal Nasrollahi, and Thomas B Moeslund, “Single-loss multi-task learning for improving semantic segmentation using super-resolution”. In: *Computer Analysis of Images and Patterns - 19th International Conference (CAIP)*, Lecture Notes in Computer Science, vol. 13053, pp. 403-411, 2021.

In addition to paper [Paper D] and paper [Paper F], the PhD student has been involved in formulating two patent applications which has been filed by Canon Europe Limited on behalf of Milestone Systems:

- [I] On the basis of: [Paper F], Inventors: **Andreas Aakerberg**, Kamal Nasrollahi, and Thomas B Moeslund, Patent application number: 2205153.6, Filing date: April 7th 2022.
- [II] On the basis of: [Paper D], Inventors: **Andreas Aakerberg**, Kamal Nasrollahi, and Thomas B Moeslund, Patent application number: 2303244.4, Filing date: March 6th 2023.

Furthermore, the PhD student has authored the following publications which are not part of the thesis.

- **[Best paper award] Andreas Aakerberg**, Kamal Nasrollahi, Christoffer Bøgelund Rasmussen, and Thomas B Moeslund, “Improving a deep learning based RGB-D object recognition model by ensemble learning”. In: *Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1-6, 2017.
- **Andreas Aakerberg**, Kamal Nasrollahi, and Thomas Heder, “Depth value pre-processing for accurate transfer learning based RGB-D object recognition”. In: *International Joint Conference on Computational Intelligence*, pp. 121-128, 2017.
- **Andreas Aakerberg**, Kamal Nasrollahi, Christoffer Bøgelund Rasmussen, and Thomas B Moeslund, “Complementing SRCNN by Transformed Self-Exemplars”. In: *Video Analytics. Face and Facial Expression Recognition and Audience Measurement: Third International Workshop*, pp. 127-136, 2017.

This thesis has been submitted for assessment in partial fulfillment of the PhD degree. The thesis is based on the submitted or published scientific papers which are listed above. Parts of the papers are used directly or indirectly in the extended summary of the thesis. As part of the assessment, co-author statements have been made available to the assessment committee and are also available at the Faculty.

Preface

This thesis is submitted as a collection of papers in partial fulfillment of a PhD study at the Section of Media Technology, Aalborg University, Denmark. This thesis covers research in three main parts: *Real-World Super-Resolution*, *Joint Low-Light Image Enhancement and Super-Resolution*, and *Improving Downstream Vision Tasks With Super-Resolution*. This work has been carried out from 2020-2023, mainly in the Visual Analysis and Perception (VAP) Lab at Aalborg University, and with collaborations with the Research Department at Milestone Systems A/S, Brøndby, Denmark, and the Image and Visual Representation Lab (IVRL) at the École polytechnique fédérale de Lausanne (EPFL), Switzerland. The PhD fellowship was funded by Danmarks Frie Forskningsfond under grant number 8022-00360B and with support from Milestone Systems A/S, through the Milestone Research Programme at Aalborg University.

The PhD period was a challenging time for me on a personal level, as I went through a divorce and experienced the loss of my brother. As a result, completing this thesis in expected time would not have been possible without the support of several individuals around me. Therefore, I am extremely grateful for everyone who has been involved and made a difference. A special thanks go to Thomas B. Moeslund for his efforts in providing a wonderful lab environment. Also thanks to my supervisor, prof. Kamal Nasrollahi, for consistently believing in me and providing me with lots of interesting opportunities to help me develop my research and leadership skills. Thanks to Sabine Süssstrunk for hosting me in the IVRL lab, and Majed El Helou for excellent supervision and interesting talks throughout our collaboration. I would also like to express my gratitude to my colleagues at the Visual Analysis and Perception Laboratory, specifically, Anders S. Johansen for our research collaboration that led us to WACV in Waikoloa, where we enjoyed Mai Tai's at the beach post-conference, Neelu Madan for providing a listening ear and uplifting me during difficult times, and to Malte Pedersen, I admire your positive attitude and really value our friendship. Additionally, I extend my heartfelt thanks to my parents, and the mother of my children for always being supportive, particularly when I needed to work overtime or travel. Thank you.

Andreas Aakerberg

Aalborg University, June 14, 2023

Preface

Dedicated to the loving memory of my dear brother Thomas Aakerberg.

Preface

Part I

Overview of the Work

Chapter 1

Introduction

With the widespread adoption of camera-enabled smartphones, digital images, and videos have become integral to our daily lives [17]. Moreover, digital images have gained significant importance in critical applications in today's society such as automation, manufacturing, healthcare, public safety, and criminal investigations, to name a few examples. Although capturing visually appealing images of everyday moments using smartphones is desirable, high-quality image data is crucial for reliable and accurate operation of automated analysis and decision-making applications. Consequently, if the image quality is insufficient, these might make wrong predictions, or fail completely [19, 35, 45]. Unfortunately, technical limitations and environmental factors can prevent digital cameras from consistently producing images of satisfactory quality to support these functions.

As an example, this is oftentimes the case for footage recorded by outdoor surveillance cameras which typically operate in harsh and uncontrollable environments. For instance, in a recent incident, where an individual went missing on February 2022 in Aalborg, Denmark, surveillance footage showing the individual entering a dark car prior to their disappearance, was used as part of the investigation [62]. However, as seen in Figure 1.1, the make and model of the dark car could not be determined due to the low image quality of the recorded footage.

Low image quality, or specifically in the surveillance scenario, low visibility of objects of interest in the image, can occur due to several factors. These factors include, but are not limited to blur, noise, under-exposure, and low resolution, and occur both due to hardware limitations and unfavorable environmental conditions [11]. While these problems can be mitigated by upgrading the camera system, and adding artificial light sources, these solutions are often expensive and impractical. Furthermore, there is always the possibility of an object of interest being positioned too far away from the camera to become



Fig. 1.1: An image captured by a surveillance camera, which fails to acquire the car in the upper right corner with sufficient detail to identify its make and model. Photo made public by Aalborg Politi.

identifiable. This problem is further complicated by the wide field-of-view lenses, typically used with surveillance cameras, which cover large areas but provide fewer pixels for resolving individual objects. Moreover, in Denmark alone, there already exist approximately 1.5 million active surveillance cameras [6], and a large part of these are incapable in terms of providing clear and detailed recordings [93]. Consequently, digital images are often rejected as evidence in a court of law due to low image quality [76].

However, it is important to note that the problem of low image quality extends beyond surveillance cameras, as no matter how high the resolution, there often exists a need to improve it. Other examples include medical imaging, such as MRI and CT scans [25], as well as satellite [68] and aerial imaging [78], where low image quality is frequently encountered, although a higher quality is often desired. Therefore, there is a growing need to enhance not only surveillance footage but also any type of low-quality, low-resolution images.

One potential solution to address the challenges of low image quality is software-based and relies on digital image processing. Digital image processing refers to a range of techniques and methods employed to manipulate digital images on a computer, aiming to improve the overall image quality or enhance specific properties or features [24]. In recent years, there has been a significant surge of interest in image processing, both in research and industry, partly driven by the desire for high image quality in compact consumer devices like smartphones. This arises from the inherent trade-off between large, high-quality camera systems and the desire for small form factors. Image processing techniques, also referred to as computational photography in this context, offer



Original



Enhanced

Fig. 1.2: $\times 4$ super-resolution of the dark car from Figure 1.1. While the overall difference is subtle, the outline of the head and taillights are now more clearly defined. Furthermore, the enhancement has revealed the position of the brand logo on the tailgate which altogether facilitates the identification of the car as a VW Golf Mk. VII. Original photo made public by Aalborg Politi.

potential solutions to overcome this trade-off [17]. Among the most promising techniques in this domain is image super-resolution which aims to increase the resolution and details of low-quality images [59]. However, as we will shed light on throughout PhD thesis, the Super-Resolution (SR) process is highly complex, due to the significant loss of information and the presence of artifacts in the Low-Resolution (LR) images. Consequently, the SR algorithms must be meticulously designed, to ensure optimal reconstruction quality in a given situation. However, despite notable progress made in the past decade, existing SR methods are still far from delivering the expected image quality improvement in real-world scenarios. An example of enhancement with a SR method applied to the dark car from Figure 1.1 can be seen in Figure 1.2. While the enhancements produced by the SR algorithm were enough to help identify the car in this case, the image still appears to be of poor quality. This is mainly attributed to the presence of various unknown factors that contribute to the degradation of the original image, thereby posing a significant challenge for the SR algorithm.

Consequently, the primary objective of this PhD thesis has been to explore approaches for improving the performance of current State-of-The-Art (SoTA) SR algorithms. Specifically, a particular emphasis has been directed towards improving their performance when applied to real-world images. This emphasis stems largely from the close collaboration with Milestone Systems, who have presented various practical challenges for consideration. Of special interest are the challenges presented through a collaborative effort involving Aalborg University, Milestone Systems, and different police departments in Denmark. Specifically, throughout the PhD period, we have been involved in supporting Sydøstjyllands Politi, Aalborg Politi, and one undisclosed de-

partment, in enhancing low-quality surveillance footage for ongoing forensic investigations. However, due to a non-disclosure agreement, no further descriptions of these cases are included in this thesis. However, it can be mentioned that the primary issue in all cases has been the insufficient visibility of the recorded footage, primarily resulting from low-resolution and unfavorable lighting conditions.

1 Thesis Structure

This PhD thesis is structured into three main parts:

- Part I: This first part provides an overview of the research conducted in this PhD thesis, divided into five chapters:
 - Chapter 1 gives a description of the main topic and scope of this PhD thesis along with an introduction to digital imaging acquisition and super-resolution, included to provide the reader with the necessary background for the subsequent sections.
 - Chapter 2 provides an introduction, an overview of the related work, and outlines the scientific contributions of this PhD thesis in the domain of *Real-World Super-Resolution*.
 - Chapter 3 provides an introduction, an overview of the related work, and outlines the scientific contributions of this PhD thesis in the domain of *Joint Low-Light Image Enhancement and Super-Resolution*.
 - Chapter 4 provides an introduction, an overview of the related work, and outlines the scientific contributions of this PhD thesis in the domain of *Improving Downstream Vision Tasks With Super-Resolution*.
 - Chapter 5 summarizes and concludes the key findings of this PhD thesis.

An overview of the papers covered in chapter 2-4, which have all been written during the PhD study, can be seen in Figure 1.3.

- Part II: This part comprises a collection of papers included in this PhD thesis, serving as the foundational work for this PhD thesis.
- Part III: This part encompasses patent applications that have been formulated on the basis of the methods developed during the PhD study.

The following section provides an introduction to image acquisition with digital cameras to offer a theoretical understanding of the challenges that must be addressed to enhance the quality of real-world images.

2. Introduction to Digital Image Acquisition

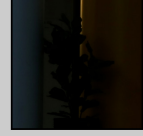
Chapter II: Real-World Super-Resolution

- [A] Real-World Super-Resolution of Face-Images From Surveillance Cameras
- [B] Real-World Thermal Image Super-Resolution (A)
- [C] Semantic Segmentation Guided Real-World Super-Resolution (G)
- [D] PDA-RWSR: Pixel-Wise Degradation Adaptive Real-World Super-Resolution



Chapter III: Joint Low-Light Image Enhancement and Super-Resolution

- [E] RELLISUR: A Real Low-Light Image Super-Resolution Dataset
- [F] RELIEF: Joint Low-Light Image Enhancement and Super-Resolution with Transformers (E)



Chapter IV: Improving Downstream Vision Tasks With Super-Resolution

- [G] Single-Loss Multi-Task Learning for Improving Semantic Segmentation Using Super-Resolution



Fig. 1.3: An overview of the publications included in this PhD thesis, wherein each publication is identified by capitalized letters enclosed in brackets. The arrangement of these publications is determined by their relevance to the three main research topics addressed in this thesis. Additionally, direct influence by preceding works is indicated by the presence of the respective publication id in parenthesis.

2 Introduction to Digital Image Acquisition

Image acquisition using digital cameras involves the process of capturing and converting analog visual information into a digital format. In simple terms, a digital image is produced by reflected light that passes through a lens to become collected by an image sensor. However, during this process, different types of degradations occur which limits the image quality. Besides environmental conditions, the main components contributing to this are the lens, digital sensor, and the Image Signal Processor (ISP). An illustration of the digital imaging pipeline can be seen in Figure 1.4. In the following section, we will present an overview of how these camera components impact the quality of the final image. However, it is important to note that the scope of this topic is extensive, with the potential for an entire PhD thesis dedicated solely to exploring the coating of the lens elements, for instance. Hence, our coverage will focus on the most crucial factors, beginning with the properties of the lens. For a more comprehensive explanation, references such as [58, 63] can be consulted.

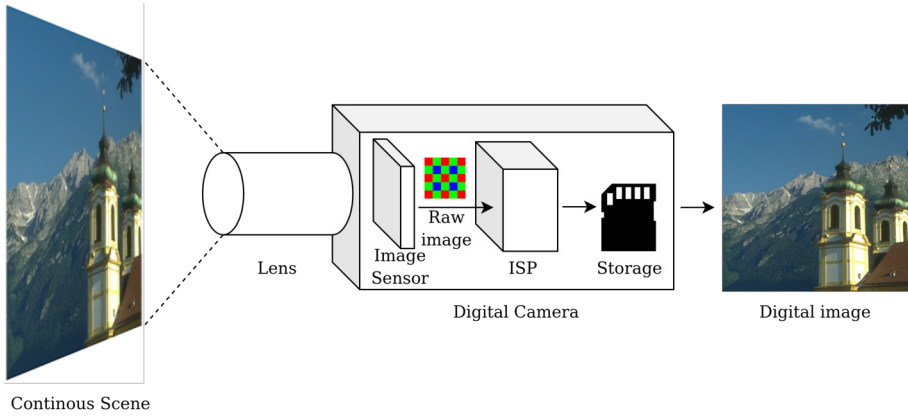


Fig. 1.4: An illustration of the digital imaging pipeline. Image from the BSD100 dataset [52].

2.1 The Lens

The purpose of the lens is to precisely focus incoming light onto the digital image sensor to produce a sharp and accurate image. However, the variations in the quality of the optical elements and the lens manufacturing process can introduce deviations that negatively affect image sharpness. One way to characterize these deviations is through the Point Spread Function (PSF), which describes how a lens spreads out an incoming point of light, into a pattern or spot in the captured image. While an ideal lens would project a point of light as a well-defined and sharp spot on the image sensor, physical lenses exhibit non-ideal PSFs that cause some blurring or spreading of the light, leading to a decrease in image sharpness. In image processing, the PSF is often represented as the blur kernel, typically a matrix, describing how the sharpness of the image was reduced and how neighboring pixels influence each other. By knowing the parameters of the blur kernel, it becomes possible to reverse the blurring effect and recover the original image.

Chromatic aberration is a specific aspect related to the PSF behavior. When the PSF varies for different wavelengths of light, it leads to chromatic aberration, characterized by colored fringes or halos around the object edges in the image. In addition to the PSF, another factor influencing image sharpness is the focusing depth of the lens. This is because the lens's ability to achieve precise focus is limited by its Depth-of-Field (DOF), which refers to the range of distances over which objects appear acceptably sharp in the captured image. Due to the intrinsic nature of optics, achieving simultaneous focus on all objects at different distances is not possible. As such, objects located outside the DOF, whether they are closer or farther from the focal plane, will appear progressively more out of focus. When capturing images, the lens can

be focused to prioritize sharpness on specific objects within the scene while accepting some degree of blur for objects outside the DOF.

The DOF can be manipulated by altering the lens aperture. Decreasing the aperture (higher f-number) expands the DOF, allowing a greater range of distances to be in acceptable focus. However, this also limits the amount of incoming light, potentially necessitating longer exposure times or higher ISO settings, which can introduce motion blur or noise, respectively. Moreover, smaller apertures increase the risk of diffraction, which negatively impacts image sharpness. The phenomenon of diffraction is also related to the PSF and occurs when the light passing through the lens and aperture interacts with the edges of the opening and causes interference. Lastly, vignetting is another phenomenon that affects image quality by a gradual decrease in brightness or illumination towards the edges of the image frame.

2.2 The Digital Image Sensor

The incoming light focused by the lens is projected to the digital image sensor which consists of a grid of photodiodes known as pixels. Two important properties influencing the quality and details of the captured image are the density of pixels and their physical size.

The pixel density determines the spatial resolution of the image and is measured in pixels per unit area. A higher spatial resolution enables capturing images with finer details, while a lower resolution leads to images with fewer details and potential artifacts. From a signal processing perspective, artifacts can arise from aliasing due to low spatial sampling frequency, particularly if the sampling rate is not at least twice the frequency of the content [61]. A visual comparison of a scene captured with different sensor resolutions is shown in Figure 1.5.

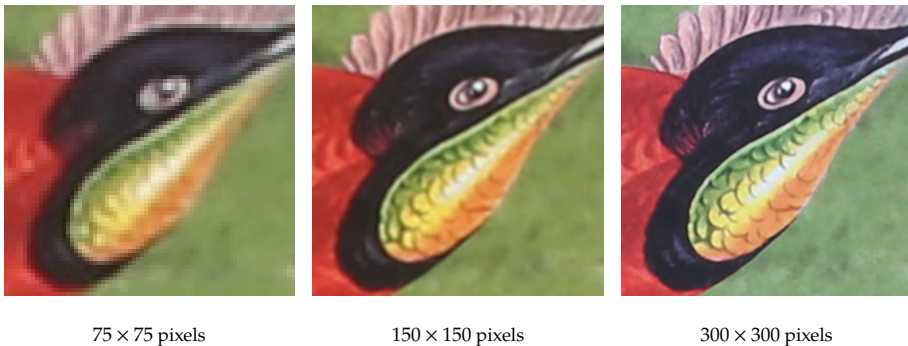


Fig. 1.5: Visual examples illustrating the improved sharpness and level of details with increased spatial sampling resolution. Images from [3], Paper E.

One solution to increase the resolution, or spatial sampling frequency, is to increase the pixel density by packing in more pixels per unit. However, as the pixel size decreases, so does their ability to collect incoming light, or specifically the photon particles making up the light, resulting in increased noise in the image. The noise increases because the Signal-to-Noise Ratio (SNR) decreases, meaning that the noise inherently part of the imaging process becomes a more significant component relative to the actual signal. The noise originates from multiple different sources. One of the most common sources is sensor noise that originates from the sensor chip itself either as random fluctuations in pixel values due to thermal energy in the sensor or fixed pattern noise due to consistent pixel-level variations as a result of manufacturing imperfections. Readout noise is another noise source that occurs during the reading process of the electronic signals in the sensor, caused by components such as signal amplifiers and analog/digital converters. Lastly, shot noise is a permanent unavoidable noise source due to the discrete nature of light, which results in random fluctuation of the number of detected photons.

The exposure time controls the period in which light is collected by the image sensor, leading to an image that is either underexposed, overexposed, or correctly exposed depending on the lighting conditions, lens aperture, and sensitivity of the image sensor. However, if the exposure time is too long and objects in the scene are moving rapidly relative to the exposure time and camera distance, it can result in motion blur. One way to mitigate this is to increase the sensor's sensitivity by increasing the digital gain, or ISO setting while reducing the exposure time. However, this not only amplifies the signal but also the unwanted noise.

Another factor influencing the image quality is the dynamic range, which represents the brightest and darkest levels that can be captured in the same image without clipping or saturation. Specifically, the dynamic range is defined as the ratio between the maximum electrical current that each photodiode can hold and its noise. Consequently, dynamic range is directly correlated to pixel size, which enables Digital single-lens reflex (DSLR) cameras with larger sensors to encode up to 14-bits, while typical smartphone cameras are limited to 10-bit encoding. If the dynamic range of the scene exceeds the camera's range,



Fig. 1.6: An example of a scene that exceeds the camera's dynamic range, resulting in parts of the image appearing too dark.

parts of the image will appear too bright or dark depending on the exposure settings, as illustrated in Figure 1.6.

Lastly, although some image sensors have dedicated photodiodes for each Red Green and Blue (RGB) color channel, the majority of cameras utilize color sub-sampling using a color filter array positioned in front of the photodiodes. The filter is typically arranged in an RGGB Bayer pattern with twice as many green filters as red or blue, resulting in lower color accuracy. The following section will elaborate on how the ISP reconstructs a 3-channel RGB image from the single-channel raw image.

2.3 The Image Signal Processor

The ISP receives the raw Bayer image data from the image sensor and executes a series of sequential processing steps to transform and enhance the quality of the image. One of the initial steps involves Bayer demosaicing, an algorithm that interpolates the missing color values within the Bayer pattern to reconstruct a complete 3-channel RGB image. However, due to the sparse nature of the Bayer pattern, two out of three colors at each pixel in the resulting output image are interpolated rather than directly sampled from the original scene.

Following demosaicing, white balancing is performed to adjust the color temperature of the image such that it appears natural to the human eye without any color cast. Subsequently, a number of vendor-specific photo-finishing steps such as color manipulation, contrast enhancement, noise reduction, and sharpening are applied to further refine the image according to a specific style.

Although the following operations prepare the image for practical use, they involve discarding a significant amount of the originally captured information. Most ISPs primarily process the image using the high-bit-depth data for optimal fidelity, but eventually, the color channel dimension is reduced, typically to 8-bit. Additionally, to further reduce storage requirements, the image is often compressed using JPEG compression. Although these last steps only cause a subtle drop in image quality to the human eye, they introduce complications for subsequent digital image processing techniques and computer vision algorithms, due to the loss of information and introduction of compression artifacts. A summary of the different factors limiting the image quality and their placement in the digital imaging pipeline can be seen below:

- **Scene:** Blur caused by fast object motion or camera shake in combination with unsuitable exposure settings. Environmental factors such as low-light.
- **Lens:** Out-of-focus, lens distortions, and wide-field of view.
- **Image sensor:** Color sub-sampling, noise, spatial quantization, limited dynamic range, and sensitivity.

- **ISP:** Quantization, Bayer demosaicing errors, and compression artifacts

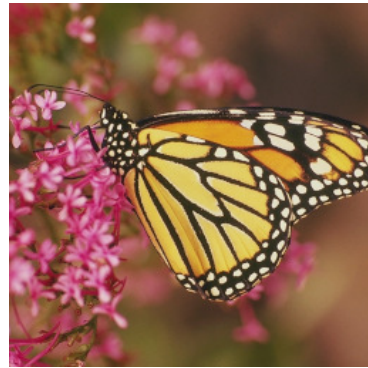
Recently, advancements in the computational photography research domain have been integrated into the modern camera pipeline to enhance image quality. These advancements include various image processing techniques such as denoising [72, 88, 92], joint demosaicing and denoising [22, 28], high-dynamic imaging [21, 34], deblurring [87, 89], and SR [20, 59, 82], among others. The following section serves as an introduction to the topic of SR, one of the most capable image-processing methods for addressing the aforementioned challenges. Additionally, the following section also provides the necessary background information for the subsequent chapters of this thesis.

3 Introduction to Super-Resolution

The concept of super-resolution has been popularized through science fiction movies like Blade Runner and TV shows such as CSI Miami, where computer-based tools to enhance the quality and details of images are demonstrated. However, while it is a simple task to increase the spatial resolution of an image by upsampling the image with interpolation methods, such as nearest-neighbor interpolation, no new information is added to the image in this process. This is also the case for more advanced methods, such as bicubic interpolation [36]. A visual example is given in Figure 1.7, where it can be seen that while bicubic interpolation correctly upsamples the smooth low-frequency areas, the image lacks sharpness, fine details, and texture. This is the fundamental difference from SR which simultaneously aims to increase both the resolution and the amount of high-frequency information in the image [1].



×4 Bicubic interpolation



High-Resolution (HR) Ground-truth

Fig. 1.7: Two images of the same spatial resolution (256×256 pixels), but with very different amounts of high-frequency details. Image from the Set14 dataset [86].

3. Introduction to Super-Resolution

In essence, SR aims to obtain this by reversing the digital image acquisition pipeline explained in Section 2 and hereby reconstruct the scene in HR based on one or more LR observations. Thus, the SR research field divides into Single-Image Super-Resolution (SISR) and Multiple Image Super-Resolution (MISR) SR based methods, where SISR methods use a single LR image to reconstruct the HR image, and MISR uses multiple. Since attempting to reverse the entire imaging pipeline is a much more elaborate problem than addressing a single factor at a time, such as denoising [72] or deblurring [89], SR is a highly challenging and complex problem.

The SISR process, which is the primary method used today, can be formulated as an inverse problem that requires constructing a forward model, commonly in the form [44]:

$$y = (x \circledast k) \downarrow_s + n \quad (1.1)$$

which involve convolution with a blur kernel k on the unknown HR image x , followed by downsampling with scale factor s , and lastly degradation by additive noise n resulting in the observed LR image y . However, inverting the forward model and estimating its parameters to obtain the HR image x based on y is not a trivial task, as the inverse problem is highly ill-posed. That is, without additional constraints, there is no unique solution to the problem, meaning that multiple plausible HR images could correspond to the same LR image. Furthermore, for an optimal reconstruction, the parameters in the forward model should ideally encompass all parameters in the imaging pipeline, which is challenging in practice.

3.1 State-of-the-Art

While the importance of super-resolution has attracted many researchers over the last decades, the problem remains unsolved due to its challenging nature. One of the initial attempts at SR, illustrated in Figure 1.8, dates back to 1984 [74] and involves alignment and fusion of multiple low-resolution images with sub-pixel motion to produce a high-resolution image. Since then, many different techniques have been proposed [59, 81], both in the frequency [75] and spatial domain and by using signal processing [38, 66], statistical [7, 73], and machine learning techniques [20, 37]. Some approaches include example-based methods which utilize a database of HR image patches to find similar patches in the low-resolution input image, for the purpose of reconstructing the missing details [84]. Self-similarity-based methods exploit the self-similarity inherently present in natural images, to find patches with similar content but at different scale levels [23, 29].

Nowadays, most researchers approach the SR problem using deep-learning methods in the spatial domain due to the high modeling capacity and flexibility of such models. In 2014, the seminal work of SRCNN [20] presented a groundbreaking approach in this regard by designing and training a Convolutional Neural Network (CNN) to learn the mapping between LR and HR image patches. The architecture was simple with only three main layers. The first layer performed shallow feature extraction, followed by a non-linear mapping layer, and lastly, a reconstruction layer, while the actual spatial upsampling was performed as a pre-processing step using bicubic interpolation. The network was trained in a supervised manner using pairs of LR and HR images and optimized to minimize the Mean Squared Error (MSE) between the networks prediction and the ground truth HR image. Due to its superior performance over existing hand-crafted methods, many follow-up works have used a similar approach. An illustration of a CNN based SR architecture commonly used today can be seen in Figure 1.9.

Here, the main difference from SRCNN is that following the initial shallow feature extraction used to convert the 3-channel input image to feature maps, multiple deep features or mapping blocks are often used. Furthermore, nowadays the upsampling is typically done as the last step together with the HR output reconstruction. However, some researchers also investigated progres-

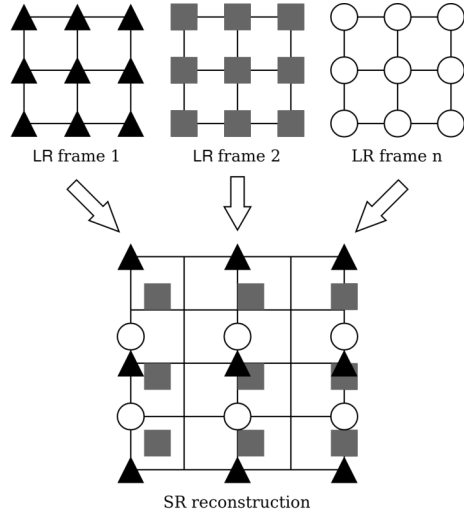


Fig. 1.8: The initial idea for SR was based on leveraging complementary information from multiple LR frames with sub-pixel motion by alignment and fusion.

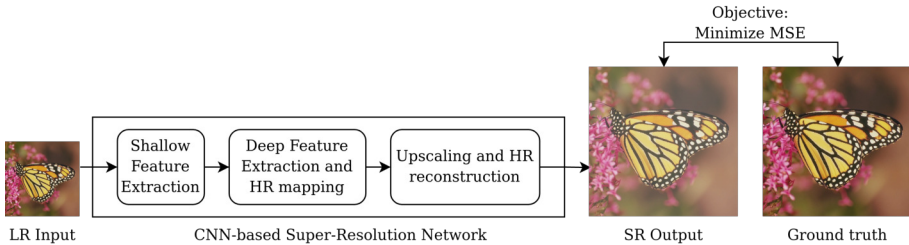


Fig. 1.9: The typical components of a CNN-based SR model and its learning objective. Image from the Set14 dataset [86].

3. Introduction to Super-Resolution

sive upsampling [5, 80] to reduce the learning difficulty in the individual part, and iterative up-and-down upsampling [46] to benefit from also learning the mapping from HR to LR.

Since the proposal of SRCNN [20], the reconstruction performance has been improved by more advanced network architectures. Key works include EDSR [43] and SRResNet [40], which is based on deep residual CNNs, and RCAN [91], SAN [16], and HAN [60] which incorporates channel and spatial attention mechanisms. More recent approaches utilize Transformers [12, 42] due to their improved long-range modeling capabilities compared to CNNs. These networks are typically optimized with L1 (Mean Average Error (MAE)) or L2 (MSE) loss, resulting in low distortion, *i.e.* the difference between the reconstructed image and the target according to a given distance metric, but not necessarily visually pleasing reconstructions according to human judgment. Later works try to overcome this problem by the use of Generative Adversarial Networks (GANs), and optimization of the models by a combination of MSE, GAN, and perceptual losses [40, 79].



(a) PSNR \uparrow : 25.72, SSIM \uparrow : 0.7377, LPIPS \downarrow : 0.368

(b) PSNR \uparrow : 22.97, SSIM \uparrow : 0.6238, LPIPS \downarrow : 0.152

Fig. 1.10: High-resolution reconstructions produced by the same SR model [79], but optimized with different loss functions. \uparrow and \downarrow indicate whether higher or lower values are desired, respectively. (a) is the outcome of optimization with L1-loss which results in better PSNR and SSIM scores than (b) which is the result of optimization with perceptually oriented losses. On the contrary, (b) appears more photo-realistic which is reflected in a better Learned Perceptual Image Patch Similarity (LPIPS) score. Image from BSD100 [52].

However, it has been proven that low distortion and high perceptual quality are at odds with each other [9], meaning that optimizing for a lower distortion comes at the expense of lower perceptual quality and vice versa. Figure 1.10 visually illustrates this phenomenon, by showing two reconstructions of the same LR image produced by the same SR model (Residual-in-Residual Dense Block (RRDB)), but optimized with different goals. Figure 1.10 (a) is the result of optimization with L1-loss while a combination of L1, perceptual [33] and GAN-loss [40] was used for (b). More recently, flow-based [48] and diffusion models [41, 65] has been used to obtain more photo-realistic results by sampling

directly from the posterior distribution, rather than estimating the mean of all possible solutions. However, while the reconstructed images might look good, the content and details are usually not aligned with the true HR image.

3.2 Current Challenges in Super-resolution Research

Deep Neural Networks (DNNs) are prone to overfitting to the training distribution, which can lead to a significant drop in performance when evaluated on out-of-distribution samples. As such, a significant limitation of most existing SoTA DNNs-based SR methods is their limited ability to generalize to unseen images [44, 47]. Part of the reason is the common practice of developing and testing the SR methods on datasets consisting of synthetic LR images obtained by simple downsampling of HR images, typically using bicubic interpolation. However, the SR models only learn to inverse the process of the predefined degradations present in the training data. Hence, this protocol can be problematic, as the downsampling process significantly changes the low-level characteristics of the images, *e.g.* by reducing noise. As a result, the LR images become an oversimplification of the real conditions, since real-world images often exhibit complex combinations of degradations, as discussed in Section 2. Consequently, images of this nature pose a much more challenging task for the SR algorithms, as they must simultaneously tackle artifact suppression and resolution/detail enhancement. However, acquiring real LR images and their corresponding Ground-Truths (GTs) for training purposes is both labor-intensive and infeasible in addressing the generalization issue. Thus, no straightforward approach currently exists to enable the SR models to generalize to real-world images. Consequently, addressing this challenge has recently received increasing attention from the research community, known as real-world super-resolution [2], and has also constituted a significant part of this PhD project.

3.3 Image Quality Assessment

Super-resolution algorithms are typically assessed by different types of distortion measures that can be grouped into Full-Reference Image Quality Assessment (FR-IQA) and No-Reference Image Quality Assessment (NR-IQA) based methods. The former requires GT target images, and are therefore usually not directly applicable for the real-world scenario. However, as they are still widely used for evaluating performance in the synthetic setting, we cover the most popular methods in this category below. On the contrary, requiring only the super-resolved image, NR-IQA methods are useful for evaluation without GTs, but since the predictions of such methods are typically based on statistics, or learned models, the accuracy is not as high as directly measuring the distortion from a GT image. Generally, the evaluation of image reconstruction

3. Introduction to Super-Resolution

methods without GT data is an unsolved problem, and the lack of reliable NR-IQA methods complicates the assessment and comparison to other approaches. However, some research relies on subjective evaluation methods, but the procedure is cumbersome and time-consuming. Lastly, there also exist other less-used methods, like reduced-reference methods, which we will not cover in the overview below.

Full-Reference Image Quality Assessment (FR-IQA)

These methods compare the quality of a processed image with a GT reference image, by measuring the difference between the two. The most widely used methods in this category are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity index (SSIM) given by the following equations, respectively:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \quad (1.2)$$

where MAX denotes the maximum possible pixel value (e.g., 255 for 8-bit images), and MSE represents the mean squared error between the original and reconstructed image.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1.3)$$

where x and y are the input and reconstructed image, respectively, μ_x and μ_y are the mean values of x and y , σ_x and σ_y are the standard deviations of x and y , σ_{xy} is the covariance between x and y , and C_1 and C_2 are constants to stabilize the division. In general, PSNR is more concerned with the pixel-level differences between the input and reconstructed image, while SSIM is more geared towards the visual quality by considering aspects of luminance, contrast, and structural information. However, both PSNR and SSIM are poorly correlated with human perception. As such, there also exist methods that are specifically targeted toward assessment of image quality similar to human judgment. These methods include learning-based methods such as LPIPS [90] and Deep Image Structure and Texture Similarity (DISTS) [18], and more recent Transformer based methods such as IQT [13]. The challenge report on the NTIRE 2022 Challenge on Perceptual Image Quality Assessment [26] provides an overview of the current SoTA within the topic.

No-Reference Image Quality Assessment (NR-IQA)

These methods assess the image quality of an image without requiring a reference image and are typically based on natural image statistics or models learned on databases of various image distortions. Popular methods in this

category include PI [8], NRQM [50], NIQE [55], PIQE [57] and BRISQUE [56], which is similar to the evaluation protocol used for the 2020 NTIRE challenge on Real-World Super-Resolution (RWSR) [2]. However, assessment without reference images is inherently difficult, and as such the performance of these metrics is still unsatisfactory and unreliable [83].

Subjective Image Quality Assessment

In the case of real images without GTs, human observers can be used to rate the perceived quality of the enhanced images for evaluation of the SR models. A commonly used method for this is Mean Opinion Rank (MOR), where the participants are asked to rank images produced by different approaches based on their perceived quality. Similarly, Mean Opinion Score (MOS) involves rating images with a score, typically from 1 (bad) to 5 (good) followed by taking the mean of all ratings. However, while these methods can be reliable when the amount of participants is sufficiently large, the process is time-consuming and cumbersome.

3.4 Ethics in Relation to Super-Resolution

Even with the use of simple hand-crafted methods, manipulation of images has raised ethical questions [67]. Consequently, in [15] 12 guidelines for scientific digital image manipulation were presented. Recently, Artificial Intelligence (AI) have become a topic that has generated controversy and raised concerns regarding privacy, ethics, and civil rights due to its promising capabilities and performance. In response to this, the European Commission has published a set of ethical guidelines for trustworthy AI [14]. In summary, a trustworthy AI should be:

- Lawful - respecting all applicable laws and regulations
- Ethical - respecting ethical principles and values
- Robust - both from a technical perspective while taking into account its social environment

Since this PhD thesis focuses on investigating AI-based SR algorithms and their practical applications, such as surveillance, it is important to discuss ethical considerations within this specific context.

Hallucinations

Recent DNN-based SR methods enhance LR images by hallucinating, or predicting, missing HR details based on their learned priors. However, uncertainties can arise due to both aleatoric (data-centric) and epistemic (model-centric)

uncertainties [30], which can result in undesirable artifacts or suppression of important details. This problem is particularly evident when the image to be enhanced is different from training data distribution [11, 44], which can lead to adverse effects. One example is SR of face images, where facial landmarks, such as birthmarks, may be lost, or wrongly inferred [32]. While this might not be critical for the overall visual quality of the reconstructed face, which might appear perceptually realistic, the consequences can be profound in forensics and legal scenarios, ultimately leading to wrongful convictions, or failure to solve criminal cases [76]. As such, this limitation should be taken into consideration when using DNN methods.

Bias and fairness

DNN based SR methods can inherit biases present in the training data. Hence, it's important to identify if the training dataset is biased towards a specific ethnicity and possibly implement methods to mitigate the reduced performance on minorities less represented in the training data. A recent example is PULSE [53], a face super-resolution algorithm that caused controversy as it had a tendency to turn dark faces white. The reason for this was found to be the reliance on racially imbalanced training data [77]. One can refer to [31] for a study on the issues of fairness for generative algorithms like SR.

Trustworthiness

DNNs are inherently black-boxes, making it difficult to understand the reasoning behind the outcome, which can introduce trust issues for end users. Moreover, most DNNs are optimized toward specific performance metrics that are rarely good indicators of a trustworthy model once trained [51]. Furthermore, while discriminative models, like image classifiers [27, 39, 69], typically output their prediction with a confidence value, current generative image restoration methods do not provide such uncertainty information to the end user. Although the emerging discipline of explainable AI (XAI) has been comprehensively studied for discriminative models [49, 64], much less attention has been given to generative models, in general [70]. The few existing works in the literature mostly deal with natural language [85], and software code inference [71], but none consider the problem of image restoration, such as SR.

Surveillance

Footage from the surveillance cameras is often used together with video analysis software to automatically monitor activities and recognize specific actions [10, 54]. Combining SR with such technologies has the potential to further enhance the performance which raises several ethical concerns, including;

how can it be ensured that potential improvements of automated surveillance systems are not used to cause increased discrimination, stigmatization, or infringement on privacy and civil rights. A more elaborate discussion and suggestion for regulations are presented in [4].

During the PhD period, we have been investigating solutions to some of the aforementioned challenges, while we for others only take the limitations into consideration during our work, and leave actual research into possible solutions for future work. The following chapter covers our work on advancing the field of RWSR.

References

- [1] *Super-resolution imaging*, ser. Digital imaging and computer vision. CRC Press, 2011.
- [2] A. Lugmayr et al., “Ntire 2020 challenge on real-world image super-resolution: Methods and results,” *CVPR Workshops*, 2020.
- [3] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, “RELLISUR: A real low-light image super-resolution dataset,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/7ef605fc8dba5425d6965fbd4c8fbe1f-Paper-round2.pdf>
- [4] A. A. Adams and J. M. Ferryman, “The future of video analytics for surveillance and its ethical implications,” *Security Journal*, vol. 28, pp. 272–289, 2015.
- [5] N. Ahn, B. Kang, and K.-A. Sohn, “Image super-resolution via progressive cascading residual network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 791–799.
- [6] S. Arildsen, “Overvågning: Halvanden million kameraer følger dig,” *Sjællandske Nyheder*. [Online]. Available: <https://www.sn.dk/danmark/overvaagning-halvanden-million-kameraer-foelger-dig/>
- [7] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Variational bayesian super resolution,” *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2010.
- [8] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, “The 2018 pirm challenge on perceptual image super-resolution,” in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 334–355.
- [9] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6228–6237. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html

References

- [10] BriefCam, “Video analytics.” [Online]. Available: <https://www.briefcam.com/technology/video-analytics/>
- [11] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, “Real-world single image super-resolution: A brief review,” *Inf. Fusion*, vol. 79, pp. 124–145, 2022. [Online]. Available: <https://doi.org/10.1016/j.inffus.2021.09.005>
- [12] X. Chen, X. Wang, J. Zhou, and C. Dong, “Activating more pixels in image super-resolution transformer,” *CoRR*, vol. abs/2205.04437, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2205.04437>
- [13] M. Cheon, S. Yoon, B. Kang, and J. Lee, “Perceptual image quality assessment with transformers,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 433–442. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021W/NTIRE/html/Cheon_Perceptual_Image_Quality_Assessment_With_Transformers_CVPRW_2021_paper.html
- [14] E. Commission. (2018) Ethics guidelines for trustworthy ai. [Online]. Available: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- [15] D. W. Crome, “Avoiding twisted pixels: Ethical guidelines for the appropriate use and manipulation of scientific digital images,” *Sci. Eng. Ethics*, vol. 16, no. 4, pp. 639–667, 2010. [Online]. Available: <https://doi.org/10.1007/s11948-010-9201-y>
- [16] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang, “Second-order attention network for single image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 11 065–11 074.
- [17] M. Delbracio, D. Kelly, M. S. Brown, and P. Milanfar, “Mobile computational photography: A tour,” *CoRR*, vol. abs/2102.09000, 2021. [Online]. Available: <https://arxiv.org/abs/2102.09000>
- [18] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *CoRR*, vol. abs/2004.07728, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07728>
- [19] S. F. Dodge and L. J. Karam, “Understanding how image quality affects deep neural networks,” in *Eighth International Conference on Quality of Multimedia Experience, QoMEX 2016, Lisbon, Portugal, June 6-8, 2016*. IEEE, 2016, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/QoMEX.2016.7498955>
- [20] C. Dong, C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [21] O. Gallo, N. Gelfandz, W.-C. Chen, M. Tico, and K. Pulli, “Artifact-free high dynamic range imaging,” in *2009 IEEE International conference on computational photography (ICCP)*. IEEE, 2009, pp. 1–7.
- [22] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, “Deep joint demosaicking and denoising,” *ACM Trans. Graph.*, vol. 35, no. 6, pp. 191:1–191:12, 2016. [Online]. Available: <https://doi.org/10.1145/2980179.2982399>

References

- [23] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 349–356.
- [24] R. C. Gonzlez and R. E. Woods, *Digital image processing, 3rd Edition*. Pearson Education, 2008. [Online]. Available: <https://www.worldcat.org/oclc/241057034>
- [25] H. Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, 2009. [Online]. Available: <https://doi.org/10.1093/comjnl/bxm075>
- [26] J. Gu, H. Cai, C. Dong, J. S. Ren, R. Timofte, Y. Gong, S. Lao, S. Shi, J. Wang, S. Yang, T. Wu, W. Xia, Y. Yang, M. Cao, C. Heng, L. Fu, R. Zhang, Y. Zhang, H. Wang, H. Song, J. Wang, H. Fan, X. Hou, M. Sun, M. Li, K. Zhao, K. Yuan, Z. Kong, M. Wu, C. Zheng, M. V. Conde, M. Burchi, L. Feng, T. Zhang, Y. Li, J. Xu, H. Wang, Y. Liao, J. Li, K. Xu, T. Sun, Y. Xiong, A. Keshari, Komal, S. Thakur, V. Jakhetiya, B. N. Subudhi, H. Yang, H. Chang, Z. Huang, W. Chen, S. Kuo, S. Dutta, S. D. Das, N. A. Shah, and A. K. Tiwari, "NTIRE 2022 challenge on perceptual image quality assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*. IEEE, 2022, pp. 950–966. [Online]. Available: <https://doi.org/10.1109/CVPRW56347.2022.00109>
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.
- [28] K. Hirakawa and T. W. Parks, "Joint demosaicing and denoising," *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2146–2157, 2006.
- [29] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206.
- [30] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, 2021. [Online]. Available: <https://doi.org/10.1007/s10994-021-05946-3>
- [31] A. Jalal, S. Karmalkar, J. Hoffmann, A. Dimakis, and E. Price, "Fairness for image generation with uncertain sensitive attributes," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 4721–4732. [Online]. Available: <http://proceedings.mlr.press/v139/jalal21b.html>
- [32] J. Jiang, C. Wang, X. Liu, and J. Ma, "Deep learning-based face super-resolution: A survey," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 13:1–13:36, 2023. [Online]. Available: <https://doi.org/10.1145/3485132>
- [33] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 694–711. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_43

References

- [34] N. K. Kalantari, R. Ramamoorthi *et al.*, “Deep high dynamic range imaging of dynamic scenes.” *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144–1, 2017.
- [35] S. Karahan, M. K. Yildirim, K. Kirta, F. S. Rende, G. Butun, and H. K. Ekenel, “How image degradations affect deep cnn-based face recognition?” in *2016 International Conference of the Biometrics Special Interest Group, BIOSIG 2016, Darmstadt, Germany, September 21-23, 2016*, ser. LNI, A. Brömmel, C. Busch, C. Rathgeb, and A. Uhl, Eds., vol. P-260. GI / IEEE, 2016, pp. 313–320. [Online]. Available: <https://doi.org/10.1109/BIOSIG.2016.7736924>
- [36] R. G. Keys, “Cubic Convolution Interpolation for Digital Image Processing,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 29, pp. 1153–1160, Jan. 1981.
- [37] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.
- [38] T. Komatsu, K. Aizawa, T. Igarashi, and T. Saito, “Signal-processing based method for acquiring very high resolution images with multiple cameras and its theoretical analysis,” *IEE Proceedings I (Communications, Speech and Vision)*, vol. 140, no. 1, pp. 19–25, 1993.
- [39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
- [40] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [41] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, “Srdiff: Single image super-resolution with diffusion probabilistic models,” *Neurocomputing*, vol. 479, pp. 47–59, 2022. [Online]. Available: <https://doi.org/10.1016/j.neucom.2022.01.029>
- [42] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *IEEE International Conference on Computer Vision Workshops*, 2021.
- [43] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [44] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, “Blind image super-resolution: A survey and beyond,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5461–5480, 2023. [Online]. Available: <https://doi.org/10.1109/TPAMI.2022.3203009>
- [45] H. Liu, H. Li, X. Wang, H. Li, M. Ou, L. Hao, Y. Hu, and J. Liu, “Understanding how fundus image quality degradation affects cnn-based diagnosis,” in *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2022, Glasgow, Scotland, United Kingdom, July 11-15, 2022*. IEEE, 2022, pp. 438–442. [Online]. Available: <https://doi.org/10.1109/EMBC48229.2022.9871507>

References

- [46] Y. Liu, S. Wang, J. Zhang, S. Wang, S. Ma, and W. Gao, "Iterative network for image super-resolution," *IEEE Transactions on Multimedia*, vol. 24, pp. 2259–2272, 2021.
- [47] A. Lugmayr, M. Danelljan, R. Timofte, M. Fritsche, S. Gu, K. Purohit, P. Kandula, M. Suin, A. N. Rajagoapalan, N. H. Joon, Y. S. Won, G. Kim, D. Kwon, C. Hsu, C. Lin, Y. Huang, X. Sun, W. Lu, J. Li, X. Gao, S. Bell-Kligler, A. Shocher, and M. Irani, "Aim 2019 challenge on real-world image super-resolution: Methods and results," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3575–3583.
- [48] A. Lugmayr, M. Danelljan, F. Yu, L. V. Gool, and R. Timofte, "Normalizing flow as a flexible fidelity objective for photo-realistic super-resolution," in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*. IEEE, 2022, pp. 874–883. [Online]. Available: <https://doi.org/10.1109/WACV51458.2022.00095>
- [49] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4765–4774. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [50] C. Ma, C. Yang, X. Yang, and M. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Underst.*, vol. 158, pp. 1–16, 2017. [Online]. Available: <https://doi.org/10.1016/j.cviu.2016.12.009>
- [51] A. Marathe, P. Jain, R. Walambe, and K. Kotecha, "Restorex-ai: A contrastive approach towards guiding image restoration via explainable AI systems," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*. IEEE, 2022, pp. 3029–3038. [Online]. Available: <https://doi.org/10.1109/CVPRW56347.2022.00342>
- [52] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 416–423 vol.2.
- [53] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: self-supervised photo upsampling via latent space exploration of generative models," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 2434–2442. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00251>
- [54] Milestone Systems, "Milestone XProtect." [Online]. Available: <https://www.milestonesys.com/video-technology/platform/xprotect/>
- [55] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
- [56] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012. [Online]. Available: <https://doi.org/10.1109/TIP.2012.2214050>

References

- [57] V. N., P. D., M. C. Bh., S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Twenty First National Conference on Communications, NCC 2015, Mumbai, India, February 27 - March 1, 2015*. IEEE, 2015, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/NCC.2015.7084843>
- [58] J. Nakamura, *Image sensors and signal processing for digital still cameras*. CRC press, 2017.
- [59] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vision Appl.*, vol. 25, no. 6, pp. 1423–1468, Aug. 2014.
- [60] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12357. Springer, 2020, pp. 191–207. [Online]. Available: https://doi.org/10.1007/978-3-030-58610-2_12
- [61] H. Nyquist, "Certain topics in telegraph transmission theory," *Proc. IEEE*, vol. 90, no. 2, pp. 280–305, 2002. [Online]. Available: <https://doi.org/10.1109/5.989875>
- [62] M. Oldager, "Bil efterlyses i forbindelse med 22-årige mias forsvinden i aalborg," *Danmarks Radio*. [Online]. Available: <https://www.dr.dk/nyheder/seneste/bil-efterlyses-i-forbindelse-med-22-aarige-mias-forsvinden-i-aalborg>
- [63] G. Pavlidis, *Foundations of Photography: A Treatise on the Technical Aspects of Digital Photography*. Springer Nature, 2022.
- [64] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should I trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, Eds. ACM, 2016, pp. 1135–1144. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [65] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *CoRR*, vol. abs/2104.07636, 2021. [Online]. Available: <https://arxiv.org/abs/2104.07636>
- [66] N. R. Shah and A. Zakhor, "Resolution enhancement of color video sequences," *IEEE transactions on Image Processing*, vol. 8, no. 6, pp. 879–885, 1999.
- [67] M. J. Shapter, "Image manipulation and the question of ethics," *Journal of Audiovisual Media in Medicine*, vol. 16, no. 3, pp. 130–132, 1993. [Online]. Available: <https://doi.org/10.3109/17453059309064840>
- [68] J. Shermeyer and A. V. Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1432–1441. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/EarthVision/Shermeyer_The_Effects_of_Super-Resolution_on_Object_Detection_Performance_in_Satellite_CVPRW_2019_paper.html
- [69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

References

- [70] J. Sun, Q. V. Liao, M. J. Muller, M. Agarwal, S. Houde, K. Talamadupula, and J. D. Weisz, "Investigating explainability of generative AI for code through scenario-based design," in *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, G. Jacucci, S. Kaski, C. Conati, S. Stumpf, T. Ruotsalo, and K. Gajos, Eds. ACM, 2022, pp. 212–228. [Online]. Available: <https://doi.org/10.1145/3490099.3511119>
- [71] —, "Investigating explainability of generative AI for code through scenario-based design," in *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, G. Jacucci, S. Kaski, C. Conati, S. Stumpf, T. Ruotsalo, and K. Gajos, Eds. ACM, 2022, pp. 212–228. [Online]. Available: <https://doi.org/10.1145/3490099.3511119>
- [72] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020. [Online]. Available: <https://doi.org/10.1016/j.neunet.2020.07.025>
- [73] M. Tipping and C. Bishop, "Bayesian image super-resolution," *Advances in neural information processing systems*, vol. 15, 2002.
- [74] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," *Multiframe image restoration and registration*, vol. 1, pp. 317–339, 1984.
- [75] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A frequency domain approach to registration of aliased images with application to super-resolution," *EURASIP journal on advances in signal processing*, vol. 2006, pp. 1–14, 2006.
- [76] S. Villena, M. Vega, J. Mateos, D. Rosenberg, F. Murtagh, R. Molina, and A. K. Katsaggelos, "Image super-resolution for outdoor digital forensics. usability and legal aspects," *Comput. Ind.*, vol. 98, pp. 34–47, 2018. [Online]. Available: <https://doi.org/10.1016/j.compind.2018.02.004>
- [77] J. Vincent, "What a machine learning tool that turns obama white can (and can't) tell us about ai bias," *The Verge*. [Online]. Available: <https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias>
- [78] T. Wang, W. Sun, H. Qi, and P. Ren, "Aerial image super resolution via wavelet multiscale convolutional neural networks," *IEEE Geosci. Remote. Sens. Lett.*, vol. 15, no. 5, pp. 769–773, 2018. [Online]. Available: <https://doi.org/10.1109/LGRS.2018.2810893>
- [79] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.
- [80] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 977–97709.
- [81] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

References

- [82] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Krainin, C. Liang, M. Levoy, and P. Milanfar, "Handheld multi-frame super-resolution," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 28:1–28:18, 2019. [Online]. Available: <https://doi.org/10.1145/3306346.3323024>
- [83] S. Xu, S. Jiang, and W. Min, "No-reference/blind image quality assessment: A survey," *IETE Technical Review*, vol. 34, no. 3, pp. 223–245, 2017.
- [84] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.
- [85] J. Yu, A. I. Cristea, A. Harit, Z. Sun, O. T. Aduragba, L. Shi, and N. A. Moubayed, "INTERACTION: A generative XAI framework for natural language inference explanations," in *International Joint Conference on Neural Networks, IJCNN 2022, Padua, Italy, July 18-23, 2022*. IEEE, 2022, pp. 1–8. [Online]. Available: <https://doi.org/10.1109/IJCNN55064.2022.9892336>
- [86] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 711–730.
- [87] J. Zhang, J. Pan, J. S. J. Ren, Y. Song, L. Bao, R. W. H. Lau, and M. Yang, "Dynamic scene deblurring using spatially variant recurrent neural networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 2521–2529. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Dynamic_Scene_Deblurring_CVPR_2018_paper.html
- [88] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017. [Online]. Available: <https://doi.org/10.1109/TIP.2017.2662206>
- [89] K. Zhang, W. Ren, W. Luo, W. Lai, B. Stenger, M. Yang, and H. Li, "Deep image deblurring: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2103–2130, 2022. [Online]. Available: <https://doi.org/10.1007/s11263-022-01633-5>
- [90] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 586–595.
- [91] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 294–310.
- [92] Y. Zhou, J. Jiao, H. Huang, Y. Wang, J. Wang, H. Shi, and T. S. Huang, "When awgn-based denoiser meets real noises," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on*

References

- Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020.* AAAI Press, 2020, pp. 13 074–13 081. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/7009>
- [93] C. Østergård, “Ringe kameraer gør overvågning ineffektiv,” *Ingeniøren*. [Online]. Available: <https://ing.dk/artikel/ringe-kameraer-goer-overvaagning-ineffektiv>

Chapter 2

Real-World Super-Resolution

1 Introduction



Fig. 2.1: An overview of the publications establishing the foundation for the findings presented in this chapter.

This chapter of the PhD thesis covers our work in the domain of RWSR. The focus has solely been on deep-learning based SR methods due to their promising performance in comparison to hand-crafted methods. In particular, CNNs have been widely used in the literature since they have shown to be capable of learning strong priors for guiding the SR process. Although the use of multiple frames, either in the form of video frames or a burst image sequence, can potentially improve the SR performance compared to using only a single image, they introduce the additional challenges of aligning the frames and compensating for motion. As such, we have been focusing exclusively on single-image based SR methods. As outlined in Section 3.2, the main difference between RWSR and classical SR is the ability to handle complex real-world images, which occurs mainly due to the domain gap between the synthetic images, typically used for training of the DNNs in the classical setting and real-world images. A visual comparison of these image domains can be seen in Figure 2.2.

Thus, when the input images deviate from the assumed training distribution, the performance of the SR model drops significantly [1, 34, 35]. This is



Fig. 2.2: A comparison between a synthetic LR image created by bicubic downsampling, a real LR image, and the corresponding $\times 4$ HR image. Images adapted from [6], Paper F.

especially pronounced if the images are corrupted by noise, blur, compression artifacts, or similar distortions, which is ill-fated as it is in scenarios like these where enhancement is most needed.

However, while classical SR is already a challenging and severely ill-posed inverse problem, due to the large solution space of plausible HR images for each LR image, RWSR pose an additional number of challenges due to the nature and complexity of real-world images. Some of these challenges include:

- Acquiring paired real-world LR and HR images is cumbersome and sometimes even impossible. This poses challenges to the supervised learning paradigm commonly employed in training DNN-based SR methods, as it relies on data pairs for forward propagation and subsequent error calculation within the network.
- The lack of HR GTs images, complicates the quantitative evaluation of the reconstruction performance of the algorithms. This hampers comparisons between different methods, assessment of algorithmic advancements during development, and quantitative evaluation.
- RWSR algorithms face the dual task of suppressing artifacts and enhancing details simultaneously. Distinguishing between details and artifacts proves challenging due to the complex distributions involved.
- Designing a RWSR model capable of generalizing to the potentially infinitely large distribution of real-world images resulting from various combinations of cameras, lenses, settings, and scenes is nontrivial.

In the following section, we will describe the SoTA within RWSR, highlighting different attempts aimed at addressing the aforementioned challenges.

2 State-of-the-Art

We begin the description of the SoTA works by establishing a taxonomy to position RWSR in the research field of image SR. By this taxonomy, we divide research in single-image SR into three groups:

- *Non-blind or classical SR*: assumes that the degradation process from HR to LR image is known (typically bicubic interpolation) [21, 29, 33].
- *Blind SR*: aims to restore images with unknown degradations, typically focused mostly on the blur kernel and experimentation on synthetic image data [11, 24, 39].
- *RWSR*: aims to restore real images with complex combinations of unknown real-world degradations, most often without access to the GT HR images [40, 48, 56].

Although consensus on the taxonomy presented above is not universally established among the research community, we distinguish blind SR from RWSR based on the following criteria. Literature on blind SR encompasses primarily theoretical and general methods that handle images without prior knowledge of the degradation process. In contrast, RWSR literature consists of more practically oriented methods specifically designed for applications in real-world scenarios. Following [17], the current body of research in RWSR can be categorized into distinct groups based on the following taxonomy:

- *Self-learning based methods*: exploits internal information in the input images [9, 39, 46].
- *Image pair based methods*: learns from collected real-world LR/HR image pairs [15, 51, 58].
- *Domain translation based methods*: aims to bridge the gap between two image domains [22, 27, 53].
- *Degradation modeling based methods*: aims to estimate the degradation parameters of the input images [47, 48, 56].

The following sections will provide an overview of the SoTA in RWSR and datasets for training and evaluation.

2.1 Real-World Super-Resolution Methods

The body of self-learning-based methods includes Shocher *et al.* [9], who propose a zero-shot approach where a small CNN is trained at test time, to learn an image specific SR model. Soh *et al.* [44] further extends this concept by

a meta-transfer learning phase which allows exploiting information from an external dataset as well. Later, Bell-Kligler *et al.* [11] designed a GAN-based mode to estimate blur kernels from LR images, which are then used in combination with the ZSSR SR model. In DAN [36], both steps are incorporated into a single model by alternating optimization. However, one drawback of self-learning-based methods is slow prediction times compared to other SR methods [50].

Another approach to RWSR is to collect real-world LR/HR image-pairs for training, either using a beam splitter [41], varying the focal length of a zoom-lens [15, 16, 51, 58], or by hardware binning [30]. While such data enables investigation and quantitative evaluation of a much harder task than SR of synthetic LR images, the models trained on such data do not perform well on images from other camera types than the ones used in the collection process since its not feasible to collect a large enough pool of images to represent all possible combinations of cameras, lenses, and settings. Furthermore, avoiding misalignment and other unwanted discrepancies, such as illumination changes, between the LR/HR image pairs is a challenging task on its own.

The idea of domain translation for the generation of realistic LR/HR training image pairs is explored in RealSR [27]. The RealSR degradation framework estimates blur kernels using KernelGAN [11], and samples realistic noise patches from a domain of real degraded images, which are used to degrade HR images in the target domain. As such, RealSR actively tries to estimate and extract image-specific degradation information from one domain, to enable translation from another image domain. This is different from the more recent brute-force-like approaches, such as BSRGAN [56], which basically tries to introduce as many as possible different combinations of degradations in the training images. However, while RealSR is able to adapt the SR model to real images from a specific domain, the trained model cannot generalize beyond this distribution. A model for translating real LR images to the bicubic domain, for leveraging generic SR models trained on such images is proposed in [42]. Moreover, CinCGAN [53] propose to use a cycle-in-cycle GAN to learn to map real noisy LR images to a domain of clean LR images followed by SR. However, these approaches essentially shift the generalization problem from the SR model to the translation model.

The current most effective and commonly employed approach in the literature for addressing the problem of RWSR is to explicitly model the image-specific degradations and incorporate them into the training data. In DSGAN [22] a GAN model is trained to introduce natural image characteristics to bicubically downsampled images, which in turn are used to train a SR model. One successful, but less elegant approach, is to create synthetic LR images with highly diverse combinations of degradation types. Following this line, BSRGAN [56] proposes a pipeline that degrades training images by random shuffling of down-sampling, blur, noise, and JPEG compression.

Real-ESRGAN [48] extends on this by proposing a second-order degradation pipeline, that applies the degradations more than once. Other works focus more on the network architectures, by trying to estimate the image-specific degradations at test time and adapt the features in the SR network accordingly [31, 40, 47, 59]. Nonetheless, a limitation that can be observed in most of the current RWSR methods is their inability to handle spatially variant degradations commonly encountered in real-world images.

2.2 Datasets

Research in RWSR often relies on the same image datasets as employed in the classical SR setting to enable training and quantitative evaluation. The most popular ones include DIV2K [7] and Flickr2K [45] which are typically used for training, and Set5 [12], Set14 [55], BSD100 [37], Manga109 [54], and Urban100 [26] that are typically used for testing. However, as these datasets only contain HR images, the LR images are often generated from the HR images by blurring, downsampling, and corruption by noise, typically using Gaussian distributions for the noise distribution and blur kernels [25]. Since such synthetic images do not resemble real-world images, researchers have also collected real-world LR/HR image pairs for both training and testing purposes. Table 2.1 provides an overview of the largest datasets of real LR/HR image pairs. However, while such datasets pose a more challenging reconstruction problem, training a SR model on such datasets does not lead to generalization beyond the camera types used in the dataset.

Table 2.1: Overview of real-world super-resolution datasets of paired real LR and HR images. Table adapted from [4], Paper E.

Name	Year	LR/HR Pairs	Type	HR resolution	Method	Content
Qu et al. [41]	2016	31	RAW	2.3MPiX	Beam-splitter	Faces
RealSR [15]	2019	595	RGB	0.48 to 5.28MPiX	Zoom lens	In/outdoor scenes
City100 [16]	2019	100	RGB	1.06MPiX	Zoom + translation	Postcards
SupER [30]	2019	5,670	Grayscale	2.2MPiX	Hardware binning	Indoor lab
SR-RAW [58]	2019	500	RAW	12MPiX	Zoom lens	In/outdoor scenes
DRealSR [51]	2020	2,507	RGB	20 to 24MPiX	Zoom lens	In/outdoor scenes

3 Scientific Contributions

The work done in this part of the PhD has been aimed at pushing the SoTA in RWSR by addressing some of the challenges outlined in Section 1. This has led to four papers, namely Paper A, Paper B, Paper C, and Paper D. The first three papers have already been published at computer vision and machine-learning-focused research outlets, while the fourth paper (Paper D) is currently under



Fig. 2.3: $\times 4$ SR of a real LR face image from the Chokepoint DB [52]. Compared to ESRGAN [49], which amplifies the corruptions, and RealSR [27], which fails to improve the amount of details, our approach exhibits superior performance in reducing artifacts and enhancing details, thereby resulting in a more visually appealing reconstruction. Images from [5], Paper A

review.

In Paper A we initiated the research on RWSR by specifically focusing on enhancing face images from surveillance cameras. Although one of the first SR methods specifically for enhancement of face images was proposed by Baker and Kanade back in 2000 [10], subsequent research in this domain has paid limited attention to the challenges associated with real-world face images [23]. Furthermore, for simplicity, most researchers have primarily focused on methods applicable only to images with a size of 16×16 pixel and centered faces [13, 14, 18], which limits the practical application of these approaches.

To address these limitations, we explored the feasibility of employing domain translation techniques to learn a SR model capable of enhancing real-world face images from surveillance cameras in Paper A. Our approach drew inspiration from [27], who proposed a degradation framework for translating clean HR images in a target domain to a source domain of degraded LR images. We build a pool of estimated translation parameters from the source domain by utilizing KernelGAN [11] for blur kernel estimation, and extraction of noise patches for estimation of the image-specific noise. By randomly selecting and applying blur kernels and noise patches from the pool, we performed the translation of clean HR images to the source domain of real surveillance images.

However, since the footage from surveillance cameras is often corrupted by compression artifacts, this pipeline is not sufficient to fully capture the degradations present in such images. Consequently, we extended the framework to also incorporate compression artifacts. Specifically, we degraded the translated images by JPEG compression, which enabled the SR model to learn JPEG artifact suppression, rather than wrongly amplifying the compression

3. Scientific Contributions

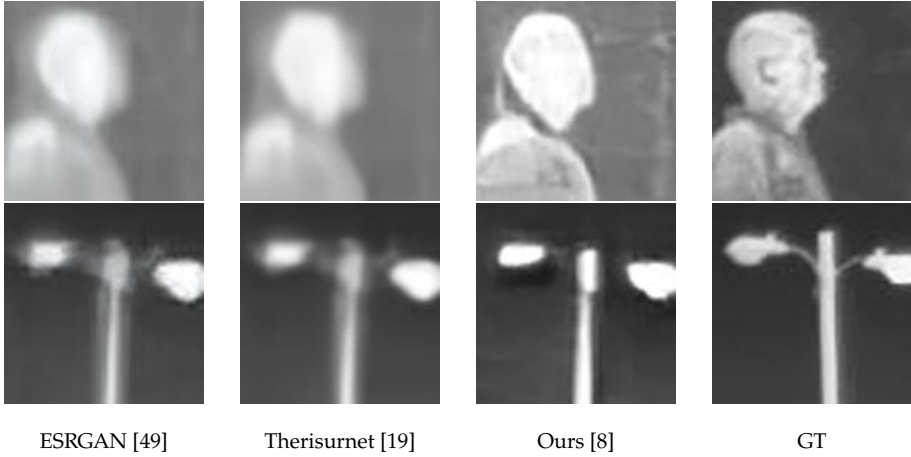


Fig. 2.4: $\times 4$ SR of real LR thermal images from the Domo validation subset of the PBVS challenge [43]. As seen, our method produces sharper and more well-defined reconstructions. Images from [8], Paper B.

patterns. We employed the GAN architecture and RRDB backbone model from ESRGAN [49], but observed that the perceptual quality could be further improved by exchanging the VGG-loss [28] with LPIPS [57] loss.

Figure 2.3 provides a comparison of our proposed method, ESRGAN, and RealSR when evaluated on a real face image obtained from a surveillance camera. As illustrated, our method excels in suppressing undesirable artifacts and enhancing high-frequency details, resulting in sharper and visually more appealing images. For evaluation on the Chokepoint testset, our approach achieved a superior MOR score of 1.43, outperforming the baseline score of 3.39. Through this study, we demonstrated the importance of incorporating as many degradation factors as possible during the model training phase for improved practical application of DNN-based SR algorithms. For future work, it could be interesting to experiment with additional factors such as chromatic aberration.

In Paper B we leverage the knowledge gained in Paper A to explore the applicability of a similar approach for SR of real thermal images. Until now, the research efforts concerning SR of thermal images, which are inherently constrained by the limitations of thermal imaging technology, have primarily concentrated on synthetic settings. As depicted in Figure 2.4, our approach is capable of producing thermal images with more clearly defined edges and better contrast and sharpness compared to the existing SoTA method Therisurnet [19] and the baseline ESRGAN [49]. Furthermore, our approach achieved a superior MOR score of 1.45, outperforming the baseline score of 3.20. As such,

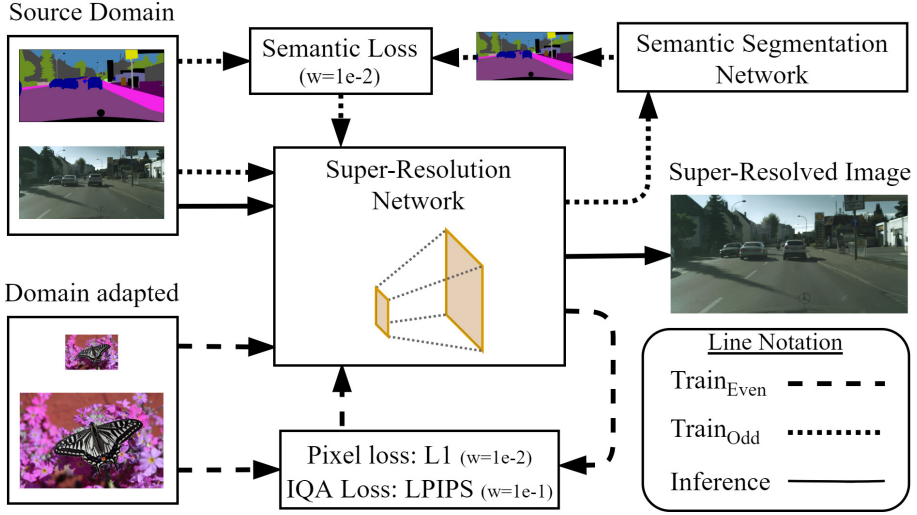


Fig. 2.5: “Schematic overview of our proposed Semantic Segmentation Guided Real-World Super-Resolution (SSG-RWSR). To learn to perform RWSR we leverage both guiding from an auxiliary semantic segmentation task and domain adaptation. At test time, the semantic segmentation network is de-coupled, and as such no semantic labels are required to super-resolve the LR test images.” [3]. Figure from [3], Paper C.

we concluded that SoTA RWSR methods for RGB images, are also beneficial for enhancement of thermal images.

In Paper C, we introduced a novel approach to improve upon the domain translation method used in our previous papers, which main limitation was the limited coverage of degradation types. Although the framework was flexible towards manual inclusion of additional degradation types, a more automated solution would likely be beneficial for the reconstruction performance. Our idea for Paper C originated from our investigations in Paper G, where we leveraged Super-Resolution (SR) to enhance the performance of a semantic segmentation model.

During that study, we observed that when optimizing a SR model solely based on the loss from the semantic segmentation task, the super-resolved images exhibited enhanced sharpness and reduced noise, compared to the LR input images. However, to maintain color consistency and rich textures, RGB image guidance remained necessary. Consequently, we hypothesized that incorporating segmentation loss as guidance during the SR learning process could lead to improved performance compared to relying solely on domain translation. As such, we undertook the first attempt at combining domain translation with guidance from pixel-wise image segmentation, aiming to address the supervised learning challenge in RWSR. An overview of the proposed framework can be seen in Figure 2.5.

3. Scientific Contributions

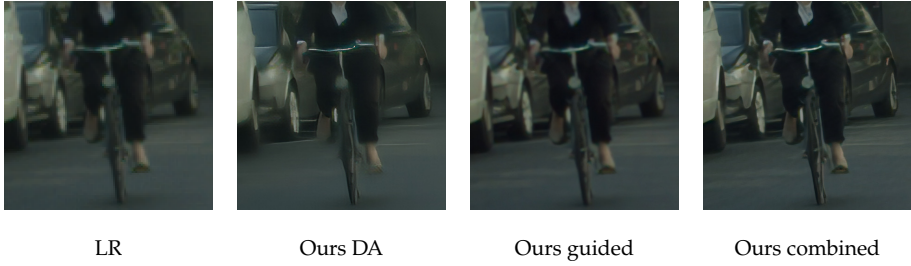


Fig. 2.6: Super-resolution ($\times 4$) of a real image from the Cityscapes dataset [20]. By combining domain adaptation (DA) and guidance by semantic segmentation, our proposed method reconstructs more visually pleasing images” [3]. Images from [3], Paper C.

Despite the fact that the semantic labels exist in the same resolution space as the LR training images, we discovered that guiding the SR model with the segmentation loss, specifically cross-entropy loss, in combination with domain translation facilitated the generation of HR images with sharper object boundaries and lower noise. Figure 2.6 visually illustrate the contribution of each component, clearly demonstrating that the combined approach yields the clearest and most detailed reconstruction of the LR input image. In the quantitative assessment conducted on the Cityscapes dataset [20], our approach demonstrated superior MOR score of 1.21, surpassing the runner-up method that achieved a score of 2.75.

Lastly, in our research on RWSR we conducted an investigation on the significance of spatially variant degradations in real LR images in Paper D. This study was initiated by an evaluation of the performance of the current SR RWSR method which exhibited significant inconsistency when faced with non-uniform degradations across an image.

To visualize this issue, Figure 2.7 presents an example where we generated an LR image with a uniform background (50% gray) and additive white Gaussian noise with $\sigma = 30$ in three different areas of the image. We then reconstructed the image using three different SoTA RWSR methods. Despite the fact that these SoTA RWSR methods were trained on images with this specific noise distribution, we hypothesized that their assumption of uniformly distributed degradations limits their ability to effectively remove the noise and reconstruct the clean image. Interestingly, certain methods partially removed noise from specific areas, while others left the noise unaltered.

This observation gave rise to an investigation into methods for introducing non-uniform noise to the training images, developing a RWSR algorithm capable of handling images with such complex degradation distributions, and evaluating whether this can help improve the reconstruction performance on real LR images.

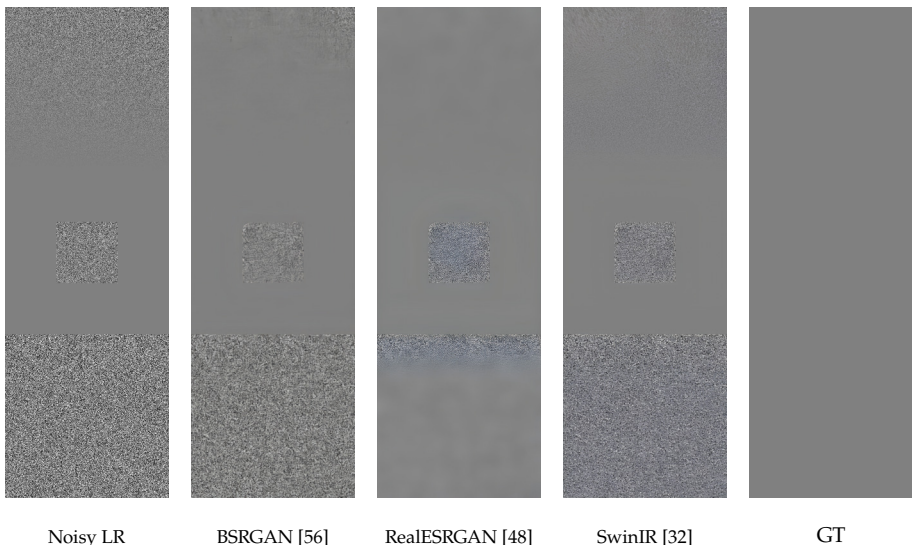


Fig. 2.7: Visualization demonstrating the inability of SoTA RWSR methods to remove patches of white Gaussian noise with $\sigma = 30$, which are superimposed on a uniform background, thus simulating spatially variant noise. This observation is noteworthy considering that the training process of these methods includes uniform noise of this specific distribution.

First, to facilitate the evaluation, we collected a dataset comprising real LR/HR image pairs with varying noise intensities, as existing real SR datasets were found to be inadequate in terms of complexity and degradation strength. In our dataset, named the SVSR dataset, we obtained the scale difference by varying the focal length of a zoom lens, while different ISO settings were employed to introduce varying noise levels. Visual examples from the dataset can be seen in Figure 2.8. However, since this is a dataset for evaluation only, due to its limited size, we further proposed a method to create LR training image with spatially varying degradations.

Drawing inspiration from the observation that noise tends to be more pronounced in darker regions of images, [38], we devised a noise degradation

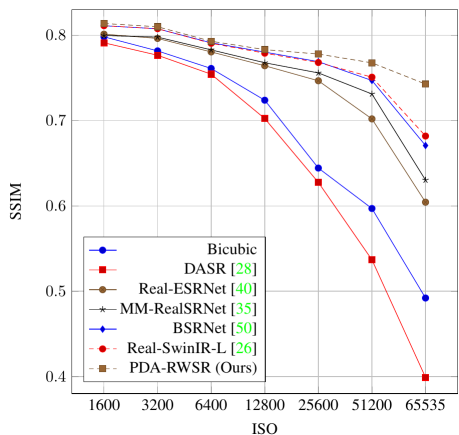


Fig. 2.10: "Plot of how the performance (SSIM) of SoTA methods decrease as the ISO (noise levels) in the SVSR benchmarking dataset increases. On the contrary, our PDA-RWSR has a more consistent performance across the range." [2]. Figure from [2], Paper D.

3. Scientific Contributions

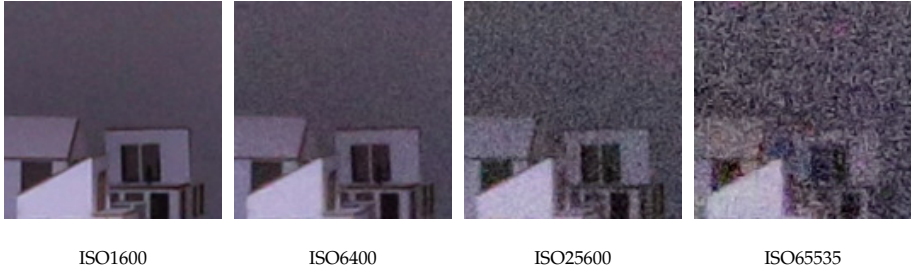


Fig. 2.8: “Visual examples from the SVSR benchmarking dataset illustrating how the noise level changes at different ISO settings for images captured with the Canon EOS 6D camera.” [2]. Images from [2], Paper D.

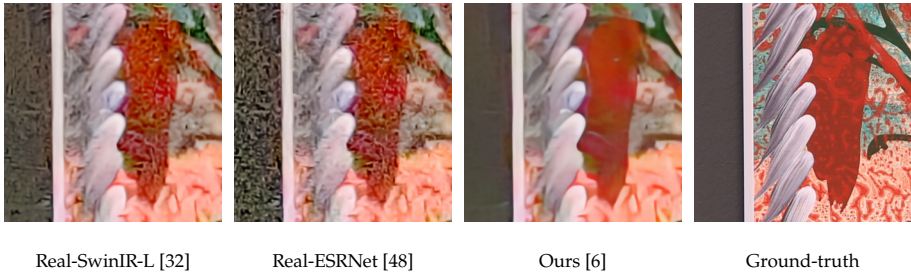


Fig. 2.9: “Visual comparison of the reconstruction performance on the SVSR dataset. In comparison to the SoTA approaches, our Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR) produces more visually faithful results with fewer artifacts.” [2]. Images from [2], Paper D.

model based on blending with masks created by image intensity thresholding, among others. To ensure that the SR model was aware of the spatially variant degradations, we designed a degradation feature extraction network capable of generating maps with pixel-wise degradation information from the degraded LR image. These maps were then used to modulate the features in the SR model, enabling the reconstruction process to adapt specifically to the degradation present in each pixel of the image. As demonstrated in Figure 2.9, our PDA-RWSR approach exhibits superior performance in removing spatially variant noise while preserving details. In terms of objective evaluation on the challenging SVSR dataset, our approach exhibited the best performance across all metrics, particularly showcasing significant enhancements on the images with the most pronounced degradation, as visualized in Figure 2.10.

To summarize, through our research efforts from Paper A to Paper D, we have thoroughly investigated critical challenges in super-resolution of real-world images and proposed novel approaches to enhance the performance and advance the field. Our main contributions can be outlined as follows:

- Insights into the challenges associated with SR of real-world images,

References

specifically surveillance footage, images heavily degraded by noise, and thermal images (Paper A, Paper B, Paper C, Paper D).

- Introduction of a novel framework for SR of real-world face images with arbitrary sizes, surpassing the performance of existing SoTA methods (Paper A).
- Exploration of the applicability of SoTA RWSR methods for the natural image domain in the thermal domain (Paper B).
- Development of a novel learning scheme for RWSR that incorporates guidance from semantic segmentation, achieving superior results compared to current SoTA methods (Paper C).
- Introduction of a novel degradation pipeline that allows introducing spatially variant degradations to LR training images (Paper D).
- Previously unpublished insights into the performance of current SoTA RWSR methods when evaluated on images with severe and non-uniform degradations (Paper D).
- Development of a novel framework for pixel-wise degradation estimation and model adaptation for SR of real-world images with spatially variant degradations, outperforming the current SoTA (Paper D).
- Creation of a challenging evaluation dataset for RWSR, intended to facilitate further advancements in the performance of RWSR algorithms (Paper D).

These contributions collectively enhance our understanding of SR in real-world scenarios and provide valuable insights for advancing the field and improving the practical applications of SR.

References

- [1] A. Lugmayr et al., "Ntire 2020 challenge on real-world image super-resolution: Methods and results," *CVPR Workshops*, 2020.
- [2] A. Aakerberg, M. E. Helou, K. Nasrollahi, S. Süsstrunk, and T. M. and, "Pda-rwsr: Pixel-wise degradation adaptive real-world super-resolution," *Under review, ICCV 2023*, 2023.
- [3] A. Aakerberg, A. S. Johansen, K. Nasrollahi, and T. B. Moeslund, "Semantic segmentation guided real-world super-resolution," in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*. IEEE, 2022, pp. 449–458. [Online]. Available: <https://doi.org/10.1109/WACVW54805.2022.00051>

References

- [4] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, "RELLISUR: A real low-light image super-resolution dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/7ef605fc8dba5425d6965fbd4c8fbe1f-Paper-round2.pdf>
- [5] —, "Real-world super-resolution of face-images from surveillance cameras," *IET Image Process.*, vol. 16, no. 2, pp. 442–452, 2022. [Online]. Available: <https://doi.org/10.1049/ipr2.12359>
- [6] —, "Relief: Joint low-light image enhancement and super-resolution with transformers," in *Scandinavian Conference on Image Analysis (SCIA)*, ser. Lecture Notes in Computer Science. Springer, 2022, pp. 157–173.
- [7] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [8] M. M. J. Allahham, A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, "Real-world thermal image super-resolution," in *Advances in Visual Computing - 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part I*, ser. Lecture Notes in Computer Science, G. Bebis, V. Athitsos, T. Yan, M. Lau, F. Li, C. Shi, X. Yuan, C. Mousas, and G. Bruder, Eds., vol. 13017. Springer, 2021, pp. 3–14. [Online]. Available: https://doi.org/10.1007/978-3-030-90439-5_1
- [9] M. I. Assaf Shocher, Nadav Cohen, "'zero-shot" super-resolution using deep internal learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [10] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," in *2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, 13-15 June 2000, Hilton Head, SC, USA. IEEE Computer Society, 2000, pp. 2372–2379. [Online]. Available: <https://doi.org/10.1109/CVPR.2000.854852>
- [11] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 284–293.
- [12] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, 2012, pp. 1–10. [Online]. Available: <http://dx.doi.org/10.5244/C.26.135>
- [13] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 109–117. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Bulat_Super-FAN_Integrated_Facial_CVPR_2018_paper.html
- [14] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018*,

References

- Proceedings, Part VI*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11210. Springer, 2018, pp. 187–202. [Online]. Available: https://doi.org/10.1007/978-3-030-01231-1_12
- [15] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, “Toward real-world single image super-resolution: A new benchmark and a new model,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3086–3095.
- [16] C. Chen, Z. Xiong, X. Tian, Z. Zha, and F. Wu, “Camera lens super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1652–1660. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Camera_Lens_Super-Resolution_CVPR_2019_paper.html
- [17] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, “Real-world single image super-resolution: A brief review,” *Inf. Fusion*, vol. 79, pp. 124–145, 2022. [Online]. Available: <https://doi.org/10.1016/j.inffus.2021.09.005>
- [18] Z. Cheng, X. Zhu, and S. Gong, “Characteristic regularisation for super-resolving face images,” in *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 2020, pp. 2424–2433. [Online]. Available: <https://doi.org/10.1109/WACV45572.2020.9093480>
- [19] V. Chudasama, H. Patel, K. Prajapati, K. P. Upla, R. Ramachandra, K. Raja, and C. Busch, “Therisurnet - a computationally efficient thermal image super-resolution network,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [21] C. Dong, C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [22] M. Fritsche, S. Gu, and R. Timofte, “Frequency separation for real-world super-resolution,” in *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.
- [23] K. Grm, M. Pernus, L. Cluzel, W. J. Scheirer, S. Dobrisek, and V. Struc, “Face hallucination revisited: An exploratory study on dataset bias,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 2405–2413. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/Biometrics/Grm_Face_Hallucination_Revisited_An_Exploratory_Study_on_Dataset_Bias_CVPRW_2019_paper.html
- [24] J. Gu, H. Lu, W. Zuo, and C. Dong, “Blind super-resolution with iterative kernel correction,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [25] M. E. Helou, R. Zhou, and S. Ssstrunk, “Stochastic frequency masking to improve super-resolution and denoising networks,” in *Computer Vision - ECCV*

References

- 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, *Proceedings, Part XVI*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12361. Springer, 2020, pp. 749–766. [Online]. Available: https://doi.org/10.1007/978-3-030-58517-4_44
- [26] J. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206.
- [27] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, “Real-world super-resolution via kernel estimation and noise injection,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [28] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 694–711. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_43
- [29] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.
- [30] T. Köhler, M. Bätz, F. Naderi, A. Kaup, A. Maier, and C. Riess, “Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2944–2959, 2020. [Online]. Available: <https://doi.org/10.1109/TPAMI.2019.2917037>
- [31] J. Liang, H. Zeng, and L. Zhang, “Efficient and degradation-adaptive network for real-world image super-resolution,” in *European Conference on Computer Vision*, 2022.
- [32] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *IEEE International Conference on Computer Vision Workshops*, 2021.
- [33] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [34] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, “Blind image super-resolution: A survey and beyond,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5461–5480, 2023. [Online]. Available: <https://doi.org/10.1109/TPAMI.2022.3203009>
- [35] A. Lugmayr, M. Danelljan, R. Timofte, M. Fritsche, S. Gu, K. Purohit, P. Kandula, M. Suin, A. N. Rajagoapalan, N. H. Joon, Y. S. Won, G. Kim, D. Kwon, C. Hsu, C. Lin, Y. Huang, X. Sun, W. Lu, J. Li, X. Gao, S. Bell-Kligler, A. Shocher, and M. Irani, “Aim 2019 challenge on real-world image super-resolution: Methods and results,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3575–3583.
- [36] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, “Unfolding the alternating optimization for blind super resolution,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

References

- [37] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 416–423 vol.2.
- [38] S. McHugh, *Understanding Photography: Master Your Digital Camera and Capture That Perfect Photo*. No Starch Press, 2018. [Online]. Available: <https://books.google.dk/books?id=TVv6DwAAQBAJ>
- [39] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 945–952. [Online]. Available: <https://doi.org/10.1109/ICCV.2013.121>
- [40] C. Mou, Y. Wu, X. Wang, C. Dong, J. Zhang, and Y. Shan, "Metric learning based interactive modulation for real-world super-resolution," in *European Conference on Computer Vision (ECCV)*.
- [41] C. Qu, D. Luo, E. Monari, T. Schuchert, and J. Beyerer, "Capturing ground truth super-resolution data," in *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*. IEEE, 2016, pp. 2812–2816. [Online]. Available: <https://doi.org/10.1109/ICIP.2016.7532872>
- [42] M. S. Rad, B. Bozorgtabar, C. Musat, U. Marti, M. Basler, H. K. Ekenel, and J. Thiran, "Benefiting from multitask learning to improve single image super-resolution," *Neurocomputing*, 2020.
- [43] R. E. Rivadeneira, "Thermal image super-resolution challenge - PBVS 2020," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 432–439.
- [44] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [45] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang *et al.*, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [46] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 3773–3782. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00383>
- [47] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," in *CVPR*, 2021.
- [48] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *International Conference on Computer Vision Workshops (ICCVW)*.
- [49] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.

References

- [50] Z. Wang, J. Chen, and S. Hoi, “Deep learning for image super-resolution: A survey,” *TPAMI*, 2020.
- [51] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, “Component divide-and-conquer for real-world image super-resolution,” in *Proceedings of the European Conference on Computer Vision*, 2020.
- [52] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, “Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition,” in *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, June 2011, pp. 81–88.
- [53] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, “Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 701–710. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018_workshops/w13/html/Yuan_Unsupervised_Image_Super-Resolution_CVPR_2018_paper.html
- [54] Y. A. A. F. T. O. T. Y. Yusuke Matsui, Kota Ito and K. Aizawa, “Sketch-based manga retrieval using manga109 dataset,” in *Multimedia Tools and Applications*, 2017, p. 76(20).
- [55] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in *Curves and Surfaces*, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 711–730.
- [56] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, “Designing a practical degradation model for deep blind image super-resolution,” in *IEEE International Conference on Computer Vision*, 2021, pp. 4791–4800.
- [57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 586–595.
- [58] X. Zhang, Q. Chen, R. Ng, and V. Koltun, “Zoom to learn, learn to zoom,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 3762–3770. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Zoom_to_Learn_Learn_to_Zoom_CVPR_2019_paper.html
- [59] Y. Zhou, C. Lin, D. Luo, Y. Liu, Y. Tai, C. Wang, and M. Chen, “Joint learning content and degradation aware feature for blind super-resolution,” in *MM ’22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, Eds. ACM, 2022, pp. 2606–2616. [Online]. Available: <https://doi.org/10.1145/3503161.3547907>

References

Chapter 3

Joint Low-Light Image Enhancement and Super-Resolution

1 Introduction

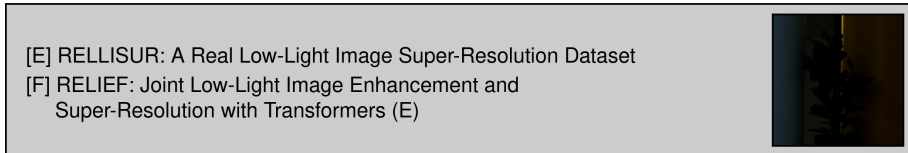


Fig. 3.1: An overview of the publications establishing the foundation for the findings presented in this chapter.

While chapter 2 presented insights and solutions for the problem of real-world image super-resolution, this chapter delves into an even more challenging image-processing problem where the visibility of the images is not only limited by low-resolution and various artifacts but also by severe under-exposure. We refer to such images as Low-Light Low-Resolution (LLLR) images.

Consequently, a substantial amount of information must be recovered to restore a LLLR image into its Normal-Light High-Resolution (NLHR) counterpart. Part of this is attributed to the significant loss of dynamic range in low-light images compared to normal-light images, as illustrated in Figure 3.3, which depicts histograms for both image groups.

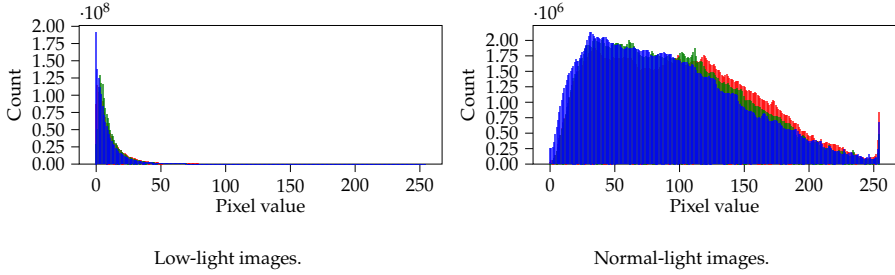


Fig. 3.3: The average RGB histograms of low-light and normal-light images in the Real Low-Light Image Super-Resolution (RELLISUR) dataset [2]. The horizontal axis represents the pixel value and the vertical axis the number of occurrences. Text and plots adapted from [2], Paper E.

As seen, the 8-bit Low-Light (LL) images contain little to no pixels with values above 50, whereas the values in Normal-Light (NL) images are more evenly distributed. Additionally, similar to the traditional SR setting, there is a notable loss of high-frequency information in the LLLR images. This is visualized in Figure 3.2 which shows the disparity in frequency content distribution between real paired LR and HR images from the SVSR dataset, obtained by the nested application of the Discrete Cosine Transform (DCT). Furthermore, due to the low photon count and low SNR, LL images often exhibit pronounced color distortion, strong noise, and low contrast. Consequently, the reconstruction process from LLLR to NLHR needs to address these challenges by filling the gaps in dynamic range and frequency content, while simultaneously suppressing the artifacts introduced during the imaging process.

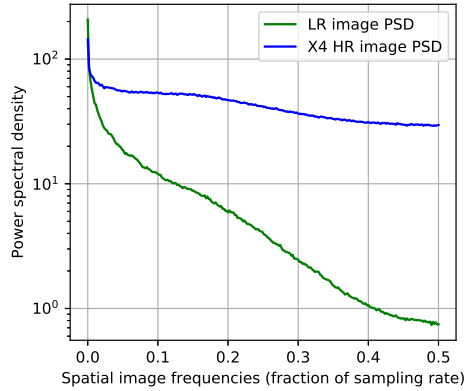


Fig. 3.2: Average power spectral density plot of LR and HR image pairs from the SVSR dataset [1], Paper F.

Yet, as current SR methods are trained on well-illuminated images, they cannot reconstruct LR images captured in low-light conditions. On the contrary, the correction of exposure levels in low-light or under-exposed images constitutes an active research domain known as low-light enhancement [15], but these methods do not address the resolution problem. Thus, despite their common occurrence, enhancing images degraded by both low-light and low-resolution is an overlooked problem in the literature with no straightforward solution. This can be partly attributed to the complexity of the problem, as both Low-Light Enhancement (LLE) and SR are ill-posed problems, and due



(a) Low-light image

(b) Normal-light image

Histogram equalization of (a)

Fig. 3.4: Illustration of histogram equalization of a low-light image. As seen, this process amplifies the noise and color distortion hidden in the dark. Furthermore, the enhanced image lacks contrast compared to the normal-light image. Images from [2], Paper E.

to the lack of real-world datasets for training and evaluation. However, one can imagine different direct ways to address this problem, but these are often followed by their own shortcomings. As an example, naively increasing the exposure of the image by histogram equalization also enhance the noise and other artifacts, while also risking over-saturation of bright areas as visualized in Figure 3.4. Another obvious solution to the problem would be to combine existing LLE and SR methods and sequentially process the LLLR image. However, as we prove in Paper E, this approach yields unsatisfactory results since errors in the LLE process might get amplified by the SR process, and crucial information for the SR process risk being discarded in the initial step. Consequently, joint processing emerges as a more promising alternative, aligning with findings in other areas of image processing research [14, 17, 27, 28]. However, this approach has not been widely studied in the body of literature which we review in the following section.

2 State-of-the-Art

This section builds upon the related work on SR and RWSR presented in Section 3.1 and Section 2, respectively. However, as the research on simultaneous LLE and SR is currently scarce, the following section primarily provides an overview of SoTA methods for LLE.

In recent years, the research field of LLE has witnessed a significant interest in the development of methods to correct the exposure of images captured in low-light conditions. Early approaches include Histogram equalization [6, 21] and gamma correction methods [12, 23], which are simple methods to improve the contrast of low-light images, typically resulting in reconstructions perceptually inconsistent with real NL images. Later approaches relied on

Retinex theory, which assumes that a color image can be decomposed into reflectance and illumination components [8, 10]. However, as these methods assume that the LL images are artifact-free, the strong levels of noise and color distortion often hidden in the dark of LL images are left untreated leading to unsatisfactory reconstruction results.

Recently, the field has started moving towards deep-learning-based approaches. In studies such as [18, 24, 26], Retinex theory is coupled with deep-learning to learn the estimation of the image illumination map. The use of multi-scale learning and attention for LLE has been explored in [25]. In [28], a network for joint LLE and deblurring is proposed. To avoid the overfitting issues with DNNs, [13] proposed an unsupervised GAN method that learns to process LL images to look similar to NL images by the use of adversarial losses. A zero-shot approach, that does not rely on paired or unpaired training data is proposed in [9]. However, neither of these approaches is capable of increasing the resolution of the LL image.

On the contrary, our RELISUR proposed in Paper E has recently facilitated the work of Cheng *et al.* [5] who proposed a U-net based SR model coupled with a module for estimation of the light distribution to guide the reconstruction process of LLLR images. However, the reconstruction performance for the joint LLE and $\times 4$ SR task on the RELISUR is 1.496dB PSNR lower than the concurrently developed method proposed by Aakerberg *et al.* [3]. Most recently, a pipeline with separate light and dark feature back-projection for learning their mutual dependencies for joint LLE and SR was proposed in [19]. However, the method is only developed for synthetic images and is therefore not applicable to reconstruction of real-world LLLR images.

3 Scientific Contributions

This section covers Paper E and Paper F which have both been published at reputable computer vision and machine-learning-focused research outlets. The papers share a common objective of bridging the gap between LLE and SR, with a particular emphasis on real-world scenarios, akin to the research presented in 3.

To facilitate research on joint LLE and SR, we initiated our work by collecting a dataset costing of real LLLR and NLHR image pairs, as described in Paper E. While synthetic LLLR images can be generated by downsampling existing low-light image datasets, they do not present the same level of reconstruction challenges as real-world images, which are crucial for advancing the field. However, constructing a dataset of the real image pairs poses a non-trivial task as such images are difficult to obtain. In Paper E, we approached this challenge by collecting real image pairs with different resolutions by adjusting the focal length of a zoom lens. This approach is feasible as the size of

3. Scientific Contributions

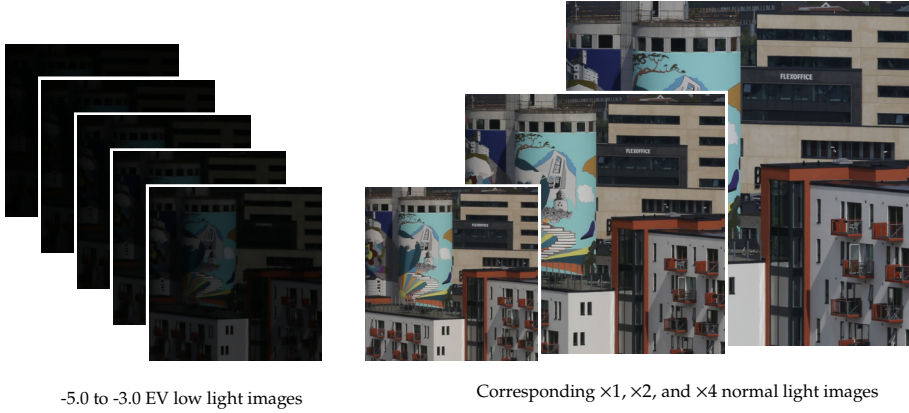


Fig. 3.5: "Example of a sequence of aligned images with different exposure (left) and scale levels (right) from the RELLISUR dataset." [2]. Images from [2], Paper E.

objects projected on the image sensor changes approximately linearly with the focal length [11]. To mimic capturing images in a low-light scenario, where there is insufficient light to accurately capture detail and color information, we reduced the exposure time below the value suggested by the camera's auto exposure. This approach simultaneously reduced the SNR compared to properly exposed images, and consequently introduced stronger levels of color distortion and noise to the images. An alternative approach could have been to collect the images at night time, but aside from being impractical, it would have made it challenging to obtain well-illuminated and colorful NLHR ground-truth images.

Another challenging aspect of the dataset collection process was to ensure a wide diversity of images in terms of color, texture, and patterns while minimizing the presence of moving objects to avoid misalignment between frames. This necessitated a significant effort in scouting suitable locations and relocating equipment. Moreover, to comply with GDPR guidelines, we took precautions to avoid capturing images that could enable the identification of individuals, such as avoiding capturing persons, license plates, or other personally identifiable information, further adding complexity to the data collection process. To ensure optimal quality and alignment of the collected image pairs, we designed a preprocessing pipeline that involved both manual screening and automated processing. In total, our dataset comprises 850 distinct scenes, each captured at three different scale levels ($\times 1$, $\times 2$, and $\times 4$) and five different under-exposure levels in addition to the auto-exposure level, all aligned towards the NLHR GT image. An example of a scene from our dataset can be seen in Figure 3.5.

Finally, the acquired dataset was divided into train, validation, and test

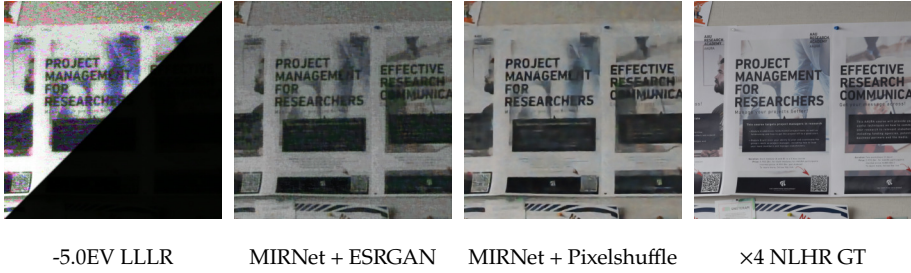


Fig. 3.6: A visual comparison of the reconstruction of a real LLLR image by sequential and joint methods. As seen joint processing (MIRNet [25] + Pixelshuffle [20]) results in a reconstruction closer to the ground truth in terms of details and artifacts. Images from [2], Paper E.

splits, adhering to a 85%/5%/10% distribution. We utilized this dataset for two primary purposes: firstly, to assess its effectiveness in training DNN methods for joint LLE and SR; and secondly, to establish benchmarks for existing approaches. Our experimental findings revealed that it is feasible to train existing image processing models to learn the direct mapping from LLLR to NLHR using the RELISUR training data. Furthermore, we found that such a joint approach leads to both lower distortion and better perceptual quality compared to sequential processing using dedicated LLE and SR methods, even when the latter methods were also trained on data from the RELISUR dataset. A visual comparison of images reconstructed by both methods can be seen in Figure 3.6. Here we compare sequential processing with MIRNet [25] and ESRGAN [22], both SoTA methods for LLE and SR respectively, to joint processing by MIRNet alone to which we added a $\times 4$ PixelShuffle [20] up-sampling module, following most existing SR works. The obtained results reveal that sequential processing yields the least visually satisfactory reconstructions, potentially due to the amplification of artifacts, introduced during the LLE process, by the SR network. On average, across the entire RELISUR, joint processing outperforms sequential processing by 0.81dB PSNR using the aforementioned method. However, it is important to note that the performance of these methods is likely sub-optimal since neither of them is specifically designed for this particular task.

In Paper F, we subsequently focused on developing one of the world’s first dedicated methods for joint LLE and SR of real LLLR images, by building upon the insights and outcomes presented in Paper E. Our approach was motivated by the observation that methods capable of capturing contextual information exhibited superior performance for this specific task. We hypothesized that this phenomenon could be attributed to the utilization of a larger pixel neighborhood, which facilitates better determination of the enhancement level for each pixel by incorporating cues from the possible upper and lower intensity boundaries. To this end, we based our solution on Transformer networks,

3. Scientific Contributions

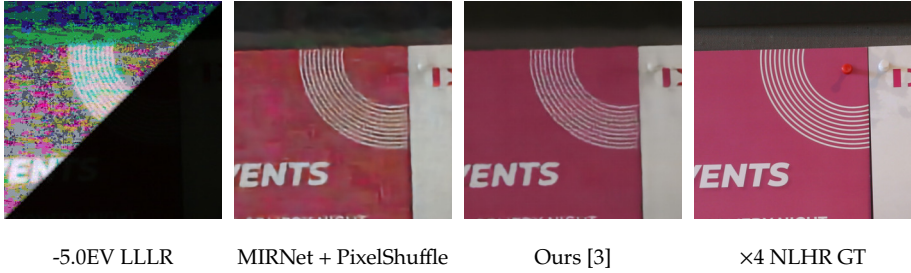


Fig. 3.7: A visual comparison of the reconstruction of a real LLLR image by our proposed method, RELIEF [3], and the second best-performing method in the experiments (MIRNet [25] + Pixelshuffle [20]). In comparison, our method produces more faithfully results with fewer artifacts. Images from [3], Paper E.

Table 3.1: “Quantitative comparison of state-of-the-art methods for joint LLE and $\times 4$ SR on the RELISUR and SICE datasets. Our Resolution and Light Enhancement Transformer (RELIEF) sets state-of-the-art results on both datasets.” [3]. Table adapted from [3], Paper E.

Method	RELLISUR [2]			SICE [4]		
	PSNR \uparrow	SSIM \uparrow	DISTS [7] \downarrow	PSNR \uparrow	SSIM \uparrow	DISTS [7] \downarrow
MIRNet [25] + Pixelshuffle [20]	21.04	0.7619	0.1609	18.02	0.6760	0.2749
ESRGAN [22]	17.49	0.6724	0.1518	16.44	0.6271	0.2611
SwinIR [16]	18.99	0.7478	0.1705	17.66	0.6867	0.2753
RELIEF	21.32	0.7686	0.1364	18.80	0.6980	0.2606

which are widely known for their ability to capture long-range dependencies. With the proposed model, we empirically validated this hypothesis by experimenting with different training patch sizes, revealing a strong correlation between larger patch sizes and higher reconstruction accuracy. Notably, an improvement of 2.24dB PSNR is obtained by using a training patch size of 384×384 pixels compared to 64×64 pixels. Our proposed method demonstrated SoTA performance on both real images from the RELISUR dataset and synthetic images from the SICE [4] dataset as seen in Table 3.1. Moreover, as illustrated in Figure 3.7, our method produced more visually pleasing reconstructions with better details and fewer artifacts, compared to the second best performing method which combined MIRNet [25] with Pixelshuffle [20].

In summary, the collective findings from our research in Paper E and Paper F represent the pioneering efforts in SR of real low-light low-resolution images. The key contributions of our work can be outlined as follows:

- To facilitate research in SR of real LLLR images we presented RELISUR, the first large-scale dataset of paired LLLR and NLHR images (Paper E).
- Through our analysis of the proposed RELISUR dataset, we provided valuable insights into the performance of existing approaches. Our find-

ings demonstrated the superiority of joint processing methods compared to sequential processing methods for LLLR image enhancement (Paper E).

- Our research highlighted the significance of utilizing global information in addressing the challenges associated with joint LLE and SR (Paper E, Paper F).
- We proposed one of the first dedicated methods for joint LLE and SR of real LLLR images, leveraging a novel Transformer architecture. This method incorporates global information effectively, leading to enhanced results in joint LLE and SR (Paper F).
- We achieved SoTA reconstruction results for both real and synthetic LLLR images, which demonstrates the effectiveness and advancements made in the field (Paper E, Paper F).

These contributions have expanded the application of SR and have provided new insights that pave the way for future research in enhancing low-light low-resolution images.

References

- [1] A. Aakerberg, M. E. Helou, K. Nasrollahi, S. Ssstrunk, and T. M. and, "Pda-rwsr: Pixel-wise degradation adaptive real- world super-resolution," *Under review, ICCV 2023.*, 2023.
- [2] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, "RELLISUR: A real low-light image super-resolution dataset," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/7ef605fc8dba5425d6965fbd4c8fbc1f-Paper-round2.pdf>
- [3] —, "Relief: Joint low-light image enhancement and super-resolution with transformers," in *Scandinavian Conference on Image Analysis (SCIA)*, ser. Lecture Notes in Computer Science. Springer, 2022, pp. 157–173.
- [4] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.
- [5] D. Cheng, L. Chen, C. Lv, L. Guo, and Q. Kou, "Light-guided and cross-fusion u-net for anti-illumination image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8436–8449, 2022. [Online]. Available: <https://doi.org/10.1109/TCSVT.2022.3194169>
- [6] D. Coltuc, P. Bolon, and J. Chassery, "Exact histogram specification," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1143–1152, 2006. [Online]. Available: <https://doi.org/10.1109/TIP.2005.864170>

References

- [7] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *CoRR*, vol. abs/2004.07728, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07728>
- [8] X. Fu, Y. Liao, D. Zeng, Y. Huang, X. S. Zhang, and X. Ding, "A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4965–4977, 2015. [Online]. Available: <https://doi.org/10.1109/TIP.2015.2474701>
- [9] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 1777–1786. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00185>
- [10] X. Guo, Y. Li, and H. Ling, "LIME: low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, 2017. [Online]. Available: <https://doi.org/10.1109/TIP.2016.2639450>
- [11] E. Hecht, "Optics," *Addison-Wesley*, 1987.
- [12] S. Huang, F. Cheng, and Y. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1032–1041, 2013. [Online]. Available: <https://doi.org/10.1109/TIP.2012.2226047>
- [13] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021. [Online]. Available: <https://doi.org/10.1109/TIP.2021.3051462>
- [14] T. Klatzer, K. Hammernik, P. Knöbelreiter, and T. Pock, "Learning joint demosaicing and denoising based on sequential energy minimization," in *2016 IEEE International Conference on Computational Photography, ICCP 2016, Evanston, IL, USA, May 13-15, 2016*. IEEE Computer Society, 2016, pp. 1–11. [Online]. Available: <https://doi.org/10.1109/ICCPHOT.2016.7492871>
- [15] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9396–9416, 2021.
- [16] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *IEEE International Conference on Computer Vision Workshops*, 2021.
- [17] Z. Liang, D. Zhang, and J. Shao, "Jointly solving deblurring and super-resolution problems with dual supervised network," in *IEEE International Conference on Multimedia and Expo, ICME 2019, Shanghai, China, July 8-12, 2019*. IEEE, 2019, pp. 790–795. [Online]. Available: <https://doi.org/10.1109/ICME.2019.00141>
- [18] L. Ma, R. Liu, Y. Wang, X. Fan, and Z. Luo, "Low-light image enhancement via self-reinforced retinex projection model," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

References

- [19] M. T. Rasheed and D. Shi, "LSR: lightening super-resolution deep network for low-light image enhancement," *Neurocomputing*, vol. 505, pp. 263–275, 2022. [Online]. Available: <https://doi.org/10.1016/j.neucom.2022.07.058>
- [20] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 1874–1883. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.207>
- [21] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Trans. Image Process.*, vol. 9, no. 5, pp. 889–896, 2000. [Online]. Available: <https://doi.org/10.1109/83.841534>
- [22] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.
- [23] Z. Wang, Z. Liang, and C. Liu, "A real-time image processor with combining dynamic contrast ratio enhancement and inverse gamma correction for PDP," *Displays*, vol. 30, no. 3, pp. 133–139, 2009. [Online]. Available: <https://doi.org/10.1016/j.displa.2009.03.006>
- [24] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 155. [Online]. Available: <http://bmvc2018.org/contents/papers/0451.pdf>
- [25] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, M. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12370. Springer, 2020, pp. 492–511. [Online]. Available: https://doi.org/10.1007/978-3-030-58595-2_30
- [26] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. ACM, 2019, pp. 1632–1640. [Online]. Available: <https://doi.org/10.1145/3343031.3350926>
- [27] R. Zhou, M. E. Helou, D. Sage, T. Laroche, A. Seitz, and S. Süssstrunk, "W2S: microscopy data with joint denoising and super-resolution for widefield to SIM mapping," in *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, A. Bartoli and A. Fusiello, Eds., vol. 12535. Springer, 2020, pp. 474–491. [Online]. Available: https://doi.org/10.1007/978-3-030-66415-2_31
- [28] S. Zhou, C. Li, and C. C. Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, ser. Lecture

References

Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13666. Springer, 2022, pp. 573–589. [Online]. Available: https://doi.org/10.1007/978-3-031-20068-7_33

References

Chapter 4

Improving Downstream Vision Tasks With Super-Resolution

1 Introduction

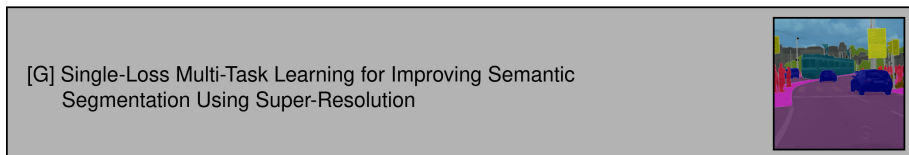


Fig. 4.1: An overview of the publications establishing the foundation for the findings presented in this chapter.

This chapter covers our work in relation to improving downstream computer vision tasks through the utilization of SR. The widespread adoption of deep-learning-based methods in computer vision, following the success of architectures like AlexNet [20], has led to remarkable advancements in a number of tasks such as object detection [11, 17, 22, 25, 34], image classification [14, 17, 20, 31, 32], semantic segmentation [7, 13, 23, 28, 39], face recognition [5, 29, 33], and human pose estimation [6, 35, 36], among others. Here, the deep-learning-based methods have surpassed the performance of traditional approaches by leveraging supervised learning on labeled datasets to automatically learn the otherwise hand-crafted computer vision pipeline, namely: feature extraction, feature selection, and decision making [4]. Consequently, this has eliminated the need for manual feature engineering and

enabled the model to learn and extract the most relevant features from the input images for the given task.

However, as discussed in Chapter 1, the quality of the input images is not always optimal in practice, which can lead to wrong predictions or complete failure of DNN-based classifiers. The phenomenon has been investigated in previous studies [10, 21], where it was demonstrated that distorting the input images with small amounts of blur, noise, and compression artifacts was enough to cause misclassifications in DNN-based classifiers. Similar findings have been reported for DNN-based face recognition in [19]. Unfortunately, computer vision research predominantly focuses on training and testing on high-quality image datasets, largely overlooking the challenge of making predictions from real-world images. Although SoTA DNN-based image restoration methods, like SR, can be used to improve the image quality, they often perform inadequately and unreliably on degraded real-world images as discussed previously in this thesis. Hence, in this chapter, we delve into Paper G, which explores if SR can be used to enhance the performance of downstream computer vision tasks when applied to real-world images. First, we provide an overview of the existing studies that investigate improving the performance of computer vision tasks through leveraging image processing techniques, such as SR.

2 State-of-the-Art

SR techniques have traditionally been employed to enhance image quality for improved visual perception by humans. However, there has been a recent surge of interest within the SR research community to explore its potential for enhancing the performance of downstream computer vision tasks.

Notably, a study [30] demonstrated 36% improvement in object detection performance by prior SR of the input images. Similar positive effects were observed in [38], when using SR in combination with classification of blurred small objects, and [27], where SR aided optical character recognition. Moreover, experiments conducted on face recognition tasks exhibited improved performance when using with SR as a pre-processing step in [3, 15, 16]. However, a comprehensive study assessing the effect of SR on four popular computer vision tasks concluded that while SR is beneficial when the input image resolution is low, it still falls short compared to using the original HR images [9].

To optimize the image enhancement for specific downstream tasks, researchers have proposed task-driven SR frameworks [12], integrating the objective of the target task into the optimization process of the SR model. Recently, [37] proposed a multi-task learning framework combining SR and semantic segmentation. In the context of small object detection, an approach combining a GAN SR network and a classifier has been proposed to enhance

accuracy [2].

Despite these advancements, most existing approaches require paired LR and HR training images, limiting their practical applicability. To address this limitation, a CycleGAN SR strategy was employed in [18] for small object detection. However, since little high-frequency information is added in the process, as a result of the weak supervision, the approach may be less suitable for improving more fine-grained tasks like semantic segmentation.

3 Scientific Contributions

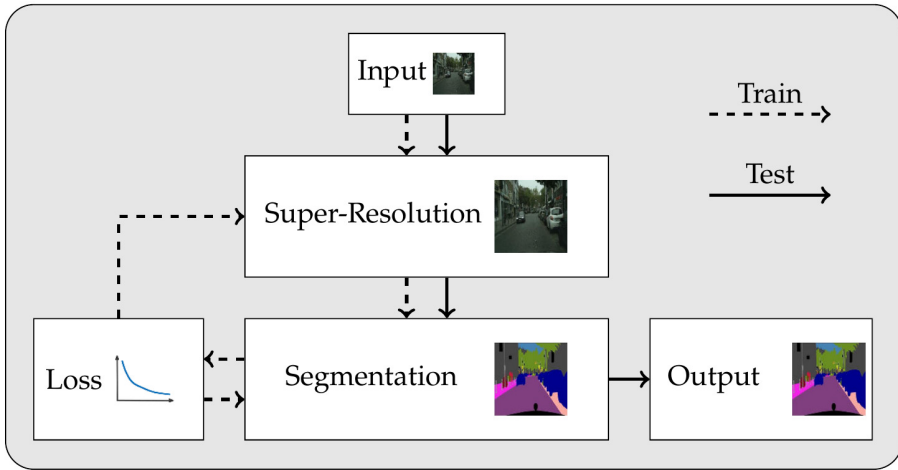


Fig. 4.2: “Our proposed framework, Multi-Task Semantic Segmentation and Super-Resolution (MT-SSSR). Dashed and full lines represent training and testing phases, respectively. The SR model learns to upsample and enhance the input image based on the segmentation task loss. The segmentation model uses the same loss to improve the accuracy of its prediction.” [1]. Figure from [1], Paper G.

In Paper G, we investigated the use of SR in combination with semantic segmentation. Semantic segmentation is a high-level computer vision task that aims to classify each pixel in an image to a specific category and hereby segmenting different objects and regions according to their semantic meaning. As such, semantic segmentation is a much more complex task compared to simple object detection, where the focus is on identifying specific objects and indicating their position with a rectangular bounding box.

As such, accurate pixel-wise labeling in semantic segmentation is dependent on high-resolution images of good quality. Hence, in Paper G we hypothesized that the semantic segmentation task can benefit from improved image quality obtained by SR. However, since most semantic segmentation benchmarking datasets typically consists of unprocessed images straight from

the camera, the traditional SR setting of training the model on synthetic LR images created with bicubic downsampling will likely be ineffective due to the domain gap.

Furthermore, since the popular semantic segmentation benchmarking datasets do not contain HR reference images to enable learning the domain-specific LR/HR mapping, we developed a method that does not require such image pairs. Specifically, we proposed a novel framework where the SR model and semantic segmentation model are optimized jointly using only the loss associated with the segmentation task. Our approach hereby differs from traditional multi-task learning as it does not require additional labeled data besides labels for the segmentation task.

As seen in Figure 4.2, our framework consists of a SR model followed by a semantic segmentation model, where the sole task of the former is to produce images that are optimal for the semantic segmentation task. Consequently, we do not use any perceptual or pixel-wise reconstruction losses for the SR model. However, a limitation of our approach is the relatively high memory requirements during training, which we addressed using mixed-precision data types.

It is widely known that the input image size affects the performance of deep-learning models [26]. As such, we compare our approach to bicubic interpolation and SR as a pre-processing step as summarized in Table 4.1. As seen, bicubic interpolation, although not providing additional information, demonstrated improved semantic segmentation performance when the image resolution was doubled. However, when upsampling the image by a factor of four using bicubic interpolation,

Method	Scale Factor	Val. (%)
HRNet	Native	69.9
HRNet	$\times 2$ Bicubic	70.9
HRNet	$\times 4$ Bicubic	67.1
HRNet	$\times 2$ SR _{ST}	71.2
MT-SSSR (ours)	$\times 2$ SR _{MT}	74.1
MT-SSSR (ours)	$\times 4$ SR _{MT}	76.3

Table 4.1: Comparison of the semantic segmentation performance on the IDD-Lite dataset [24] when utilizing bicubic interpolation, single-task (_{ST}) and multi-task (_{MT}) super-resolution to increase the input resolution. Accuracy is reported in mean Intersection over Union (mIoU). Table adapted from [1], Paper G.

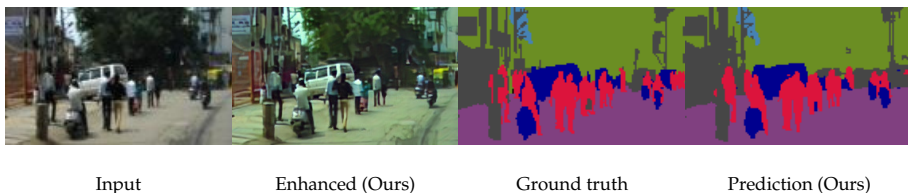


Fig. 4.3: Visualization of the image enhanced with our method, and the resulting segmentation results on IDD-Lite [24]. Images from [1], Paper G.

3. Scientific Contributions

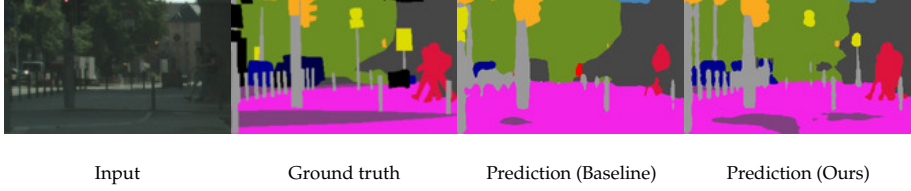


Fig. 4.4: Visual comparison of the semantic segmentation masks generated by the baseline HRNet [36] and our approach. Images from [1].



Fig. 4.5: Example images from the CityScapes dataset [8] processed with different upsampling methods.

the performance fell below the baseline, which we attribute to the loss of sharp edges resulting from the interpolation process. On the contrary, our proposed method consistently outperformed the baseline for both the $\times 2$ and $\times 4$ upscaling factors, yielding a significant 6.9% improvement in mIoU.

One noteworthy observation regarding the performance improvement achieved by our proposed solution is illustrated in Figure 4.3, where we have extracted the enhanced image during testing. As seen, in the upper left portion of the images, the ground truth label incorrectly classifies the center part of the

triangular construction on the pole top as a roadside object instead of the sky.

However, through the image enhancements made by our approach, the model correctly classified these pixels as sky. Furthermore, we also observed significant improvements in the segmentation of small objects in general, as visualized in Figure 4.4. Specifically, it can be seen that the segmentation of poles aligns more accurately with the GT labels, indicating improved performance.

Furthermore, we observe that our jointly trained model produces sharper and more contrast-rich and detailed images, compared to both bicubic interpolation and pre-processing with SR, as visualized in Figure 4.5. In particular, it's interesting that the sharpness and details are improved even though the semantic labels used to guide the SR learning reside in the LR domain.

In summary, our research contributions on the topic of improving computer-vision tasks with image-processing techniques can be outlined as follows:

- We introduced a novel framework that leverages the loss associated with the semantic segmentation task to jointly optimize a combined SR and semantic segmentation model. This framework enhances segmentation accuracy by enhancing the input images to become optimal for the segmentation task (Paper G).
- Since our proposed method only requires GT semantic labels, our method is applicable to practical scenarios where no LR/HR image pairs are available (Paper G).
- We achieved SoTA results in semantic segmentation on the challenging CityScapes and IDD-Lite datasets, surpassing the baseline by 4.2% and 2.2%, respectively (Paper G).

References

- [1] A. Aakerberg, A. S. Johansen, K. Nasrollahi, and T. B. Moeslund, "Single-loss multi-task learning for improving semantic segmentation using super-resolution," in *Computer Analysis of Images and Patterns - 19th International Conference, CAIP 2021, Virtual Event, September 28-30, 2021, Proceedings, Part II*, ser. Lecture Notes in Computer Science, N. Tsapatsoulis, A. Panayides, T. Theoharides, A. Lanitis, C. S. Pattichis, and M. Vento, Eds., vol. 13053. Springer, 2021, pp. 403–411. [Online]. Available: https://doi.org/10.1007/978-3-030-89131-2_37
- [2] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: small object detection via multi-task generative adversarial network," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217. Springer, 2018, pp. 210–226. [Online]. Available: https://doi.org/10.1007/978-3-030-01261-8_13

References

- [3] E. Bilgazyev, B. A. Efraty, S. K. Shah, and I. A. Kakadiaris, "Sparse representation-based super resolution for face recognition at a distance," in *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*, J. Hoey, S. J. McKenna, and E. Trucco, Eds. BMVA Press, 2011, pp. 1–11. [Online]. Available: <https://doi.org/10.5244/C.25.52>
- [4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
- [5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*. IEEE Computer Society, 2018, pp. 67–74. [Online]. Available: <https://doi.org/10.1109/FG.2018.00020>
- [6] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021. [Online]. Available: <https://doi.org/10.1109/TPAMI.2019.2929257>
- [7] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211. Springer, 2018, pp. 833–851. [Online]. Available: https://doi.org/10.1007/978-3-030-01234-2_49
- [8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [9] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?" in *WACV*, 2016.
- [10] S. F. Dodge and L. J. Karam, "Understanding how image quality affects deep neural networks," in *Eighth International Conference on Quality of Multimedia Experience, QoMEX 2016, Lisbon, Portugal, June 6-8, 2016*. IEEE, 2016, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/QoMEX.2016.7498955>
- [11] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013.
- [12] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *Neural Information Processing - 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8-12, 2021, Proceedings, Part V*, ser. Communications in Computer and Information Science, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds., vol. 1516. Springer, 2021, pp. 387–395. [Online]. Available: https://doi.org/10.1007/978-3-030-92307-5_45
- [13] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October*

References

- 22-29, 2017. IEEE Computer Society, 2017, pp. 2980–2988. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.322>
- [14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.
- [15] P. H. Hennings-Yeomans, S. Baker, and B. V. K. V. Kumar, “Simultaneous super-resolution and feature extraction for recognition of low-resolution faces,” in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 24-26 June 2008, Anchorage, Alaska, USA. IEEE Computer Society, 2008. [Online]. Available: <https://doi.org/10.1109/CVPR.2008.4587810>
- [16] P. H. Hennings-Yeomans, B. V. K. V. Kumar, and S. Baker, “Robust low-resolution face identification and verification using high-resolution features,” in *Proceedings of the International Conference on Image Processing, ICIP 2009, 7-10 November 2009, Cairo, Egypt*. IEEE, 2009, pp. 33–36. [Online]. Available: <https://doi.org/10.1109/ICIP.2009.5413920>
- [17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: <http://arxiv.org/abs/1704.04861>
- [18] H. Ji, Z. Gao, X. Liu, Y. Zhang, and T. Mei, “Small object detection leveraging on simultaneous super-resolution,” in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 2020, pp. 803–810. [Online]. Available: <https://doi.org/10.1109/ICPR48806.2021.9413058>
- [19] S. Karahan, M. K. Yildirim, K. Kirta, F. S. Rende, G. Butun, and H. K. Ekenel, “How image degradations affect deep cnn-based face recognition?” in *2016 International Conference of the Biometrics Special Interest Group, BIOSIG 2016, Darmstadt, Germany, September 21-23, 2016*, ser. LNI, A. Brömmel, C. Busch, C. Rathgeb, and A. Uhl, Eds., vol. P-260. GI / IEEE, 2016, pp. 313–320. [Online]. Available: <https://doi.org/10.1109/BIOSIG.2016.7736924>
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
- [21] H. Liu, H. Li, X. Wang, H. Li, M. Ou, L. Hao, Y. Hu, and J. Liu, “Understanding how fundus image quality degradation affects cnn-based diagnosis,” in *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2022, Glasgow, Scotland, United Kingdom, July 11-15, 2022*. IEEE, 2022, pp. 438–442. [Online]. Available: <https://doi.org/10.1109/EMBC48229.2022.9871507>
- [22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, “SSD: single shot multibox detector,” in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9905. Springer, 2016, pp. 21–37. [Online]. Available: https://doi.org/10.1007/978-3-319-46448-0_2

References

- [23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3431–3440. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298965>
- [24] A. Mishra, S. Kumar, T. Kalluri, G. Varma, A. Subramaian, M. Chandraker, and C. V. Jawahar, "Semantic segmentation datasets for resource constrained training," in *Computer Vision, Pattern Recognition, Image Processing, and Graphics*, R. V. Babu, M. Prasanna, and V. P. Namboodiri, Eds. Springer Singapore, 2020.
- [25] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 779–788. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.91>
- [26] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: <http://arxiv.org/abs/1804.02767>
- [27] V. Robert and H. Talbot, "Does super-resolution improve OCR performance in the real world? A case study on images of receipts," in *ICIP*, 2020.
- [28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28
- [29] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 815–823. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298682>
- [30] J. Shermeyer and A. V. Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1432–1441. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/EarthVision/Shermeyer_The_Effects_of_Super-Resolution_on_Object_Detection_Performance_in_Satellite_CVPRW_2019_paper.html
- [31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
- [32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.
- [33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on*

References

- Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014.* IEEE Computer Society, 2014, pp. 1701–1708. [Online]. Available: <https://doi.org/10.1109/CVPR.2014.220>
- [34] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *ICML*, 2019, pp. 6105–6114.
- [35] A. Toshev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014.* IEEE Computer Society, 2014, pp. 1653–1660. [Online]. Available: <https://doi.org/10.1109/CVPR.2014.214>
- [36] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, “Deep high-resolution representation learning for visual recognition,” *TPAMI*, 2019.
- [37] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, “Dual super-resolution learning for semantic segmentation,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020.* IEEE, 2020, pp. 3773–3782. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00383>
- [38] T. Wang, W. Sun, H. Qi, and P. Ren, “Aerial image super resolution via wavelet multiscale convolutional neural networks,” *IEEE Geosci. Remote. Sens. Lett.*, vol. 15, no. 5, pp. 769–773, 2018. [Online]. Available: <https://doi.org/10.1109/LGRS.2018.2810893>
- [39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *CVPR*, 2017.

Chapter 5

Conclusion

This PhD thesis, conducted in collaboration with Milestone Systems A/S under the Milestone Research Programme at Aalborg University during the period from 2020 to 2023, focused on investigating novel methods for enhancing the visibility and quality of real-world images through super-resolution. Specifically, three main themes have been covered:

- Real-World Super-Resolution
- Joint Low-Light Image Enhancement and Super-Resolution
- Improving Downstream Vision Tasks With Super-Resolution

SR is an example of a technology that over the last decades has seen significant advancements in the academic literature, but has not yet been widely adopted in practical applications. One reason for this is the lack of attention on the domain gap between synthetic LR images, used in most research studies, and the more complex and sometimes heavily corrupted LR images encountered in real-world settings.

However, most recently there has been an increased focus in SR research on developing methods that can generalize to such real-world images. In this PhD thesis, we have continued this line of investigation by exploring even more challenging scenarios and proposing novel insights, methodologies, and datasets to advance the field.

Specifically, within our research on RWSR we have investigated SR of face images from real-world surveillance footage. Through this study, we found that existing methods failed to model the strong compression artifacts often found in such images. To this end, we proposed a method based on domain adaptation, which we also evaluated on thermal images, and observed improved performance in both cases. To overcome the lack of paired real-world LR and HR training image pairs, we explored the use of guidance by

semantic segmentation, which we found to be highly effective. Based on an observation that current SoTA RWSR methods struggled with images containing spatially variant degradations, we developed a method that estimates per-pixel degradations and adapts the SR reconstruction process accordingly. To enable quantitative evaluation of the reconstruction of such images, and evaluate the effectiveness of our approach, we collected the first dataset of paired real LR and HR images with varying degrees of spatially variant noise. Our proposed method demonstrated superior performance compared to all existing RWSR methods when evaluated on the challenging dataset.

In the area of Joint Low-Light Image Enhancement and Super-Resolution, we initiated our research by collecting the first large-scale dataset of paired LLLR and NLHR images. We used this dataset to establish baselines with existing methods and demonstrated that joint LLE and SR significantly outperformed sequential processing. Building upon these findings, we investigated the development of a dedicated method for joint LLE and SR and showed that improved performance can be obtained by leveraging the long-range modeling capabilities of Transformer networks.

Lastly, in our research on Improving Downstream Vision Tasks With Super-Resolution, we found that the semantic segmentation performance can be significantly improved by combining SR and semantic segmentation and optimizing both solely using the loss associated with the segmentation task.

These contributions collectively pushed the capabilities of SR, bringing it closer to reliable, robust, and high-quality reconstruction performance. However, despite these significant advancements, RWSR remains a challenging and unsolved problem for severely degraded images, particularly those with multiple types of degradations such as blur, noise, low-light, and low-resolution. While we have proposed solutions that improve generalization to real-world images, there is still a long way to go in developing a "one-fits-all" SR model, capable of reliably enhancing all types of real-world images. Since it is infeasible to collect real-world training data to cover this wide distribution, a promising future research direction could be to explore the use of self-supervised learning which has proven to be highly effective in image generation and natural language processing.

Furthermore, ethical concerns and trustworthiness are largely overlooked issues in the research community regarding generative models like SR. These unsolved challenges likely contribute to a limited adaptation of SR methods in safety-critical and ethically sensitive applications such as medical practice and forensics. Therefore, these issues also give opportunities for future work.

In conclusion, this thesis has made significant contributions to the field of image super-resolution for real-world applications. It enhances our understanding of the challenges involved by providing valuable new insights and

addresses them through the introduction of novel solutions and benchmark datasets, thereby advancing the overall performance. Collectively, this has made us able to reveal more details in low-quality images than ever before.

Chapter 5. Conclusion

Part II

Papers

Paper A

Real-World Super-Resolution of Face-Images From Surveillance Cameras

Andreas Aakerberg, Kamal Nasrollahi, and Thomas B Moeslund

The paper has been published in
IET Image Processing, Volume 16, Issue 2, pp. 442-452, 2022.

© 2022 by the authors.

The layout has been revised.

Abstract

Most existing face image Super-Resolution (SR) methods assume that the Low-Resolution (LR) images were artificially downsampled from High-Resolution (HR) images with bicubic interpolation. This operation changes the natural image characteristics and reduces noise. Hence, SR methods trained on such data most often fail to produce good results when applied to real LR images. To solve this problem, we propose a novel framework for generation of realistic LR/HR training pairs. Our framework estimates realistic blur kernels, noise distributions, and JPEG compression artifacts to generate LR images with similar image characteristics as the ones in the source domain. This allows us to train a SR model using high quality face images as Ground-Truth (GT). For better perceptual quality we use a Generative Adversarial Network (GAN) based SR model where we have exchanged the commonly used VGG-loss [1] with LPIPS-loss [2]. Experimental results on both real and artificially corrupted face images show that our method results in more detailed reconstructions with less noise compared to existing State-of-The-Art (SoTA) methods. In addition, we show that the traditional non-reference Image Quality Assessment (IQA) methods fail to capture this improvement and demonstrate that the more recent NIMA metric [3] correlates better with human perception via Mean Opinion Rank (MOR).

1 Introduction

Face Super-Resolution (SR) is a special case of SR which aims to restore High-Resolution (HR) face images from their Low-Resolution (LR) counterparts. This is useful in many different applications such as video surveillance and face enhancement. Current State-of-The-Art (SoTA) face SR methods based on Convolutional Neural Networks (CNNs) are able to reconstruct images with photo-realistic appearance from artificially generated LR images. However,

these methods often assume that the LR images were downsampled with bicubic interpolation, and therefore fail to produce good results when applied to real-world LR images. This is mostly due to the fact that the downsampling operation with bicubic downscaling changes the natural image characteristics and reduces the amount of artifacts. Hence, when using algorithms trained with supervised learning on such artificial LR/HR image pairs, the reconstructed images usually contains strong artifacts due to the domain gap.



Fig. A.1: $\times 4$ SR of a real low-quality face image (100×128 pixels) from the Chokepoint DB [5]. Our method enhances details and removes noise while the ESRGAN [4] amplifies the corruptions.

This paper is about SR of real low-resolution, noisy, and corrupted images, also known as Real-World Super-Resolution (RWSR). We apply our proposed method to face images, but the method is also applicable to other image domains. To create a SR model that is robust against the corruptions found in real images, we create a degradation framework that can produce LR images that have the same image characteristic as the images that we want to super-resolve, *i.e.* the source domain images. By creating LR images from clean high-quality images, *i.e.* the target domain, allows us to train a SR model that learns to super-resolve images with similar characteristics. This approach is inspired by the work of Ji *et al.* [6] who propose to perform RWSR via kernel estimation and noise injection. However, we observe that their framework for image degradation is not ideal for SR of LR face images from surveillance cameras, as these are often also corrupted by compression artifacts. Hence, we extend the degradation framework from [6] to include JPEG compression artifacts. We use the ESRGAN [4] model, which is one of the SoTA models for perceptual quality, as our backbone SR model. However, we find that the combination of loss functions for the ESRGAN is not ideal for optimal perceptual quality. To this end, we exchange the VGG-loss [1] with PatchGAN [7] loss for the discriminator similar to [6]. Inspired by Jo *et al.* [8], we additionally exchange the VGG-loss [1] with Learned Perceptual Image Patch Similarity (LPIPS) loss [2] for better perceptual quality. Different from existing models for face SR [9–11], we do not restrict our model to only work for face images of fixed input sizes, which makes our model more useful in practice. To the best of our knowledge, we are the first to propose a method for SR of real LR face images of arbitrary sizes.

We evaluate our method on two different face image datasets and one dataset of general images. To enable comparison of the SR performance against Ground-Truth (GT) reference images, we artificially corrupt high-quality images from Flickr-Faces-HQ Dataset (FFHQ) [12] and DIV2k [13] and report quantitative results using conventional Image Quality Assessment (IQA) methods and the most recent methods for assessment of the perceptual quality. For evaluation on real LR face image from surveillance cameras we use the Chokepoint DB [5]. In this case, as no GT image is available, we report the results using Mean Opinion Rank (MOR) and several non-reference based IQA methods. In both cases we show the effectiveness of our method via quantitative and qualitative evaluations. Furthermore, our evaluations show that most existing non-reference based IQA methods correlate poorly with human perception, while the recent Neural Image Assessment (NIMA) [3] metric provides a good correlation with human judgment as proven with MOR.

In summary, our contributions are:

- A novel framework for generation of LR/HR training pairs, where we introduce realistic image compression artifacts, and improve upon the

2. Related Work

noise collection method from [14], for noise injection, by adding additional constraints.

- Improving the ESRGAN [4] SR model with a novel combination of loss functions including local patch-wise adversarial loss [7], perceptual loss calibrated towards human judgement [2], and pixel-wise loss for better visual quality.
- A comprehensive evaluation on real LR face images from the Chokepoint DB [5] and artificially corrupted face images from the FFHQ DB [12]. Furthermore, we also evaluate on general images from the DIV2K dataset [13], to demonstrate that our method is also applicable to other image domains.
- Quantitatively, we evaluate our method using the most popular non-reference based IQA methods, and find only the recent NIMA [3] metric to correlate with human judgment via MOR.
- Our work highlights the importance of accurate modeling of the degradation parameters for practical applications of GAN-based SR.

2 Related Work

Recent advancements within deep-learning have proven very successful for use within super-resolution, and models of this type often achieve SoTA results. The first deep-learning based method for super-resolution was proposed by Dong *et al.* [15] who successfully trained a CNN to learn a non-linear mapping from LR to HR images. Later proposals relied on deeper networks and residual learning [16, 17], recursive learning [18], multi-path learning [19], and different loss functions [20] to reduce the reconstruction error between the super-resolved image and the GT image. However, while these methods yield high Peak Signal-to-Noise Ratio (PSNR) values, they tend to produce over-smoothed images which lack high-frequency details. To overcome this, Ledig *et al.* [21] proposed to use Generative Adversarial Networks (GANs) for SR with the SRGAN, to achieve realistic looking images according to human perception. The ESRGAN [4] further improves the SRGAN [21] by several changes to the discriminator and generator. The LR images needed for training the aforementioned deep-learning based super-resolution models are typically created by downsampling HR images with an ideal downscaling kernel, typically bicubic downscaling. However, the images generated by this kernel do not necessarily match real SR images. Additionally, in the downscaling process, important natural image characteristics, such as image sensor noise is removed, which the super-resolution algorithms are then prevented from learning. This results in poor reconstruction results and unwanted artifacts

when a real-world noisy LR image is super-resolved [22].

Real-World Super-Resolution One way to address the the lack of a proper imaging model for RWSR, is to create datasets that consist of real LR/HR image pairs captured using two cameras with different focal lengths [23–25]. However, this method is cumbersome and has inherent problems with the alignment of the image pairs. To overcome the problem of missing real-world training data, Shocher *et al.* [26] propose a zero-shot approach where a small CNN is trained at test time on LR/HR pairs extracted from the LR image itself. Soh *et al.* [27] extend the work of [26] by using meta-transfer learning phase to exploit information from an external dataset. Gu *et al.* [28] train a kernel estimator and corrector CNNs under the assumption that the downscaling kernel belongs to a certain family of Gaussian filters and uses the estimated kernel as input to a super-resolution model. To super-resolve LR images with arbitrary blur kernels, Zhang *et al.* [29] propose a deep plug-and-play framework which takes advantage of existing blind deblurring methods for blur kernel estimation. Bell-Kligler *et al.* [30] trains a GAN to estimate blur kernels from LR images and combines it with the ZSSR SR model [26]. Fritsche *et al.* [31] train a GAN to introduce natural image characteristics to images downsampled with bicubic downscaling, which is then used to train a super-resolution for improved performance on real-world images. Zhang *et al.* [32] propose an iterative network for SR of blurry, noisy images for different scaling factors by leveraging both learning and model-based methods. Most recently Ji *et al.* [6] propose a degradation framework for the creation of LRHR image pairs for training. The degradation framework estimates blur kernels and noise distributions from real LR images in the source domain which are used to degrade HR images in the target domain. This enables training of a GAN based SR model which is shown to perform better on real LR images. However, a key limitation of this method is that it does not address the compression artifacts often found in real-world images.

Face Super-Resolution Face SR is a SR technique specialized for reconstruction of face images. One of the first methods for face SR was proposed by Baker and Kanade [33]. This method reconstructed face details by searching for the most optimal mapping between LR and HR patches. Wang *et al.* [34] used an eigen transformation to map between LR and HR faces. Yang *et al.* [35] use a facial landmark detector to localize facial components which are subsequently reconstructed from similar HR reference components.

More recent work relies on deep learning based methods with CNNs and GANs. Dahl *et al.* [36] use pixel recursive learning with two CNNs to synthesize realistic hair and skin details. Chen *et al.* [37] combine face SR and face alignment to achieve previously unseen PSNR values. By searching the latent space of a generative model for images that downscale correctly,

3. The Proposed Framework

Menon *et al.* [38] are able to create face images of high resolution and perceptual quality. However, the problem with this approach is that the generated faces are often far from the true identity of the actual person, as illustrated in Figure A.2. Additionally, none of the above mentioned methods are robust against noise or other corruptions in the input images [39]. There are very few publications available in the literature which address the problem of RWSR of face-images [39]. Furthermore, the few existing face RWSR methods are only compatible with LR images that have been squared to 16×16 pixels, meaning that the reconstructed image will be only 64×64 or 128×128 pixels depending on the scaling factor [9–11]. Hence, these models cannot perform true RWSR directly on the LR images. This means that the actual usefulness of the existing face SR models is limited. On the contrary, our work presents one possible solution for $\times 4$ RWSR of face images of arbitrary sizes, which we evaluate on real LR face images from surveillance cameras without any prior re-scaling.

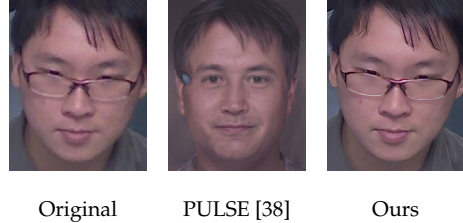


Fig. A.2: An example of SR of a real low-quality face image from the Chokepoint DB [5], where it can be seen that the PULSE [38] method changes the identity of the person, while our method preserves the identity and enhances details.

3 The Proposed Framework

This section describes our two-step framework for RWSR. The first step aims to generate LR images from clean HR images in the target domain \mathcal{Y} , such that these have similar image characteristics as the ones in the source domain \mathcal{X} . The second step involves training a SR model on the constructed paired data, and optimizing for perceptual quality.

3.1 Novel Image Degradation

Traditional approaches for SR assumes that a LR image I_{LR} is the result of a downscaling operation of the corresponding HR image I_{HR} using some kernel k and scaling factor s , namely:

$$I_{LR} = (I_{HR} * k) \downarrow_s \quad (\text{A.1})$$

However, real LR images from cameras are influenced by multiple other factors that degrade the image as well. The RealSR [6] framework tries to address this issue by considering realistic noise distributions and blur kernels in the downscaling process. However, we observe that real images from surveillance

cameras are often also degraded with compression artifacts, which makes the RealSR framework perform poorly on such images. To this end, we extend the degradation framework from [6] to include JPEG compression artifacts in addition to estimation of realistic noise distributions and blur kernels. Thus, we extend the basic SR formulation from Equation 3.1, and assume that the following image degradation model was used to create I_{LR} .

$$I_{LR} = c((I_{HR} * k) \downarrow_s + n) \quad (\text{A.2})$$

where k , s , n , and c denotes the blur kernel, scaling factor, noise, and compression function, respectively. I_{HR} is unknown together with k , n , and c . In our degradation framework, we estimate the kernel and noise directly from the images in the source domain X . We build a pool of the estimated kernels and noise patches which is used to generate corrupted LR images from clean HR images and finally JPEG compress the images, in order to create image pairs for training the SR model.

3.2 Blur Kernel Estimation

For estimation of realistic blur kernels, we adopt the KernelGAN method by Bell-Kligler *et al.* [30]. This method estimates an image specific SR kernel k_i using an unsupervised approach. More specifically, a GAN is trained to down-scale the input image in a way that best preserves the image patch distributions across scales. We estimate realistic blur kernels from all training images in X to form a pool of kernels that can be used to degrade the HR images in Y .

Downsampling To create the downsampled image I_D we randomly choose a blur kernel k_i from the pool of estimated kernels and perform cross-correlation with images in Y . More formally the process is described as:

$$I_D = (Y_n * k_i) \downarrow_s, i \in \{1, 2 \dots m\} \quad (\text{A.3})$$

where I_D is the downsampled image, Y_n is a HR image, k_i refers to a kernel from the degradation pool $\{k_1, k_2, \dots k_m\}$ and s is the scaling factor.

3.3 Noise Estimation

For degradation with realistic image noise, we adopt the method from [14] to extract noise patches from the source images X . Here the assumption is that an approximate noise patch can be obtained from a noisy image by extracting an area with weak background and then subtracting the mean. We define two patches p_i and q_j^i . We obtain p_i by a sliding window approach across images in X , and similarly for q_j^i by scanning p_i . p_i is considered a smooth patch if the

3. The Proposed Framework

following constraints are met:

$$|Mean(q_j^i) - Mean(p_i)| \leq \mu \cdot Mean(p_i) \quad (A.4)$$

and

$$|Var(q_j^i) - Var(p_i)| \leq \gamma \cdot Var(p_i) \quad (A.5)$$

where $Mean$ and Var denotes the mean and variance respectively, and μ and γ are scaling factors. Different from [14] we add an additional constraint to ensure that saturated patches are not extracted:

$$Var(p_i) \geq \phi \quad (A.6)$$

where ϕ denotes a minimum variance threshold. If all constraints are satisfied, p_i will be considered a smooth patch. We then create a pool of noise patches n_i by subtracting the mean value from all valid p_i .

Degradation with Noise We degrade the LR images by injecting real noise patches from the noise pool. For better regularization of the SR model we randomly pick a noise patch from the noise pool and inject it to the LR image during training. The downscaled and noisy LR image I_N is created as follows:

$$I_N = I_D + n_i, i \in \{1, 2 \dots l\} \quad (A.7)$$

where I_D is a downscaled image, and n_i is a noise patch from the noise pool $\{n_1, n_2, \dots n_l\}$

3.4 Degradation with Compression artifacts

Finally, we introduce compression artifacts to the LR training images to close the domain gap between these and the real JPEG compressed LR images in the source domain X . As there are no way of determining the compression strength of existing JPEG images we empirically compare images from X to similar images with different JPEG compression strengths applied and find that a compression strength of 30 results in similar compression artifacts.

3.5 Backbone Model

We base our SR model on the ESRGAN [4], which is one of the SoTA networks for perceptual SR with $\times 4$ upscaling, and train it on the paired LR and HR images generated with our degradation framework. Different from the SRGAN [21], the ESRGAN uses Residual-in-Residual Dense Blocks (RRDBs) in the generator network and the discriminator predicts the relative realness instead of an absolute value. Additionally, the ESRGAN removes the batch

normalization layers used in SRGAN.

Loss Functions While traditional supervised SR models are trained with pixel loss to minimize the Mean Squared Error (MSE) between the reconstructed HR image and the GT image, we rely on loss functions that maximize the perceptual quality. The original ESRGAN [4] model uses several different loss functions during training. More specifically, the generator uses adversarial loss \mathcal{L}_{adv} [40] in combination with VGG perceptual loss \mathcal{L}_{vgg} [1] and pixel loss \mathcal{L}_{pix} , while the discriminator use VGG-128 [41] loss \mathcal{L}_{vgg} . However, we find that this combination of loss functions is not ideal for high perceptual quality. Following the work of [6], we first exchange the VGG-128 [41] discriminator loss with a PatchGAN discriminator from [7] to reduce the amount of artifacts in the reconstructed images. Different from the VGG loss, the PatchGAN loss \mathcal{L}_{patch} has a fully convolutional structure, and only penalizes structure differences at the scale of patches, to determine if an image is real or fake. For optimization of the generator, the loss from all patches are averaged and fed back to the generator. Continuing this track, we seek to also replace the VGG-loss in the generator. Inspired by [8], we find that using the LPIPS perceptual loss \mathcal{L}_{lips} [2] results in less noise and richer textures compared to using VGG-loss for the generator. This is mainly because the VGG network is trained for image classification, while LPIPS is trained to score image patches based on human perceptual similarity judgements. The LPIPS perceptual loss is formulated as:

$$\mathcal{L}_{lips} = \sum_k \tau^k(\phi^k(I_{gen}) - \phi^k(I_{gt})) \quad (\text{A.8})$$

where I_{gen} is a generated image, I_{gt} is the corresponding GT image, ϕ is a feature extractor, τ is a transformation from embeddings to a scalar LPIPS score. The score is computed from k layers and averaged. In our implementation of LPIPS we use the pre-trained AlexNet model provided by the authors. In total, our full training loss for the generator is as follows:

$$\mathcal{L}_{generator} = \lambda_{pix} \cdot \mathcal{L}_{pix} + \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{lips} \cdot \mathcal{L}_{lips} \quad (\text{A.9})$$

where λ_{pix} , λ_{adv} and λ_{lips} are scaling parameters.

3.6 Datasets

This section describes the datasets used for training and testing. For our experiments on real LR face images from surveillance cameras we use the Chokepoint Dataset [5] as our source domain images X . This dataset contains images of 29 different persons captured with three cameras in a real-world surveillance setting. All images have a resolution of 800×600 . We use a face detection algorithm to extract the faces from the images, and randomly

3. The Proposed Framework

split the dataset, to obtain 72,282 images for training and 3,805 images for testing. The average resolution of the cropped faces is $\approx 92 \times 92$. We only use the Chokepoint training images to estimate realistic blur kernels and noise distributions for our degradation framework, and not for direct training of our SR model.

For the target domain of high-quality face images Y , we combine 571 face images from the SiblingsDB [42], 8,040 face images from the Radboud Faces Database [43] and 5,000 randomly selected face images from FFHQ database [12] for a total of 13,611 images. Both the SiblingsDB and Raboud Face Database contains portrait face images professionally captured in a studio setting with controlled lighting. The face images from the FFHQ are more diverse in appearance, and ethnicity of the subjects. We augment all images in the target domain by downsampling by 25, 50 and 75% with bicubic downscaling to obtain a more diverse dataset. We then apply our degradation framework described in Section 3.1 on the images in Y to obtain LR/HR image pairs for training of our SR model.

We also evaluate on both synthetically created LR face and general images. The synthetic setting enables comparison with the traditional full-reference IQA metrics commonly used in SR while the experiments on general images can be used to show the generalization abilities of our method. For evaluation face images, we use the first 1,000 images from the FFHQ dataset. For evaluation on general images we use the DIV2K validation set [13] consisting of 100 images. To generate realistic LR/GT image pairs, we introduce three kinds of corruptions, namely, downsampling, sensor noise, and compression artifacts. For downsampling, we first convolve the image with an 11×11 Gaussian blur kernel with a standard deviation of 1.5. For modeling of sensor noise we follow the protocol from [44] and use pixel-wise independent Gaussian noise, with zero mean and a standard deviation of 8 pixels. For compression artifacts, we convert the images to JPEG using a compression strength of 30.

3.7 Evaluation Metrics

Real-World Images Due to the nature of RWSR, no GT reference image exists, which makes it impossible to compare the different methods using traditional SR IQA methods *e.g.* PSNR and Structural Similarity index (SSIM). To this end, we follow the no-reference based IQA evaluation protocol from the NTIRE2020 RWSR challenge [45]. In particular, we assess the image quality using NIQE [46], BRISQUE [47], PIQE [48], NQRM [49] and PI [50]. PIQE and NIQE are non-learnable metrics which relies only on image statistics. BRISQUE and NQRM are learned metrics, trained on a database of different distortion types. However, for reliable scoring, the image to be scored must contain at least one of the distortions types present in the training data. Finally, PI is a weighted score computed as $\frac{1}{2}((10 - NQRM) + NIQE)$. As no-reference

based IQA is a challenging problem, the aforementioned methods are known to correlate poorly with human ratings [45]. To address this issue, we supplement our evaluation protocol with MOR and NIMA [3], where the latter is a learned metric based on human opinion scores, capable of quantifying image quality with high correlation to human judgement. We use the pre-trained NIMA model for rating of the technical image quality [51]. For the MOR, we ask the participants to rank overall image quality of the SR results. To simplify the ranking, we only include the predictions of the top-5 methods based on NIMA scores. To avoid bias, the order of the methods are randomly shuffled. We average the assigned rank of each method over all images and participants to compute the MOR. Since the MOR is a direct measure of human judgement we use this metric for final assessment of the different methods.

4 Experiments and Results

Implementation Details We perform all our experiments with a scaling factor $s = 4$. For our SR model we jointly train the generator and discriminator for 400K iterations with a batch size of 16. We initialize the weights from the PSNR optimized RRDB model from [4]. We use LR patches of size 32×32 , and empirically set λ_{pix} , λ_{adv} and λ_{lips} to 0.01, 0.005 and 0.001 respectively. For noise estimation we set p_i to match the LR patch size and q_j^i to 8. Similar to [14] we set μ and γ to 0.1 and 0.25 respectively. We empirically set the minimum variance threshold ϕ to 0.5. For degradation with compression artifacts we JPEG compress the LR training images with with a random strength of [15, 30].

4.1 Comparison with State-of-the-Art

We did not find any other $\times 4$ face image specific RWSR methods in the literature. Instead, we compare our method to bicubic upscaling, as well as with different groups of SoTA super-resolution methods including two generic SR models (ESRGAN [4], EDSR [17]), two SR methods for arbitrary blur kernels (DPSR [52], USRNet [32]), two real-world SR models (MZSR [27], and RealSR [6]). We fine-tune or adjust the competing models for optimal performance for a fair comparison. For the unsupervised MZSR [27], we enable back-projection with 10 iterations and set a noise level of 0.5. We re-train the RealSR [6] using the framework provided by the authors. The remaining methods all requires paired training data, which is not available in the real-world SR setting. Due to this, these models cannot be re-trained for our experiments, and as such we use the pre-trained weights provided by the authors. Specifically for USRNet [32] and DPSR [52], we input blur kernels estimated with KernelGAN [30], and set noise levels for real images as recommended by the authors.

4. Experiments and Results

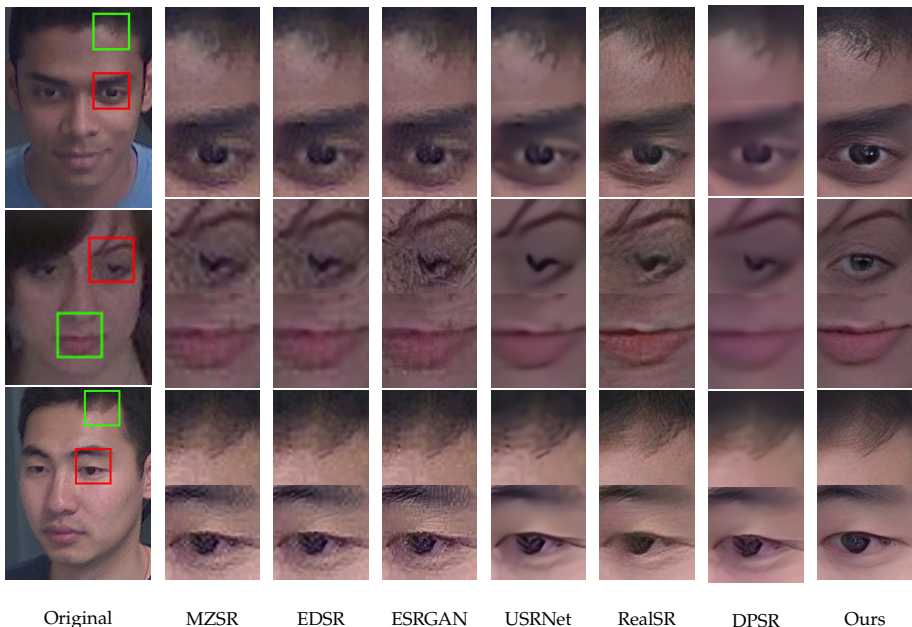


Fig. A.3: Comparison with SoTA methods for $\times 4$ SR of real low-quality face images from the Chokepoint DB [5]. As visible, our method generates superior reconstructions over the existing methods for different faces.

Method	NIQE ↓	BRISQUE ↓	PIQE ↓	NRQM ↑	PI ↓	NIMA ↑	MOR ↓
Bicubic [53]	5.77	56.77	86.28	3.09	6.34	3.92	-
MZSR [27]	7.36	50.09	77.63	3.75	6.81	3.97	-
EDSR [17]	5.43	50.63	81.97	3.82	5.81	4.08	-
ESRGAN [4]	3.75	19.35	19.20	7.08	3.34	4.34	4.72
USRNet [32]	6.10	59.13	87.70	3.19	6.46	4.75	3.11
RealSR [6]	3.50	17.20	9.11	5.45	4.00	4.93	3.39
DPSR [52]	5.58	55.52	60.99	3.38	6.10	5.15	2.71
Ours	4.56	19.07	14.61	7.62	3.47	5.92	1.43

Table A.1: Quantitative results on the Chokepoint testset. ↑ and ↓ indicate whether higher or lower values are desired, respectively. Our model scores lower on the traditional IQA metrics while being superior on the more recent NIMA metric and MOR which indicate that the traditional IQA metrics are not ideal for evaluation of perceptual quality.

Artificially Corrupted Images For our experiments on artificially corrupted images we evaluate the performance using three conventional IQA methods, PSNR, SSIM, and the later Multi Scale Structural Similarity index (MS-SSIM) [54]. However, these metrics focus more on signal fidelity rather than

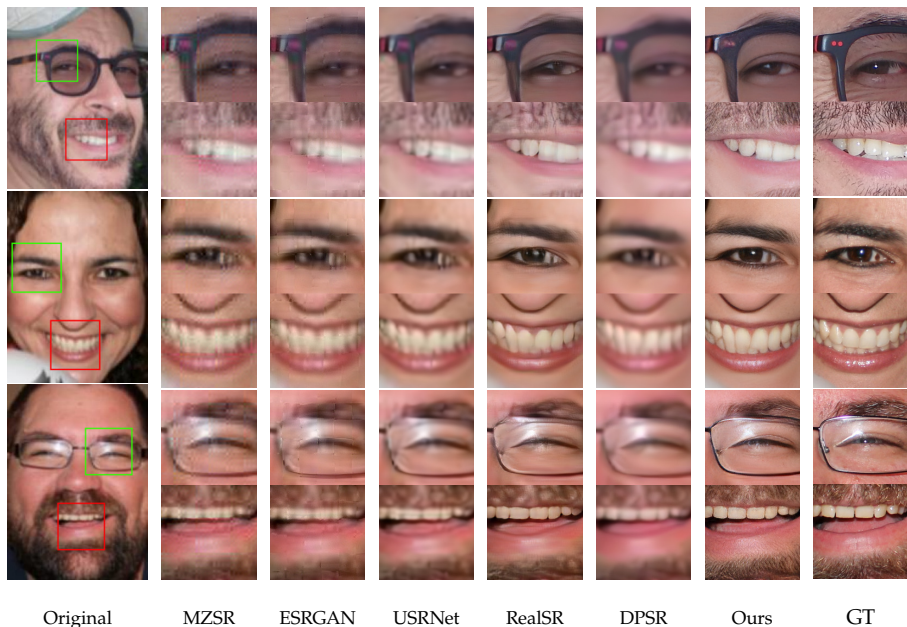


Fig. A.4: Comparison with SoTA methods for $\times 4$ SR of artificially corrupted face images from the FFHQ [12] testset. As seen, our method hallucinates faces with richer detail and less artifacts compared to the existing methods.

Method	PSNR \uparrow	SSIM \uparrow	MS-SSIM \uparrow	NLPD \downarrow	LPIPS \downarrow	DISTS \downarrow
Bicubic [53]	28.39	0.79	0.88	0.32	0.52	0.20
MZSR [27]	29.56	0.78	0.89	0.29	0.43	0.18
EDSR [17]	28.27	0.78	0.88	0.33	0.50	0.19
ESRGAN [4]	28.09	0.77	0.88	0.34	0.40	0.19
USRNet [32]	28.53	0.80	0.89	0.32	0.53	0.21
RealSR [6]	29.14	0.79	0.90	0.29	0.29	0.18
DPSR [52]	27.45	0.79	0.88	0.33	0.51	0.25
Ours	30.20	0.79	0.91	0.28	0.25	0.16

Table A.2: Quantitative results on the FFHQ testset. \uparrow and \downarrow indicate whether higher or lower values are desired, respectively.

perceptual quality [55]. As our method is optimized towards perceptual quality, we also include three of the most recent full-reference metrics targeting perceptual quality, namely Normalized Laplacian Pyramid Distance (NLPD) [56], LPIPS [2], and Deep Image Structure and Texture Similarity (DISTS) [57].

4. Experiments and Results

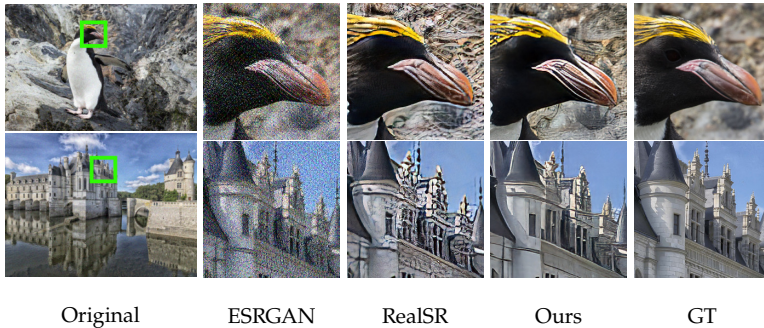


Fig. A.5: Super-resolution results of artificially corrupted LR images from the DIV2K dataset.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Bicubic	25.16	0.65	0.67
ESRGAN [4]	16.40	0.14	0.99
RealSR [6]	18.37	0.50	0.34
Ours	20.95	0.58	0.31

Table A.3: Quantitative results on the DIV2K validation set. \uparrow and \downarrow indicate whether higher or lower values are desired, respectively.

Real-World Face Images In this experiment we evaluate the SR performance on real LR face images from the Chokeypoint testset. Quantitative and qualitative results can be seen in Table A.1 and Figure A.3, respectively. As seen, our method clearly outperforms the other methods in terms of perceptual quality, by producing more detailed reconstructions with less artifacts. However, while the traditional no-reference IQA methods (NIQE [46], BRISQUE [47], PIQE [48] and NQRM [49]) fails to capture this, scores from the more recent NIMA [3] method correlates well with the qualitative results. Finally, the MOR, a direct measure of human judgement, shows that the study participants prefer the reconstructions of our method, over the ones from the competing methods, by a large margin. This further highlights the need for better no-reference IQA metrics for judgement of the perceptual quality.

Artificially Corrupted Face Images This experiment evaluate the SR performance on artificially corrupted images from the FFHQ testset. We show quantitative and qualitative results in Table A.2 and Figure A.4, respectively. As seen, our method produces sharp and detailed images with fewer unpleasant artifacts, which closely resembles the GT images. This is also reflected in the quantitative results. Most noteworthy are the DIST and LPIPS scores, which are known to be highly correlated with human judgement. These highlight the advantage of our method in terms of reconstruction with high perceptual

quality. At the same time, our method results in the best PSNR scores which shows that our reconstructions are also the most accurate.

Artificially Corrupted General Images Finally, we also evaluate on artificially corrupted generic images from the DIV2K [13] validation set. Quantitative and qualitative results can be seen in Table A.3, and Figure A.5, respectively. As seen our method is also applicable to other image domains, where it produces noise free reconstructions with better visual quality compared to ESRGAN and RealSR. Furthermore, our method achieves the best PSNR score, which shows that the reconstructions by method is closer to the ground truth.

4.2 Ablation Study

We evaluate the effect of our proposed method for realistic image degradation and our improved ESRGAN based SR model in the same setting as described in Section 4.1. A qualitative comparison can be seen in Figure A.6.

Baseline Here, we use kernel estimation and noise injection to generate training data for the ESRGAN with patch discriminator, similar to [6]. This SR model is fine-tuned to our face image dataset, and serves as our baseline. The resulting HR images contain unpleasing noise and lack detail.

Compression Artifacts In this setting, we add JPEG compression artifacts to the LR images during training of the baseline model. This results in more noise-free reconstructions compared to the baseline.

LPIPS loss Here, we use the LPIPS loss function for the generator instead of VGG-loss combined with the addition of compression artifacts. When the baseline model is re-trained under these settings the resulting reconstructions becomes sharper with better texture and details.

4.3 Failure Cases

While our method produces reconstructed faces of better visual quality than the compared SoTA methods, it does not solve the problem of RWSR of face images. Figure A.7 shows several failure cases of our method. These occur when the input image is severely corrupted *e.g.* by motion blur or harsh lighting, or when out-of-focus. In these cases, our method might only super-resolve some parts of the face, *e.g.* a single eye, or even hallucinate unrealistic facial features.

5. Conclusion

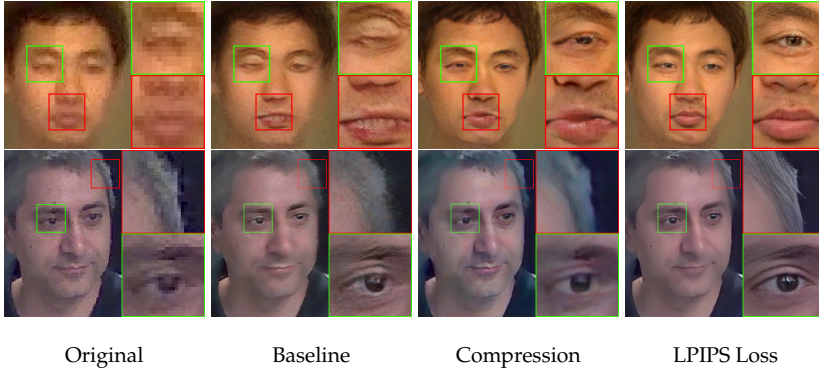


Fig. A.6: Ablation study of the effect of including compression artifacts in the degradation framework and exchanging the VGG-loss with LPIPS-loss for the generator in the SR model, compared to the baseline and the original LR images.

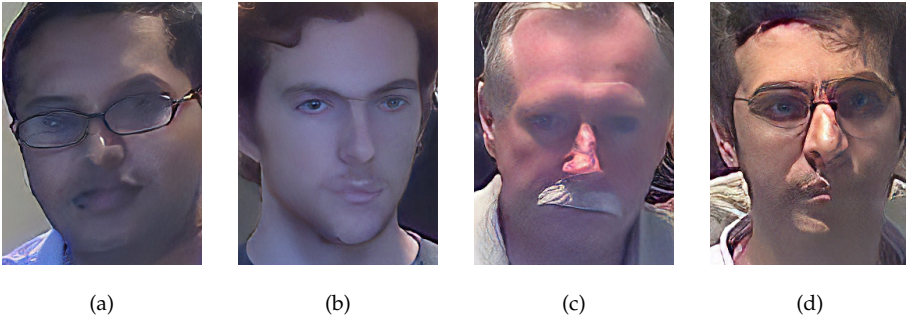


Fig. A.7: Examples of failure cases. Figure (a) and (b) illustrate cases where only parts of the image is super-resolved. Figure (c) shows a case where almost no high-frequency details are restored. Figure (d) shows a case where unrealistic facial features are introduced.

5 Conclusion

In this paper, we have presented a novel framework for RWSR, which we have evaluated on low-quality face images from surveillance cameras, and artificially corrupted face and general images. Our method shows SoTA performance in both cases, which is achieved by making the SR model robust against the most common degradation types present in real LR images, and our novel combination of loss functions. Moreover, our model is the first to perform SR on real LR face images of arbitrary sizes, which makes it useful for practical applications. In the future, even better reconstructions could possibly be obtained by adding attention mechanisms to enable the SR model focus more on the facial components and by including more image degradation types.

6 Acknowledgments

This work was supported by Danmarks Frie Forskningsfond under grant number 8022-00360B.

References

- [1] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 694–711. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_43
- [2] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 586–595.
- [3] H. T. Esfandarani and P. Milanfar, "NIMA: neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, 2018. [Online]. Available: <https://doi.org/10.1109/TIP.2018.2831899>
- [4] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops, L. Leal-Taixé and S. Roth, Eds.* Cham: Springer International Publishing, 2019, pp. 63–79.
- [5] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, June 2011, pp. 81–88.
- [6] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, "Real-world super-resolution via kernel estimation and noise injection," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [7] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2242–2251. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.244>
- [8] Y. Jo, S. Yang, and S. J. Kim, "Investigating loss functions for extreme super-resolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. IEEE, 2020, pp. 1705–1712. [Online]. Available: <https://doi.org/10.1109/CVPRW50498.2020.00220>
- [9] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses

References

- with gans,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 109–117. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Bulat_Super-FAN_Integrated_Facial_CVPR_2018_paper.html
- [10] Z. Cheng, X. Zhu, and S. Gong, “Characteristic regularisation for super-resolving face images,” in *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 2020, pp. 2424–2433. [Online]. Available: <https://doi.org/10.1109/WACV45572.2020.9093480>
- [11] A. Bulat, J. Yang, and G. Tzimiropoulos, “To learn image super-resolution, use a GAN to learn how to do image degradation first,” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11210. Springer, 2018, pp. 187–202. [Online]. Available: https://doi.org/10.1007/978-3-030-01231-1_12
- [12] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.
- [13] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [14] J. Chen, J. Chen, H. Chao, and M. Yang, “Image blind denoising with generative adversarial network based noise modeling,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 3155–3164.
- [15] C. Dong, C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [16] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.
- [17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [18] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.
- [19] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, “Image super-resolution via dual-state recurrent networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1654–1663.
- [20] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

References

- [21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [22] A. Lugmayr, M. Danelljan, R. Timofte, M. Fritsche, S. Gu, K. Purohit, P. Kandula, M. Suin, A. N. Rajagoapalan, N. H. Joon, Y. S. Won, G. Kim, D. Kwon, C. Hsu, C. Lin, Y. Huang, X. Sun, W. Lu, J. Li, X. Gao, S. Bell-Kligler, A. Shocher, and M. Irani, "Aim 2019 challenge on real-world image super-resolution: Methods and results," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3575–3583.
- [23] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3086–3095.
- [24] H. Vaezi Joze, I. Zharkov, K. Powell, C. Ringler, L. Liang, A. Roulston, M. Lutz, and V. Pradeep, "Imagepairs: Realistic super resolution dataset via beam splitter camera rig," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. [Online]. Available: <https://www.microsoft.com/en-us/research/publication/imagepairs-realistic-super-resolution-dataset-via-beam-splitter-camera-rig/>
- [25] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2020.
- [26] M. I. Assaf Shocher, Nadav Cohen, "zero-shot" super-resolution using deep internal learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [27] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [28] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [29] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1671–1681. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Deep_Plug-And-Play_Super-Resolution_for_Arbitrary_Blur_Kernels_CVPR_2019_paper.html
- [30] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 284–293.
- [31] M. Fritsche, S. Gu, and R. Timofte, "Frequency separation for real-world super-resolution," in *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

References

- [32] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3217–3226.
- [33] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," in *2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, 13-15 June 2000, Hilton Head, SC, USA. IEEE Computer Society, 2000, pp. 2372–2379. [Online]. Available: <https://doi.org/10.1109/CVPR.2000.854852>
- [34] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst. Man Cybern. Part C*, vol. 35, no. 3, pp. 425–434, 2005. [Online]. Available: <https://doi.org/10.1109/TSMCC.2005.848171>
- [35] C. Yang, S. Liu, and M. Yang, "Structured face hallucination," in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 1099–1106. [Online]. Available: <https://doi.org/10.1109/CVPR.2013.146>
- [36] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 5449–5458. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.581>
- [37] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [38] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: self-supervised photo upsampling via latent space exploration of generative models," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 2434–2442. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00251>
- [39] K. Grm, M. Pernus, L. Cluzel, W. J. Scheirer, S. Dobrisek, and V. Struc, "Face hallucination revisited: An exploratory study on dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 2405–2413. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2019/html/Biometrics/Grm_Face_Hallucination_Revisited_An_Exploratory_Study_on_Dataset_Bias_CVPRW_2019_paper.html
- [40] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 2672–2680. [Online]. Available: <http://papers.nips.cc/paper/5423-generative-adversarial-nets>
- [41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
- [42] T. F. Vieira, A. Bottino, A. Laurentini, and M. De Simone, "Detecting siblings in image pairs," *The Visual Computer*, vol. 30, no. 12, pp. 1333–1345, Dec 2014. [Online]. Available: <https://doi.org/10.1007/s00371-013-0884-3>

References

- [43] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.
- [44] A. Lugmayr, M. Danelljan, and R. Timofte, "Unsupervised learning for real-world super-resolution," in *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE, 2019, pp. 3408–3416. [Online]. Available: <https://doi.org/10.1109/ICCVW.2019.00423>
- [45] A. Lugmayr et al., "Ntire 2020 challenge on real-world image super-resolution: Methods and results," *CVPR Workshops*, 2020.
- [46] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
- [47] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012. [Online]. Available: <https://doi.org/10.1109/TIP.2012.2214050>
- [48] V. N., P. D., M. C. Bh., S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Twenty First National Conference on Communications, NCC 2015, Mumbai, India, February 27 - March 1, 2015*. IEEE, 2015, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/NCC.2015.7084843>
- [49] C. Ma, C. Yang, X. Yang, and M. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Underst.*, vol. 158, pp. 1–16, 2017. [Online]. Available: <https://doi.org/10.1016/j.cviu.2016.12.009>
- [50] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 334–355.
- [51] C. Lennan, H. Nguyen, and D. Tran, "Image quality assessment," <https://github.com/idealo/image-quality-assessment>, 2018.
- [52] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1671–1681.
- [53] R. G. Keys, "Cubic Convolution Interpolation for Digital Image Processing," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 29, pp. 1153–1160, Jan. 1981.
- [54] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar, 2003)*, pp. 1398–1402.
- [55] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6228–6237. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html

References

- [56] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized laplacian pyramid," in *Human Vision and Electronic Imaging, HVEI 2016, San Francisco, California, USA, February 14-18, 2016*, H. de Ridder, T. N. Pappas, and B. E. Rogowitz, Eds. Ingenta, 2016, pp. 1–6. [Online]. Available: <http://ist.publisher.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000016/art00008>
- [57] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *CoRR*, vol. abs/2004.07728, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07728>

References

Paper B

Real-World Thermal Image Super-Resolution

Moaaz Allahham, Andreas Aakerberg, Kamal Nasrollahi, and
Thomas B Moeslund

The paper has been published in
Advances in Visual Computing - 16th International Symposium, ISVC 2021,
Lecture Notes in Computer Science, vol. 13017. Springer, pp. 3-14, 2021.

© 2021 by the authors.

The layout has been revised.

Abstract

Thermal cameras are used in various domains where the vision of RGB cameras is limited. Thermographic imaging enables the visualizations of objects beyond the visible range, which enables its use in many applications like autonomous cars, nightly footage, military, or surveillance. However, the high cost of manufacturing this type of camera limits the spatial resolution that it can provide. Real-World Super-Resolution (RWSR) is a topic that can be used to solve this problem by using image processing techniques that enhance the quality of a real-world image by reconstructing lost high-frequency information. This work adapts an existing RWSR framework that is designed to super-resolve real-world RGB images. This framework estimates the degradation parameters needed to generate realistic Low-Resolution (LR) and High-Resolution (HR) image pairs, then the SR model learns the mapping between the LR and HR domains using the constructed image pairs and applies this mapping to new LR thermal images. The experiments results show a clear improvement in the perceptual quality in terms of clarity and sharpness, which surpasses the performance of the current SoTA method for thermal image SR.

1 Introduction

In recent years, thermal imaging has grown considerably and is being used in various domains where a typical RGB camera can not get the job done, like nightly footage, surveillance, or in autonomous cars. However, thermal images generally have some shortcomings like insufficient details and blurred edges, and most importantly considerably low-resolution. This makes it too hard to observe the structure and recognize objects in an image. However, having a thermal camera that is capable of capturing high-resolution images is not as affordable as using RGB cameras. Even the most expensive thermal cameras, which can vary from US\$200 to more than US\$20,000 [1], still can not deliver sufficient resolutions. To the best of our knowledge, the highest resolution that a thermal camera can provide as for today is 1920×1200 pixels for the Vayu HD [2], thus enhancing real images captured by thermal cameras is therefore important. However, although increasing the resolution of a thermal image with an image processing algorithm would not compensate for the true information that is not captured by the camera’s sensor, having an enhanced and higher resolution image makes it easier to recognize objects and structure in an image. The efficiency of this process can be improved by taking advantage of computer vision techniques that can assist in enhancing these images. Many methods were developed to perform image super-resolution, however, most of these methods perform poorly when used on real LR images. This is because they follow the approach of downsampling HR images to construct LR and HR pairs and then they super-resolve the LR image back to match the

HR image quality. Such methods fail when given a real-world image as the degradation process is not entirely known. Therefore recent studies have been working on developing methods that would be more robust to previously unseen real-world images that are acquired directly from cameras with unknown degradation parameters. This RWSR issue also applies to the thermal imaging domain, making it an interesting area to investigate since it has not been widely explored. Hence, the goal of this project was to explore the State-of-The-Art (SoTA) SR algorithms that deal with RGB images and investigate its usability in the thermal imaging domain, and explore the possibility of tuning these methods to fit the thermal domain. The main contributions of this work are:

- A comparison of the performance of existing RGB-based RWSR solutions in the thermal imaging domain.
- SoTA results within the real-world thermal SR domain are achieved.

2 Related Work

2.1 RGB Image Super-Resolution

Zero-shot Methods

In 2017, ZSSR [3] was introduced as the first blind SR algorithm (self-learning-based) that performed SR on LR real-world images without relying on any prior image examples or prior training. Instead, ZSSR trains an image-specific CNN using the recurrence of small patches across different scales within the same image at test time. This was done by downscaling the test image to smaller versions of itself, then applying data-augmentation to the smaller versions to fulfill the need of having multiple examples as a training dataset. The image-specific CNN learns to reconstruct the original LR image using the downscaled examples, then they finally apply the trained CNN to the original test image to construct the desired HR output. ZSSR outperformed external-based SoTA methods in some regions when tested on images with salient recurrence of information. A drawback of ZSSR is the fact that the learning process fully depends on the internal information in the test image, which makes it require thousands of back-propagation gradient updates. This yields slow testing time as well as poor results in some regions compared to other external-based methods [4]. Inspired from ZSSR, Meta-Transfer Learning for Zero-Shot Super-Resolution (MZSR) [4] was introduced, where the authors of MZSR utilize the powerful parts of ZSSR and improve upon it by introducing the concept of Meta-Transfer learning. The idea behind how meta-learning works is to make the model adapt fast to new blur kernel scenarios by adding a meta-training step, then utilize transfer-learning by pre-training the SR network using a

2. Related Work

large-scale dataset DIV2K [5]. The combination of Meta-transfer learning and ZSSR exploits both the internal (the test image) and external (the DIV2k) information. The main advantage that was introduced in the MZSR work, was the flexibility and fast running time compared to the ZSSR method, as well as outperforming other supervised SoTA algorithms such as CARN [6] and RCAN [7]. Different zero-shot methods were designed following the ZSSR principle, however the most recent study that was able to achieve competitive SoTA results was Dual Super-resolution (DualSR) [8]. DualSR addresses the RWSR problem in a similar way to the way it was addressed in the ZSSR work, where they learn the image-specific LR-HR relations by training their proposed network at the test time using patches extracted from the test image. Their proposed network is split into mainly two parts, the downsampler which learns the degradation process using a generative adversarial network (GAN), and an upsampler that learns to super-resolve the LR image. Both the up-sampler and down-sampler are trained simultaneously by improving each other using the cycle-consistency loss, the masked interpolation loss, and the adversarial loss.

Learned Degradation based Super-resolution

Many supervised SR approaches make the assumption that LR images are a bicubically downsampled version of their HR counterpart, and that Gaussian noise is usually used to simulate the sensor noise. However, these approaches fail when tested on real images because those images were not degraded using ideal degradation operation (bicubic kernel + Gaussian noise). For this reason, Fritsche et al. [9] introduced DSGAN(the winner of AIM2019 RWSR challenge [10]), which is a GAN network that learns to generate the appropriate LR images, which have the same corruptions as the original HR images. Bell-Kligler et al. [11] introduced another realistic degradation method KernelGAN, an image-specific Internal-GAN, which trains solely on the LR test image at test time and learns its internal distribution of patches. The generator of the network is trained to produce a lower resolution image such that the network’s discriminator can not distinguish between the patch distribution of the generated image and the patch distribution of the original LR image. Ji et al. [12] proposed their method RealSR, which is divided into two stages. They first use KernelGAN to estimate the degradation from the real data and use it to construct the LR images, and then they train an SR model based on the constructed data. RealSR method was the winner of the NTIRE 2020 challenge [10], and by the time of doing this work, RealSR is considered to be the SoTA in the real-world super-resolution field for RGB images.

2.2 Thermal Image Super-Resolution

All the methods mentioned in 2.1 are examples of super-resolution methods that deal with images in the RGB spectrum. However, there are only a few studies that developed methods for super-resolving LR thermal images. Cho et al. [13] conducted a study where they tried to enhance thermal images by training a CNN using different image spectrums aiming to find the best representation that would fit the thermal domain. They found that a grayscale trained network provided the best enhancement. Lee et al. [14] proposed a similar CNN-based on enhancement for thermal images, where they evaluated four RGB-based domains with a residual-learning technique. That improved the enhancement in comparison to the previous work by [13]. Rivadeneira et al. [15] was motivated by the two previously proposed methods, so he proposed the Thermal Enhancement Network (TEN), which was the first CNN-based method to be trained specifically using thermal dataset unlike the two previous proposals by [13, 14]. TEN was based on the SRCNN model [16], which utilizes the residual net and dense connections technique. TEN was able to outperform the previously proposed methods, which was due to training the network using thermal images instead of RGB-based domains. Recently, Rivadeneira et al. [1] proposed another thermal SR method that is based on the well-known CycleGAN [17] architecture. Two-way Generative-Adversarial-network (CycleGAN) is a technique that is used to map information from one domain to another. So the authors of [1] used the CycleGAN network to map information from the LR domain to the HR domain. They trained their proposed network to perform x2 scale SR following two scenarios, LR to medium-resolution (MR) and MR to HR. Chudasama et al. [18] proposed TherISuRNet, which is another method to super-resolve thermal images by progressively upscaling the LR test image to obtain the final SR image. They achieve different upscaling factors (x2, x3, and x4) by applying residual learning. The TherISuRNet network consists of four main modules: low-frequency feature extraction modules, high-frequency feature extraction modules, second high-frequency feature extraction modules, and finally an image reconstruction module that is responsible for reconstructing the final SR image. They measured the performance of their proposed method by comparing its performance to the most common SoTA methods [7, 15, 19–21] and bicubic interpolation, and they were able to surpass all the other methods when testing on thermal images. TherISuRNet was the winning method of the Thermal Image Super-Resolution Challenge PBVS 2020 [22], which makes the TherISuRNet the SoTA method for the thermal image SR domain.

Constraints Noted from Related Works

Having reviewed the relevant literature on super-resolution applied to both RGB and thermal images, we witness that, to the best of our knowledge:

- None of the studies try to investigate the performance of RGB-based SR methods in the thermal domain.
- All the existing thermal SR methods were trained using synthetically constructed image pairs.

3 Dataset

One of the challenges when working with RWSR methods is the lack of ground truth data that could be used for supervised learning and to evaluate the performance of the SR methods leading to unreliable performance when testing on single real world images. For this work, the PBVS dataset [1, 22] was used as it offers three subsets called *Domo*, *Axis* and *GT* with different native resolutions (160×120 , 320×240 , 640×512 , respectively), which were acquired using three different cameras. For this work, the *Domo* and *GT* subsets are used as the source and target domains respectively. Each of these subsets includes a total of 951 training images and 50 images for validation. The *Axis* subset was discarded since the goal of this work was to super-resolve a given resolution with an upscaling factor of $s = 4$ and later evaluate the performance by comparing it to the ground-truth, which has a native resolution that matches the SR output images. Therefore, it was decided to super-resolve the input images (*Domo* validation subset) and compare the output with the ground truth (*GT* validation subset). However, one of the problems with the PBVS dataset is the limited number of images in each subset, which is considered too little to be used for training a neural network. Therefore, we used the augmented version of the PBVS dataset, which was provided by the authors of the TherISuR-Net [18]. The augmentation operations they apply on the original dataset are horizontal flipping, 180° rotation, and two affine operations, resulting in a total of 4755 training images for each subset.

4 Thermal RealSR

This section describes the two-step pipeline that T-RealSR uses to achieve the final SR results. The first step aims to realistically degrade the HR from the target domain Y , such that the degraded images have the same image characteristics as the LR images in the source domain X . The second step is to use the LR-HR image pairs to train a SR model that can be used to super-resolve real-world thermal images.

4.1 Realistic Degradation using KernelGAN and Noise Injection

To understand how we can construct a realistic LR image that does not have ideal blurring and noise characteristics, let's assume an LR image is obtained following the degradation operation [12]:

$$I_{LR} = (I_{HR} * k) \downarrow_s + n \quad (\text{B.1})$$

Where k denotes the kernel used to blur the image, n denotes the noise added to the image, and s denotes the downscaling factor. Instead of using ideal kernels (e.g. Bicubic downscaling), T-RealSR explicitly utilizes KernelGAN to create a pool of kernels, and it extract noise patches from a real LR images to create a noise patches pool. Then both these pools are used to construct the realistic LR-HR image pairs.

Kernel degradation

In general, KernelGAN is an image-specific Internal-GAN [23] that trains solely on a given LR image at test time and learns its internal distribution of patches. Its generator (G) is trained to generate a downsampled version of the given image, such that its discriminator (D) can not distinguish between the patch-distribution of the generated image and the patch distribution of the original image. D is trained to output a heat map, referred to as $D\text{-map}$, indicating for each pixel how likely is its surrounding patch to be drawn from the original patch-distribution. The loss is the pixel-wise MSE difference between the output $D\text{-map}$ and the label map. Where the label map is all the ones in the crops extracted from the original image, and all the zeros in the crops extracted from the downsampled image [11].

Noise Extraction

In addition to creating the kernel pool, T-RealSR introduces a simple filtering rule for extracting noise patches from source images. The idea behind extracting these noise patches is to inject them into the degraded images, so LR images from the two different domains (source LR and generated LR images) will have similar noise distribution. The filtering rule used to choose the relevant noise patch is as follows:

$$\sigma(n_i) < v \quad (\text{B.2})$$

Where $\sigma(\cdot)$ denotes the function used to calculate the noise variance, and v is the max value of variance.

Having created a series of kernels $\{k_1, k_2, \dots, k_l\}$ and a series of noise patches $\{n_1, n_2 \dots n_m\}$, the degradation process is performed as follows:

$$I_{LR} = (I_{HR} * k_i) \downarrow_s + n_j, i \in 1, 2, \dots, l, j \in 1, 2, \dots, m \quad (\text{B.3})$$

Where s denotes the sampling stride.

4.2 Super-Resolution Model

As mentioned in section 2.1, T-RealSR consists of two phases, the first is constructing the realistic image pairs using KernelGAN and the second phase is training the SR model, which is based on ESRGAN with some modification. To understand the T-RealSR SR backbone, we need to first understand how ESRGAN works and then understand how T-RealSR adjust the ESRGAN architecture to make it more flexible to different image sizes. ESRGAN [24] stands for Enhanced Super-Resolution Generative Adversarial Networks, which is a generative adversarial network that is based on SRGAN [25]. SRGAN is a GAN network that is capable of generating realistic textures during single-image SR, whose discriminator aims to base its prediction on perceptual quality. However, ESRGAN improves SRGAN by adjusting the SRGAN architecture where they introduce their Residual-in-Residual Dense Block (RRDB) without batch normalization, as well as improving the SRGAN discriminator by making it judge whether an image is more realistic than another rather than judging whether an image is real or fake. ESRGAN improvement over SRGAN resulted in sharper and more visually pleasing results [24].

From the name Enhanced Super-Resolution GAN, we can tell that the architecture should contain the two main modules, discriminator D and generator G networks. The G network takes a low-resolution image (LR) as input, and it passes it through a 2D convolutional layer (Conv1) with small 3×3 kernels and 64 feature maps. It is then passed through 23 Residual in Residual Dense Blocks (RRDB). The image is then passed through another convolutional layer (Conv2) in which its output is summed with the output of the first (Conv1). At this stage, the image gets upsampled with a factor of 4 by passing it through an upsampling block that consists of two convolutional layers for reconstruction, with LeakyReLU (LReLU) activation ($\alpha = 0.2$) on each layer. After upsampling, the image is passed through another convolutional layer (Conv3) with LReLU activation ($\alpha = 0.2$). Finally, the image is passed through the final convolutional layer (Conv4) that final super-resolved image. The other part of the network is the discriminator D , and to be more specific it is called the Relativistic Discriminator [26]. Following [24] this specific discriminator was used rather than using the standard discriminator used in SRGAN [25]. This is because the relativistic discriminator estimates the probability that a real image x_r is relatively more realistic than a fake one x_f . Where a standard discriminator estimates only whether an image x is natural enough to be real.

We adapted the ESRGAN structure and trained it using the constructed paired data $\{I_{LR}, I_{HR}\}$. Several losses were used during the training including:

- **Pixel loss** L_1 : or so called Mean Absolute Error (MAE), which measures the mean absolute pixel difference of all pixels in two given images.
- **Perceptual loss** L_{per} : proposed to enhance the visual quality by minimizing the error in feature space instead of pixel space. It uses the inactive features of VGG-19 [27] and aims to enhance the visual quality of low-frequency information like edges.
- **Adversarial loss** L_{adv} This loss is used to enhance the texture details to make the image look more realistic.

The final loss function was the weighted sum of all the above losses as follows:

$$L_{total} = \lambda_1 L_1 + \lambda_{per} L_{per} + \lambda_{adv} L_{adv} \quad (B.4)$$

Where λ_1 , λ_{per} , and λ_{adv} are constants used to specify the weight of each of the losses on the total loss.

PatchGAN Discriminator

The discriminator (VGG-128) used in the ESRGAN may introduce many artefacts, so PatchGAN [28] was used instead for two reasons: First is that VGG-128 used by ESRGAN limits the size of the generated image to 128, making multi-scaling training not as simple, Second is that the VGG-128 fixed fully connected layer makes the discriminator pays more attention to the global features and ignore the local ones. Where the PatchGAN has a fully convolutional structure that maintains a fixed receptive field that restricts the discriminator’s attention to the local image patches. The structure of PatchGAN only penalizes structure at the scale of patches, meaning that it tries to classify if each $N \times N$ patch in an image is real or fake. The responses of all patches get averaged afterward forming the final D output to guarantee global consistency, then gets fed back to the generator.

5 Experiments and Results

5.1 Evaluation Metrics

Usually, the most challenging part when dealing with RWSR images is the lack of GT reference images. However, despite having the GT images, which the PBVS dataset provides, the SR and GT images are not perfectly aligned together. Making it difficult to use reference-based IQA methods such as SSIM, PSNR, or LPIPS, however we still use them for reference purposes. Additionally, it was decided to take another evaluation approach by following the IQA evaluation protocol from the NTIRE2020 challenge, where they used

non-reference-based IQA methods including PIQE, NIQE, and BRISQUE. In addition to that, the Mean Opinion Score (MOS) method was used to support the previously mentioned non-reference-based methods, which correlate poorly with human opinion. For the MOS, a total of 20 participants were given a set of 13 SR images that were generated using different methods. Then the participants were asked to give unique scores that range between 1 and 6 (best to worse respectively) to each individual images based on the perceived clarity and sharpness of the images. The results of 6 different SR methods were used, where the methods were shuffled randomly when presented to the participants to avoid bias. The scores were then averaged for the individual images for each method, and were then used to calculate the final MOS scores.

5.2 Comparison with the State of the Art

To the best of our knowledge, an evaluation of the adapted T-RealSR method as well as the other mentioned SoTA SR rgb-methods within the thermal domain, in comparison to the SoTA thermal SR method has not be done before. Therefore, we compare the adapted method to bicubic upscaling, as well as with a number of RWSR methods including two zero-shot SR methods (DualSR [8], KernelGAN+ZSSR [3, 11]) and the ESRSGAN [24] RWSR method, and for the thermal SoTA SR method TherISuRNet [18]. To ensure a fair comparison, ESRSGAN [24] was retrained using the same dataset used to train the adapted T-RealSR, and employing the settings suggested by the authors of the ESRSGAN. For DualSR [8] and KernelGAN+ZSSR [3, 11], a training is not needed, as it is a part of the inference phase; the settings suggested by the authors were used. For the TherISuRNet [18], the retraining was needed as pretrained weights were not provided by the authors, and the same settings were adapted because the method was designed specifically for the utilized PBVS dataset.

Image Registration

We explained in Section 3 how the PBVS subsets (Domo and GT) were acquired using different cameras. Despite the effort by the authors to acquire two identical pictures of the same scene using different cameras, the process was physically impossible. That introduced some challenges when having to evaluate the performance of the different SR methods. Besides the different light conditions and different sensors' noise that resulted in brightness and contrast differences, the images were not perfectly aligned together. The imperfect alignment of the images meant that reference-based IQA methods in general and PSNR in specific, will be inaccurate to be used on their own. Therefore, we decided to apply image registration between the SR images and the GT reference images prior to evaluating the images using the non-reference-based

methods. To do so, the ORB detector [29] with a target number of features $N = 5000$ was used to align the images together as illustrated in Figure B.1. The central crop (50%) of both the SR and GT images was used for evaluation. This was done to discard the black areas around the registered images and to make the comparison as fair as possible, since lens distortion is at its minimum in the central part of the image.

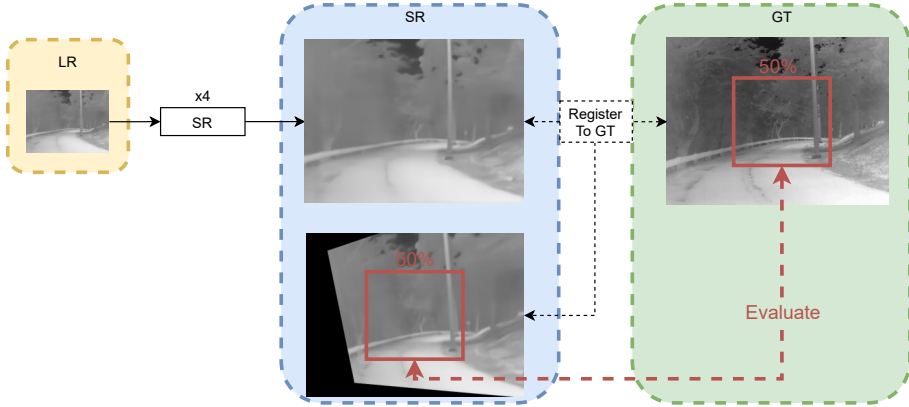


Fig. B.1: The evaluation pipeline used to evaluate the super-resolved LR image in comparison to the GT.

Quantitative and Qualitative Evaluation

We evaluate the performance of the methods on the PBVS test dataset, where we show the quantitative results in table B.1. For the qualitative results a number of patches taken from some test images are shown in Figure B.2. The adapted T-RealSR method outperforms the other thermal and rgb-based SR methods by a large margin. Where it is possible to see that the traditional non-reference-based IQA methods (PIQE, NIQE, BRISQUE) correlate well with the human-opinion based MOS method. However, the reference-based IQA methods (SSIM and PSNR) correlate poorly with the other IQA methods. This is due to brightness and contrast differences. A method such as PSNR, will penalize the performance in case the registered image is shifted one pixel in any direction, and we know for sure that this is most likely the case with our test data.

6 Conclusion

In this work we investigate the possibility of using rgb-based RWSR methods to super-resolve real-world thermal images. The images used for evaluation were

6. Conclusion

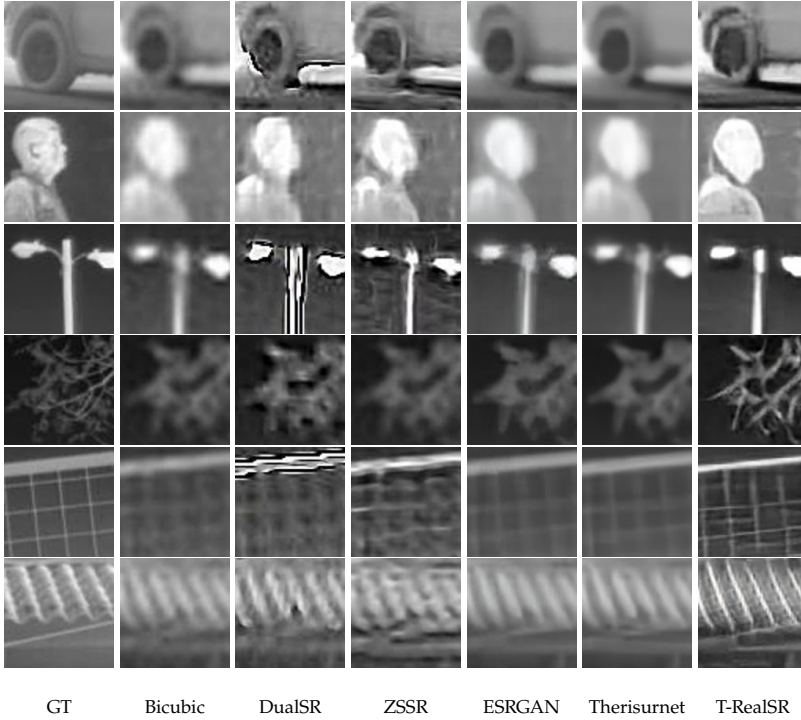


Fig. B.2: Qualitative comparison of SoTA methods for x4 SR of real LR images from the Domo validation subset.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PIQE \downarrow	NIQE \downarrow	BRISQUE \downarrow	MOS \downarrow
Bicubic	20.11	0.70	0.46	67.39	5.55	57.20	4.10
DualSR [8]	18.77	0.59	0.43	56.48	4.18	43.03	4.74
ZSSR+KernelGAN [11]	19.01	0.57	0.44	60.79	5.71	46.14	4.15
ESRGAN [24]	18.37	0.65	0.43	76.77	5.72	53.74	2.98
TherISuNet [18]	20.10	0.71	0.42	88.69	5.20	55.34	3.20
T-RealSR [12]	18.78	0.52	0.37	36.33	3.31	34.31	1.45

Table B.1: Comparison between the SotA methods that have been tested. The best values are in bold text.

upscaled with a factor of 4, and we found that tuning the T-RealSR by training it using thermal images is able to achieve SoTA performance that surpasses the current SoTA thermal-based SR method by a large margin in terms of perceived quality. This was proven by the different IQA methods, which showed results that correlate with the human-based MOS evaluation method. This work is, up to our knowledge, the first work that train on thermal images using realistically degraded image pairs, making it robust to real images that contain some of the most common degradation types (blurring and sensor noise).

References

- [1] R. E. Rivadeneira, A. D. Sappa, and B. X. Vintimilla, "Thermal image super-resolution: a novel architecture and dataset," in *International Conference on Computer Vision Theory and Applications*, 2020, pp. 1–2.
- [2] Sierra-Olympic, "Vayu HD feature specification," accessed: 02/06-2021. [Online]. Available: https://sierraolympic.com/wp-content/uploads/2020/06/2020_VayuHD_Sell-Sheet_FINAL.pdf
- [3] A. Shocher, N. Cohen, and M. Irani, "'Zero-Shot' super-resolution using deep internal learning," 2017.
- [4] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," 2020.
- [5] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [6] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," 2018.
- [7] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 294–310.
- [8] M. Emad, M. Peemen, and H. Corporaal, "Dualsr: Zero-shot dual learning for real-world super-resolution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 1630–1639.
- [9] M. Fritsche, S. Gu, and R. Timofte, "Frequency separation for real-world super-resolution," in *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.
- [10] A. Lugmayr, M. Danelljan, R. Timofte, N. Ahn, D. Bai, J. Cai, Y. Cao, J. Chen, K. Cheng, S. Chun, W. Deng, M. El-Khamy, C. Ho, X. Ji, A. Kheradmand, G. Kim, H. Ko, K. Lee, J. Lee, and X. Zou, "NTIRE 2020 challenge on real-world image super-resolution: Methods and results," 2020.
- [11] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 284–293.
- [12] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, "Real-world super-resolution via kernel estimation and noise injection," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [13] Y. Cho, N. Bianchi-Berthouze, N. Marquardt, and S. J. Julier, "Deep thermal imaging," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Apr 2018.
- [14] K. Lee, J. Lee, J. Lee, S. Hwang, and S. Lee, "Brightness-based convolutional neural network for thermal image enhancement," *IEEE Access*, vol. 5, pp. 26 867–26 879, Nov. 2017.

References

- [15] Y. Choi, N. Kim, S. Hwang, and I. S. Kweon, "Thermal image enhancement using convolutional neural network," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 223–230.
- [16] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [17] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2242–2251. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.244>
- [18] V. Chudasama, H. Patel, K. Prajapati, K. P. Upla, R. Ramachandra, K. Raja, and C. Busch, "Therisurnet - a computationally efficient thermal image super-resolution network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [19] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [20] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *The European Conference on Computer Vision (ECCV)*, September 2018.
- [21] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "Srfeat: Single image super-resolution with feature discrimination," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [22] R. E. Rivadeneira, "Thermal image super-resolution challenge - PBVS 2020," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 432–439.
- [23] A. Shocher, S. Bagon, P. Isola, and M. Irani, "Ingan: Capturing and remapping the DNA" of a natural image," 2019.
- [24] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "Esrgan: Enhanced super-resolution generative adversarial networks," 2018.
- [25] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [26] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," 2018.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
- [28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.
- [29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 2564–2571.

References

Paper C

Semantic Segmentation Guided Real-World Super-Resolution

Andreas Aakerberg, Anders S. Johansen, Kamal Nasrollahi, and
Thomas B Moeslund,

The paper has been published in the
IEEE/CVF Winter Conference on Applications of Computer Vision Workshops,
WACV - Workshops, pp. 449-458, 2022.

© 2022 IEEE. Reprinted, with permission, from:

Andreas Aakerberg, Anders S. Johansen, Kamal Nasrollahi, and Thomas B Moeslund. "Semantic segmentation guided real-world super-resolution", 2022 *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops*, pp. 449-458, 2022.

The layout has been revised.

Abstract

Real-world single image Super-Resolution (SR) aims to enhance the resolution and reconstruct High-Resolution (HR) details of real Low-Resolution (LR) images. This is different from the traditional SR setting, where the LR images are synthetically created, typically with bicubic downsampling. As the degradation process for real-world LR images are highly complex, SR of such images is much more challenging. Recent promising approaches to solve the Real-World Super-Resolution (RWSR) problem include the use of domain adaptation to create realistic training-pairs, and self-learning based methods which learn an image specific SR model at test time. However, as domain adaptation is an inherently challenging problem in itself, SR models based solely on this approach are limited by the domain gap. In contrast, while self-learning based methods remove the need for paired-training data by utilizing internal information in the LR image, these methods come with the cost of slow prediction times. This paper proposes a novel framework, Semantic Segmentation Guided Real-World Super-Resolution (SSG-RWSR), which uses an auxiliary semantic segmentation network to guide the SR learning. This results in noise-free reconstructions with accurate object boundaries, and enables training on real LR images. The latter allows our SR network to adapt to the image specific degradations, without Ground-Truth (GT) reference images. We support the guidance with domain adaptation to faithfully reconstruct realistic textures, and ensure color consistency. We evaluate our proposed method on two public available datasets, and present State-of-the-Art results in terms of perceptual image quality on both real and synthesized LR images.

1 Introduction



Fig. C.1: Super-resolution ($\times 4$) of a real image from the Cityscapes dataset [1]. By combining domain adaptation (DA) and guidance by semantic segmentation, our proposed method reconstructs visually pleasing images. In contrast, ESRGAN fails to handle the corruptions in the real image, resulting in many artifacts.

Single image Super-Resolution (SR) aims to upsample a Low-Resolution (LR) image and reconstruct the missing high-frequency details. SR has been a widely studied problem for decades, due to its vast number of applications in fields such as medical imaging, remote sensing, and surveillance. In latter,

SR are often used to improve the performance of down-stream vision tasks, such as object detection and tracking, by improving the visibility of the images which often suffer from low-resolution due to the wide field-of-view and large object to camera distance. Traditionally, most work has been focusing on improving the fidelity of the images by minimizing the Mean Squared Error (MSE). However, recently more focus has been put into generating realistic High-Resolution (HR) images as perceived by humans [2]. Current State-of-The-Art (SoTA) deep learning-based SR methods most often require paired LR/HR images to be trained by supervised learning. Commonly, researchers have been using artificial LR images created by downsampling HR images, typically using bicubic interpolation. However, this strategy changes the natural image characteristics, such as sensor noise and other corruptions, which limits a SR model trained on such data to perform well on real LR images. Blind SR tries to address this problem by assuming an unknown downsampling kernel, but it still relies on Ground-Truth (GT) reference images for supervised learning.

Recent promising approaches to solve the Real-World Super-Resolution (RWSR) problem, where there aren't any LR/HR pairs for training, includes methods based on domain adaptation [3–5], where [3] was the winner of the NTIRE 2020 Challenge on RWSR [6]. These methods aim at creating synthetic LR images with similar characteristics as the real LR images. However, SR models relying solely on this approach are limited by the domain gap, due to the inherently challenging domain adaptation process. Self-learning based methods [7, 8] removes the need for paired training images, by learning an image specific SR model at test time, using only internal information available in the input image. However, this comes with a significant cost in terms of increased inference time [9].

In this work, we propose a novel framework, Semantic Segmentation Guided Real-World Super-Resolution (SSG-RWSR), to handle SR of real LR images without GT references or prior knowledge about the image formation model. We address the lack of training data by a combination of domain adaptation and guiding the SR learning by the loss of an auxiliary semantic segmentation network. Semantic Segmentation (SS) is a computer vision technique that provides scene understanding by dense labeling of pixels in an image. We argue that the loss of the SS task provides strong cues about the fidelity of the images, which can be used to jointly optimize the SR model towards producing more accurate, and noise-free HR images. The loss of the SS task also enables training on real LR images, without the need for GT reference image, which we argue can help the SR model adapt to the image-specific degradations. To reconstruct realistic textures, and ensure color consistency with the LR images, we propose to simultaneously train on synthetically generated LR/HR image pairs. To this end, we leverage domain adaptation to obtain LR images, with similar characteristics and corruptions as the real images. At test time, we

2. Related Work

decouple the SS network, which allows for faster inference times. To the best of our knowledge, we are the first to propose a framework for RWSR guided by the loss of a semantic segmentation network. We demonstrate the effectiveness of our proposed Semantic Segmentation Guided Real-World Super-Resolution (SSG-RWSR) on two publicly available datasets, using both real and synthesized LR images, and show that our method outperforms the existing SoTA approaches. Visual results of our method can be seen in Figure C.1. In summary, the contributions of our work are as follows:

- We propose a novel framework for RWSR which allows learning from real LR images without requiring the corresponding GT images.
- We propose to guide the learning of the RWSR task with the loss of a semantic segmentation network, which helps to reconstruct sharp and noise-free HR images.
- We show that domain adaptation and guidance by the segmentation loss is complementary to each other, and improves the texture and fine details of the reconstructed images, compared to using guidance by the segmentation loss alone.
- Our method is trained end-to-end without any manual parameter tweaking.
- We show SoTA results for RWSR on two publicly available datasets of both real and synthesized LR images.

2 Related Work

2.1 Single image super-resolution

Current SoTA methods for single image SR most often rely on deep Convolutional Neural Network (CNN) based SR architectures, which achieve impressive performance on artificially created LR images. Some of the most recent work includes EDSR [10], which is based on a deep residual CNN, the ResNet based SRResNet proposed by [2], and RCAN [11], which employs channel attention to re-scale features and recover HR details. These networks are optimized with MSE loss, which leads to good Peak Signal-to-Noise Ratio (PSNR) values, but fail to preserve the natural appearance of the images [12]. This problem is addressed in [2], which presents an SR model based on Generative Adversarial Networks (GANs), optimized with a combination of MSE, GAN, and VGG loss [13]. This approach leads to more photo-realistic images with better correlation to human perception of good image quality. In ESRGAN [14] this idea is further developed, mainly by improving the generator and adopting

a relativistic discriminator. However, the performance of the aforementioned methods degrade significantly when used on real LR images [15]. This is mainly due to the domain gap between the real and synthetic LR images. To overcome this issue, ZSSR [16] introduced a zero-shot approach which learns an image specific SR model at test time. In MSZR [8] this concept is extended to exploit information from an external dataset as well. In KernelGAN [7], ZSSR is used together with a GAN based network for estimation of image-specific blur kernels. DAN [17] proposed to address both steps in a single model using an alternating optimization algorithm that jointly estimates blur kernels and performs SR. However, these image-specific learning methods come with the cost of extremely slow prediction times compared to other SR methods [9]. In contrast, the prediction times of our method are similar to [14]. In [3], a domain adaptation based approach to RWSR is presented. First, a pool of realistic blur-kernels and noise patches is collected. These are then used to transform clean HR images into realistic LR images with similar appearance as real LR images. Next, a SR model is trained on the constructed data. However, since the domain adaptation is a challenging task in itself, the SR model is limited by the domain gap between the synthesized and real LR images. In DPSR [18], de-blurring and de-noising are combined with SR to deal with blurry and noisy LR images. However, without sufficient prior information about the image-specific degradations, the effectiveness of the method is limited.

2.2 Guided super-resolution

Lutio *et al.* [19], proposed a method for super-resolution of depth images guided by RGB images. By considering it a pixel-to-pixel transformation problem, they learn a mapping between the LR and HR images that are also applicable to the depth image. Inversely [20] proposed a zero-shot approach that extracts LR and HR patches using corresponding depth maps. Subsequently they train a GAN that employs SR- and Degredation Simulation Network (DSN)-modules in a cyclical manner that alternates between $LR \rightarrow HR \rightarrow LR$ and $HR \rightarrow LR \rightarrow HR$ mapping. In image generation tasks, such as [21–23] it has been shown that semantic information can be utilized to generate detailed textures and realistic looking images. In [24], semantic information is used to guide a SR network towards creating textures in areas where this is important, and creating sharper lines at object boundaries. Condition networks that employ SS probability maps to actively guide the SR network at a feature-map level is proposed in [25] and [26]. It is shown in [25] that the conditions can strongly influence the textures generated and result in much more realistic looking textures that are more semantically appropriate. While [24] shows that CNNs learn some categorical information, [27] propose that more categorical information can be learned by treating SR as a multi-task problem where a parallel network head that predicts a semantic map is added. The shared back-

bone is then forced to learn the categorical information necessary for accurate segmentation, which benefits the SR head. The work most closely related to ours is [28], which use multi-task learning to jointly perform SS and SR, and control the balance between SS and SR performance by adaptive weighting. However, when the SR task is given the highest weight, the performance does not benefit much from the semantic information, and drops further as more priority is given to the SS task. Furthermore, a key difference from this, and all of the existing methods utilizing semantic information for SR, is that they require paired LR and HR images for training, which makes them unsuitable for the RWSR problem. On the contrary, we show that semantic information can be leveraged to solve the RWSR problem where no GT reference images are available, making our method applicable to scenarios where real-world images, such as the ones from surveillance cameras, need to be improved by super-resolution.

2.3 Semantic segmentation

Much like in SR, SS architectures tends to follow an encoder-decoder architecture, that first encodes information with feature extraction network, typically a ResNet variant, and then decodes it again to recover spatial information and resolution. Learning to recover spatial information is difficult [29, 30], and as such SoTA SS methods have tended towards architectures that retain spatial resolution to some extent. PSPNet [29] proposed using a pyramid pooling module where the input feature-map would be pooled across different regions varying from 1×1 to 6×6 sub-regions, to get varying degrees of detail in the pooled feature-maps. They further employ 1×1 convolution to reduce the channel depth before concatenation. To recover the initial resolution lost from repeated convolution, the feature-maps are upsampled with bilinear interpolation to match the original input size. DeepLabv3 [30] proposed using atrous-convolution in the encoder to create coarse feature-representations before employing a spatial pooling pyramid to recover information at different scales. This was further expanded in [31] with depth-wise-separable convolutions resulting in the network being able to learn more fine-grained control of the details in each layer. HRNet [32] proposed an architecture that retains the spatial resolution of one branch, and parallel branches that perform further convolutions, rather than sequential repeated convolutions. Retaining the resolution with further convolutions in a parallel branch allows for the retention of fine-grained detail, while still obtaining deep representational information. However while HRNet attempts to keep a higher resolution, the initial convolutions result in an output prediction which is one-fourth of the size of the input image, which means that the prediction has to be up-sampled to compute the prediction accuracy. By super-resolving the input image, the need for up-sampling of the prediction is avoided, which leads to more accurate predic-

tions [33], which in turn improves the guiding of a SR network by the semantic loss. In [34], an auxiliary super-resolution branch is used to improve the performance on a semantic segmentation model. The SS model shares encoder weights with the SR model, which are optimized during training with MSE loss, before being removed at test time. The training process requires paired LR and HR images, and the method is therefore not applicable to real-world applications.

3 The Proposed Method

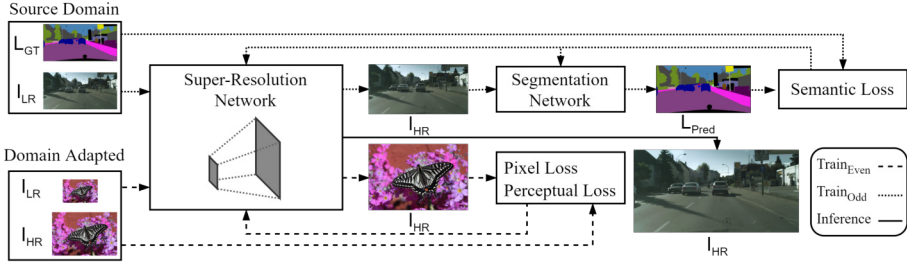


Fig. C.2: Schematic overview of our proposed SSG-RWSR. To learn to perform RWSR we leverage both guiding from an auxiliary semantic segmentation task and domain adaptation. At test time, the semantic segmentation network is de-coupled, and as such no semantic labels are required to super-resolve the LR test images.

The fundamental challenge in RWSR is the lack of real natural LR/HR image pairs which can be used to learn a SR network with supervised learning. Current RWSR methods often constrain the SR problem by assuming that the LR image is the result of an imaging model described as:

$$I_{LR} = (I_{HR} * k) \downarrow_s + n \quad (\text{C.1})$$

where k , s , and n denotes blur kernel, scaling factor, and noise, respectively. However, in reality, the image formation of real images is much more complicated.

A block diagram of our proposed SSG-RWSR framework can be seen in Figure C.2. We propose to combine domain adaptation and guiding of the SR learning by the loss of an auxiliary semantic segmentation network. The benefit of guiding the SR learning by the segmentation loss is two-fold. First, this helps our SR network to adapt to the natural image characteristics of the LR images in the source domain, without the need for GT reference images. This is important as these can be cumbersome, and sometimes even impossible to obtain. Conversely, LR images can always be annotated with semantic labels. Secondly, the loss of the segmentation task can provide strong cues about

3. The Proposed Method

the level of noise in the images, and the quality of object boundaries that can help guide the SR network towards producing more accurate reconstructions. We support the SR learning by training on image pairs created with domain adaptation. This helps our model to reconstruct realistic textures and accurate colors. During training, we alternate between training on real LR images in the source domain X , guided by SS, and LR images created by our domain adaptation approach, to leverage information from both domains. Both concepts are elaborated in the following subsections.

3.1 Guiding with semantic segmentation

We argue that a SS model can benefit from input images with low noise and high levels of detail, which can be provided by a carefully trained SR model. Hence the accuracy of a SS model can be used to guide the SR network towards producing better image quality. Based on this assumption, we structure our SSG-RWSR such that the SS network is fully dependant on the SR output. This is different from [27], where a separate semantic head is used, as we argue that for optimal guidance, the two networks should be directly linked. During training on real images, the input LR image is sequentially processed by the SR and SS networks. The SS loss is then used to optimize both the SR and SS models. This means that the SR model is getting increasingly better at producing HR images that are optimal for the segmentation task, and in addition, the SS model continuously adapts to the improved input images to further optimize the segmentation accuracy.

3.2 Domain adaptation

To ensure that our SR network learns to reconstruct HR images with realistic textures and maintain consistency with the LR input images in terms of color, we also train our SR model on paired LR/HR images. To obtain LR images with similar image characteristic as the real LR images in the source domain X , we utilize domain adaptation [35]. The procedure is elaborated in the following.

Estimation of degradation parameters We map clean HR images from the target domain Y to the real LR source domain X to minimize the domain gap between real and synthesized LR images. Our approach is based on kernel estimation and sampling of realistic noise patches [3]. For estimation of realistic blur kernels, we use KernelGAN [7], on real LR images in X to build a pool of image-specific blur kernels that can be used to degrade the clean HR images in Y .

To generate artificial LR images which are more similar to the real LR images we employ the method from [36] to sample noise from the real LR images in X . This approach assumes that realistic noise can be obtained from

an image by extracting patches from uniform areas, and then subtracting the mean. To this end, we define two patches p_i and q_j^i . p_i is obtained by a sliding window approach across images in X . Similarly q_j^i is obtained by scanning p_i . We consider p_i a uniform patch if the following constraints are met:

$$|Mean(q_j^i) - Mean(p_i)| \leq \mu \cdot Mean(p_i) \quad (C.2)$$

and

$$|Var(q_j^i) - Var(p_i)| \leq \gamma \cdot Var(p_i) \quad (C.3)$$

where $Mean$ and Var denote the mean and variance, respectively, and μ and γ are scaling factors. Different from [36] we add an additional constraint to ensure that saturated patches are not extracted:

$$Var(p_i) \geq \phi \quad (C.4)$$

where ϕ denotes a minimum variance threshold. If all constraints are satisfied p_i is considered a valid noise patch, from which we subtract the mean value and then add to a pool of noise patches n_i .

Realistic image degradation We degrade clean HR images from the target domain Y with the estimated blur kernels and noise patches following the image formation model described in Equation C.1. More specifically, we create artificial LR images I_D , by first convolving a HR image in Y with a randomly selected kernel k_i from the pool of estimated blur kernels, followed by a downsampling operation. The process can formally be described as:

$$I_D = (Y_n * k_i) \downarrow_s, i \in \{1, 2 \dots m\} \quad (C.5)$$

where I_D is the downsampled image, Y_n is a HR image, k_i refers to a kernel from the degradation pool $\{k_1, k_2, \dots k_m\}$ and s is the scaling factor.

During training of our SR network, we inject noise to the synthesized LR images by applying a randomly selected noise patch from the pool of noise patches n_i . The processes can be described as:

$$I_N = I_D + n_i, i \in \{1, 2 \dots l\} \quad (C.6)$$

where I_D is a downsampled image, and n_i is a noise patch from the noise pool $\{n_1, n_2, \dots n_l\}$.

3.3 Backbone networks

Super-resolution Our SR network consist of 23 Residual-in-Residual Dense Blocks (RRDBs) [14]. To better utilize the semantic information we use a LR patch size of 128×128 pixels. We use a combination of L1 pixel loss, \mathcal{L}_{pix} , and

4. Implementation details

Learned Perceptual Image Patch Similarity (LPIPS) loss, \mathcal{L}_{lrips} , to optimize the network when training on the domain adapted images. The L1 loss ensures color consistency between the prediction and the GT image, while LPIPS loss helps to improve the perceptual quality with strong correlation to human perception [12]. The total loss for learning the SR model from the domain adapted images is defined as:

$$\mathcal{L}_{domain-adapted} = \lambda_{pix} \cdot \mathcal{L}_{pix} + \lambda_{lrips} \cdot \mathcal{L}_{lrips} \quad (C.7)$$

where λ_{pix} , and λ_{lrips} are scaling parameters.

Semantic segmentation To maintain a high spatial resolution throughout the segmentation network we use an architecture with multiple parallel high-to-low resolution subnetworks with information exchange [32] as our SS backbone. We optimize the segmentation model with cross-entropy loss, \mathcal{L}_{ce} , which is also used for guiding the SR model. The loss for guiding the SR learning is defined as:

$$\mathcal{L}_{guided} = \lambda_{ce} \cdot \mathcal{L}_{ce} \quad (C.8)$$

where λ_{ce} is a scaling parameter.

4 Implementation details

Similar to recent RWSR literature [6, 15, 37] we perform our experiments with $\times 4$ scaling factor. For the creation of realistic training image pairs, as described in Section 3.2, we use the DF2K dataset as target domain Y of clean HR images. The DF2K is a merge of 800 and 2650 images from DIV2K [38] and Flickr2K [39], respectively.

Training details To train our SR and SS backbones, we initialize from models pre-trained on DF2K and Cityscapes, respectively. We jointly train both models, alternating between updating both models based on the cross-entropy loss, and updating only the SR model based on pixel and LPIPS loss. We denote the two update cycles as $Train_{Odd}$ and $Train_{Even}$ respectively. We use a batch size of 12 and train for 100000 iterations on randomly cropped LR patches and semantic labels using four V100 GPUs. We use the ADAM optimizer with an initial learning rate of 1×10^{-4} for both models. Through experimentation, we find suitable weights for the loss functions and set λ_{pix} , λ_{lrips} , λ_{ce} to 0.01, 0.1, and 0.01 respectively. For extraction of realistic noise patches from X , we set p_i to match the LR patch size and set q_j^i to 32, μ to 0.1, γ to 0.3, and ϕ to 0.5 which we find appropriate for real images.

Inference At test time, we de-couple the segmentation network, and as such, semantic labels are no longer required. We obtain super-resolved images by running our trained SR on the full LR input image. Hence the inference time of our SSG-RWSR is similar to [14].

5 Experiments and results

We compare our proposed method to four recent SoTA methods for SR of real images, namely MZSR [8], DPSR [18], RealSR [3], and DAN [17]. We adjust the competing models for optimal performance for a fair comparison. We use KernelGAN [7] to estimate blur kernels for use with MZSR [8]. For DPSR [18] and DAN [17], we set noise levels as recommended by the authors. With RealSR [3] we use the degradation framework provided by the authors, and re-train the model to the respective datasets. We also include the ESRGAN [14] in our comparison, to highlight the effect of applying a SR model trained on bicubically downsampled LR images on real LR images. For this, we use the pre-trained weights provided by the authors.

5.1 Datasets

Evaluation on real images For evaluation on real images we use the Cityscapes [1] and IDD [40] datasets, which both contain images and appertaining semantic labels. The Cityscapes dataset has 19 different classes and is divided into 2975 training, 500 validation, and 1525 test images, respectively, which have a resolution of 2048×1024 pixels. We use the validation set to evaluate the performance of our method. The IDD dataset has 30 different classes and contains both images of 1920×1080 and 1280×720 pixels. For our experiments, we use the 1280×720 pixels images from the training and validation set which amount to 1876 and 442 images respectively.

Evaluation on synthesized images To validate the performance of the proposed SSG-RWSR on images with known GTs, we conduct experiments on synthetically degraded LR images. This allows for evaluation with Full-Reference Image Quality Assessment (FR-IQA) metrics. To simulate realistic LR images we first degrade the images by convolving an 11×11 Gaussian blur kernel with a standard deviation of 1.5 before downsampling. Following the protocol from [15], we model sensor noise by adding Gaussian noise, with zero mean and a standard deviation of 8 pixels. This simulates real-world LR images acquired with a low-quality camera, in poor lighting conditions. For consistency, we also downsample the appertaining semantic labels. During training, only the degraded LR images and labels are available, and the degradation process and GTs are kept hidden. We perform our experiments with synthesized LR images on the Cityscapes dataset.

5.2 Quantitative Evaluation metrics

Due to the lack of GT reference images, it impossible to compare the reconstruction performance on real images with traditional SR FR-IQA metrics. As such we mainly rely on Mean Opinion Rank (MOR), which is a direct measure

5. Experiments and results

of human perceived perceptual quality [6]. We ask the participants to rank the super-resolved images based on overall image quality. We randomly shuffle the presented images to avoid bias. Readers can refer to our supplementary material for more details about our evaluation with MOR. Furthermore, we also evaluate the performance using two SoTA learning based No-Reference Image Quality Assessment (NR-IQA) methods, namely, NIMA [41] and MetaIQA [42] as these show a good correlation to human judgement. For both methods, we use the pre-trained weights for evaluation of the technical image quality.

For our experiments on synthesized LR images, we use two traditional SR metrics, PSNR and SSIM, and two perceptually oriented metrics, LPIPS [12], and DISTS [43]. Out of these, we mainly consider the LPIPS and DISTS metrics as indicators of the image quality due to their high correlation with human judgement [12]. Note that low distortion and high perceptual quality are at odds with each other, making it impossible to two obtain both [44]. With the use of GAN training and perceptual loss, our method is optimized to obtain a good trade-off with a slight bias towards perceptual quality.

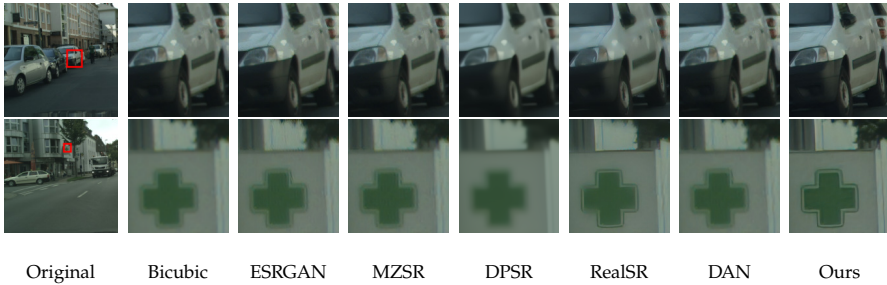


Fig. C.3: Comparison with SoTA methods for $\times 4$ SR of **real** images from the Cityscapes dataset. As visible, our method reconstructs sharper and more visually appealing results compared to the existing methods.

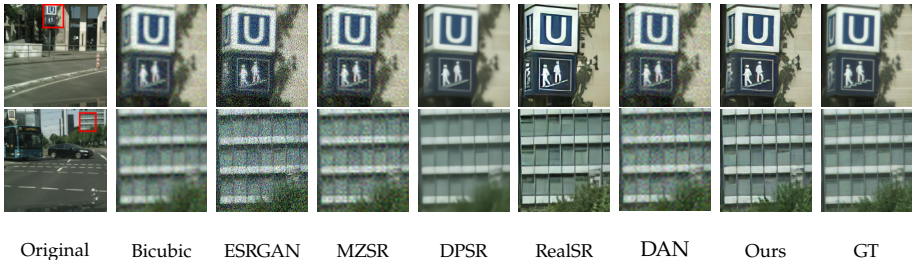


Fig. C.4: Comparison with SoTA methods for $\times 4$ of **synthetically** degraded images from the Cityscapes dataset. As visible, our method reconstructs sharp images with low noise compared to the existing methods.



Fig. C.5: Comparison with SoTA methods for $\times 4$ SR of **real** images from the IDD dataset. As visible, our method reconstructs more detailed images with less artifacts compared to the existing methods.

5.3 Qualitative results

Real images In Figure C.3 and C.5 we visualize super-resolution results of real LR images. We see that most methods fail to handle the highly complex degradation process present in the real images, which results in many artifacts (ESRGAN, MZSR, RealSR) or blurry images (DPSR, DAN). In comparison, our method generates sharper images with better visual quality and less noise.

Synthesized images In Figure C.4 we see that ESRGAN, MZSR and DAN cannot properly handle the noisy LR image which causes a high degree of artifacts to be present in the super-resolved images. DPSR performs better in that regard, but the images appear blurry and lack high-frequency details. In contrast, both RealSR and our method produces artifact-free, sharp, and natural appearing images.

5.4 Quantitative results

Cityscapes (Real LR images)			
Method	NIMA \uparrow	Meta-IQA \uparrow	MOR \downarrow
Bicubic [45]	4.62	0.245	-
ESRGAN [14]	4.95	0.247	-
MZSR [8]	4.88	0.231	3.33
DPSR [18]	4.83	0.240	4.41
RealSR [3]	4.87	0.236	2.75
DAN [17]	4.65	0.246	3.47
Ours	5.04	0.254	1.21

Table C.1: Quantitative results on the Cityscapes validation sets. \uparrow and \downarrow indicate whether higher or lower values are desired, respectively. As seen, our method obtains both the best MOR and NIMA and Meta-IQA results.

5. Experiments and results

IDD (Real LR images)			
Method	NIMA \uparrow	Meta-IQA \uparrow	MOR \downarrow
Bicubic [45]	4.73	0.330	-
ESRGAN [14]	4.94	0.325	-
MZSR [8]	5.00	0.330	2.96
DPSR [18]	4.92	0.330	3.16
RealSR [3]	4.83	0.296	4.88
DAN [17]	4.77	0.330	2.48
Ours	5.03	0.323	1.45

Table C.2: Quantitative results on the IDD validation sets. \uparrow and \downarrow indicate whether higher or lower values are desired, respectively. As seen, our method obtains both the best MOR and NIMA results, and the second best Meta-IQA results.

Cityscapes (Synthesized LR images)				
Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow
Bicubic [45]	27.51	0.62	0.64	0.19
ESRGAN [14]	18.17	0.11	1.29	0.20
MZSR [8]	26.68	0.55	0.73	0.16
DPSR [18]	33.11	0.90	0.42	0.13
RealSR [3]	25.88	0.77	0.26	0.10
DAN [17]	27.16	0.58	0.60	0.20
Ours	29.08	0.83	0.19	0.07

Table C.3: Quantitative results on the artificially degraded Cityscapes validation set. \uparrow and \downarrow indicate whether higher or lower values are desired, respectively. Our method achieves a good trade-off between low distortion and high perceptual quality with the second best PSNR and SSIM results, and the best perceptual quality as measured by the LPIPS and DISTS metrics.

Real images As show in Table C.1 and C.2 our method results in the most visually pleasing reconstructions of both real images from the CityScapes and IDD datasets according to the MOR. This is also supported by the NIMA and Meta-IQA scores, where only the DAN [17] is slightly better according to the Meta-IQA scores on the IDD dataset. However, this is in contrast to the visual appearance of the images, as the digits on the licence plates shown in Figure C.5 are more well defined in the image produced by our method, compared to the ones produced by DAN.

Synthesized images As shown in Table C.3 our method achieves a good compromise between fidelity and perceptual quality, by obtaining the best LPIPS and DISTS scores, which indicate that our super-resolved images are closer to the GT in terms of visual quality, and the second best results on the hand-crafted metrics (PSNR, SSIM). The latter is expected, as our method is

optimized towards perceptual image quality, which are at odds with a low reconstruction error [44].

5.5 Ablation study

To study the effect of the individual components in our proposed SSG-RWSR framework we compare ablations of the framework to the full system. Figure C.1 and Table D.5 shows the visual difference, and quantitative results for the different settings, respectively. As seen, training only on the synthetically created LR/HR pairs results in HR images with more high-frequency details than the LR image. However in some areas, the hallucinated details appear to be incorrect or missing. On the contrary, training only on the real LR images guided by the SS loss, produces less detailed images, but the reconstructions are more consistent with the objects and shapes present in the LR image. In comparison, our combined SSG-RWSR produces images that are both sharp, detail rich, and with a photo-realistic appearance.

Method	NIMA \uparrow	Meta-IQA \uparrow
Ours (DA)	4.33	0.206
Ours (Guided only)	5.00	0.251
Ours	5.04	0.254

Table C.4: The effect of the different components in our proposed method on the Cityscapes validation set. \uparrow and \downarrow indicate whether higher or lower values are desired, respectively.

6 Conclusion

In this paper, we address the RWSR problem where no ground truth data are available. To this end, we introduce a novel framework, SSG-RWSR, where the SR learning is guided by an auxiliary semantic segmentation network. This enables our SR model to adapt to the image specific degradations present in real LR images, and enables reconstruction of sharp object boundaries and noise-free images. We combine guidance by the segmentation loss with domain adaptation, to reconstruct realistic textures and ensure color consistency. Our experimental results on both real and synthesized LR images demonstrate a significant improvement over the SoTA methods, resulting in less noise and better visual quality. This is supported by human ranking of the super-resolved images, where our method outperforms other methods by large margins.

Disclosure of Funding This research was funded by Milestone Systems A/S, Brøndby Denmark and the Independent Research Fund Denmark, under grant number 8022-00360B.

References

- [1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [2] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [3] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, "Real-world super-resolution via kernel estimation and noise injection," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [4] R. Zhou and S. Süssstrunk, "Kernel modeling super-resolution on real low-resolution images," in *ICCV*, 2019.
- [5] A. Lugmayr, M. Danelljan, and R. Timofte, "Unsupervised learning for real-world super-resolution," in *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE, 2019, pp. 3408–3416. [Online]. Available: <https://doi.org/10.1109/ICCVW.2019.00423>
- [6] A. Lugmayr et al., "Ntire 2020 challenge on real-world image super-resolution: Methods and results," *CVPR Workshops*, 2020.
- [7] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 284–293.
- [8] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [9] Z. Wang, J. Chen, and S. Hoi, "Deep learning for image super-resolution: A survey," *TPAMI*, 2020.
- [10] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [11] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 294–310.

References

- [12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 586–595.
- [13] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 694–711. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_43
- [14] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops, L. Leal-Taixé and S. Roth, Eds.* Cham: Springer International Publishing, 2019, pp. 63–79.
- [15] A. Lugmayr, M. Danelljan, R. Timofte, M. Fritsche, S. Gu, K. Purohit, P. Kandula, M. Suin, A. N. Rajagoopalan, N. H. Joon, Y. S. Won, G. Kim, D. Kwon, C. Hsu, C. Lin, Y. Huang, X. Sun, W. Lu, J. Li, X. Gao, S. Bell-Kligler, A. Shocher, and M. Irani, "Aim 2019 challenge on real-world image super-resolution: Methods and results," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3575–3583.
- [16] M. I. Assaf Shocher, Nadav Cohen, "'zero-shot" super-resolution using deep internal learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [17] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.
- [18] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1671–1681.
- [19] R. De Lutio, S. D'Aronco, J. D. Wegner, and K. Schindler, "Guided super-resolution as pixel-to-pixel transformation," in *ICCV*, 2019.
- [20] X. Cheng, Z. Fu, and J. Yang, "Zero-shot image super-resolution with depth guided internal degradation learning," in *ECCV*, 2020.
- [21] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.
- [22] S. Zhu, R. Urtasun, S. Fidler, D. Lin, and C. Change Loy, "Be your own prada: Fashion synthesis with structural coherence," in *ICCV*, 2017.
- [23] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *ICCV*, 2017.
- [24] M. S. Rad, B. Bozorgtabar, U. Marti, M. Basler, H. K. Ekenel, and J. Thiran, "SROBB: targeted perceptual loss for single image super-resolution," in *ICCV*, 2019.

References

- [25] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *CVPR*, 2018.
- [26] L. Liu, S. Wang, and L. Wan, "Component semantic prior guided generative adversarial network for face super-resolution," *IEEE Access*, 2019.
- [27] M. S. Rad, B. Bozorgtabar, C. Musat, U. Marti, M. Basler, H. K. Ekenel, and J. Thiran, "Benefiting from multitask learning to improve single image super-resolution," *Neurocomputing*, 2020.
- [28] Q. Zhang, G. Yang, and G. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2021.
- [29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.
- [30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, 2017.
- [31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.
- [32] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2019.
- [33] A. Aakerberg, A. S. Johansen, K. Nasrollahi, and T. B. Moeslund, "Single-loss multi-task learning for improving semantic segmentation using super-resolution," in *Computer Analysis of Images and Patterns - 19th International Conference, CAIP 2021, Virtual Event, September 28-30, 2021, Proceedings, Part II*, ser. Lecture Notes in Computer Science, N. Tsapatsoulis, A. Panayides, T. Theodoridis, A. Lanitis, C. S. Pattichis, and M. Vento, Eds., vol. 13053. Springer, 2021, pp. 403–411. [Online]. Available: https://doi.org/10.1007/978-3-030-89131-2_37
- [34] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 3773–3782. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00383>
- [35] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *ICCV*, 2013.
- [36] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 3155–3164.
- [37] J. Cai, S. Gu, R. Timofte, and L. Zhang, "Ntire 2019 challenge on real image super-resolution: Methods and results," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [38] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

References

- [39] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang *et al.*, “Ntire 2017 challenge on single image super-resolution: Methods and results,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [40] G. Varma, A. Subramanian, A. M. Namboodiri, M. Chandraker, and C. V. Jawahar, “IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments,” in *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*. IEEE, 2019, pp. 1743–1751. [Online]. Available: <https://doi.org/10.1109/WACV.2019.00190>
- [41] H. T. Esfandarani and P. Milanfar, “NIMA: neural image assessment,” *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, 2018. [Online]. Available: <https://doi.org/10.1109/TIP.2018.2831899>
- [42] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, “MetaIqa: Deep meta-learning for no-reference image quality assessment,” in *CVPR*, 2020.
- [43] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *CoRR*, vol. abs/2004.07728, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07728>
- [44] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6228–6237. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html
- [45] R. G. Keys, “Cubic Convolution Interpolation for Digital Image Processing,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 29, pp. 1153–1160, Jan. 1981.

Paper D

PDA-RWSR: Pixel-Wise Degradation Adaptive Real-World Super-Resolution

Andreas Aakerberg, Majed El Helou, Kamal Nasrollahi, Sabine
Süsstrunk, and Thomas B Moeslund,

The paper has submitted for review at the
International Conference on Computer Vision (ICCV), 2023

© 2023 by the authors.

The layout has been revised.

Abstract

While many methods have been proposed to solve the Super-Resolution (SR) problem of Low-Resolution (LR) images with complex unknown degradations, their performance still drops significantly when evaluated on images with real-world degradations. One often overlooked factor contributing to this is the presence of spatially varying degradations in real LR images. To address this issue, we propose a novel degradation modeling pipeline capable of generating paired LR/High-Resolution (HR) images with spatially varying noise, a key contributor to reduced image quality. Furthermore, to fully leverage such training data, we novelly propose a Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR) framework. Specifically, we design a new Transformer-based Real-World Super-Resolution (RWSR) model capable of adapting the reconstruction process based on pixel-wise degradation features extracted by a new supervised degradation estimation model. Along with our proposed method, we also introduce a new challenging real-world Spatially Variant Super-Resolution (SVSR) benchmarking dataset, where the images are degraded by complex non-independent and identically distributed (i.i.d) noise, to evaluate the robustness of existing RWSR methods. Comprehensive experiments on synthetic and the proposed challenging real dataset demonstrates the superiority of our method over the current State-of-The-Art (SoTA).

1 Introduction

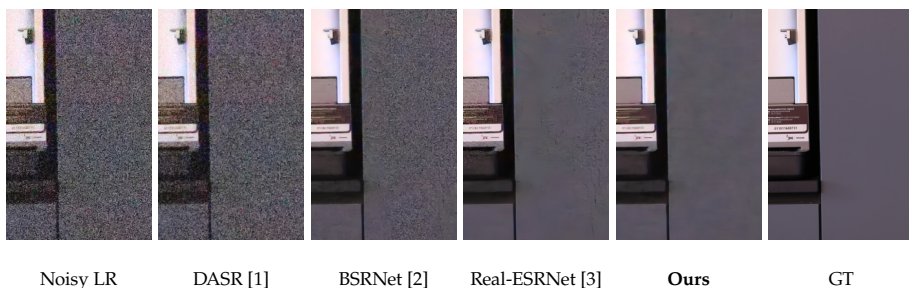


Fig. D.1: Visualization of how the assumption of uniform degradations in current State-of-The-Art (SoTA) Real-World Super-Resolution (RWSR) methods [1–3] limits the reconstruction performance, compared to our proposed pixel-wise degradation adaptive method that produces more faithful reconstructions.

Image Super-Resolution (SR) enhances the resolution and details of Low-Resolution (LR) images. Most recent SR methods accomplish this by learning a mapping from a LR image, synthetically generated by bicubic downsampling, to the corresponding High-Resolution (HR) image [4–9]. However, since Deep

Neural Network (DNN)-based methods tend to overfit to the training data distribution, this often results in poor generalization ability to real images where the degradations are much more complex. Recent attempts to improve the performance on real images include elaborate degradation models [2, 10, 11], and network conditioning based on degradation estimation [1]. Nevertheless, they all assume uniformly distributed degradations and hereby ignore the phenomenon of spatially varying noise levels present in real images. This key factor compromising the image quality occurs mainly due to the naturally varying Signal-to-Noise Ratio (SNR)-levels caused by different reflective properties of the scene. A related problem has been investigated for the task of deblurring images with spatially variant blur [12]. An example of a failure case when super-resolving images with spatially variant degradations with the current SoTA methods can be seen in Figure D.1.

We introduce a novel degradation modeling pipeline capable of introducing spatially variant degradations. Specifically, we propose a mask blending technique that synthesizes LR images with varying degrees of noise across the image to model the signal dependent noise present in real images. To fully leverage such complex training data, we also propose a Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR) framework. Specifically, our novel framework consists of a DNN that learns to extract pixel-wise degradation features from the LR image in a supervised manner, and a Transformer-based RWSR model that conditions the reconstruction process based on the pixel-wise degradation features.

Our main motivation is that while many SoTA SR methods try to address the problem of enhancing real natural LR images, they surprisingly often fail in challenging and practical applications where SR is most needed. This issue has received little attention in the research community, partly due to the lack of a sufficiently realistic and challenging real-world SR datasets that can be used for benchmarking. While datasets of real image pairs do exist, they either only consider the resolution difference [13–15], or contain noisy/clean image pairs [16, 17] without scale difference, and hereby excluding more challenging scenarios such as LR images corrupted by strong and signal dependant noise. To this end, we propose a new Spatially Variant Super-Resolution (SVSR) dataset, that contains LR images of multiple different scenes captured with varying noise levels and types, and the corresponding noise-free HR Ground-Truth (GT) images, to enable qualitative evaluation of RWSR methods in practical scenarios. We summarize our contributions as follows:

- A novel image degradation model that enables degradation at pixel level, as opposed to existing models that mostly operate on image level.
- A new Transformer-based RWSR model capable of adapting the reconstruction process based on pixel-wise degradation features extracted by a new supervised degradation estimation model.

- A novel real-world Spatially Variant Super-Resolution (SVSR) benchmarking dataset that challenges all existing SR methods.
- We highlight the importance of spatially variant degradation modeling and adaptation by demonstrating SoTA performance on the SVSR dataset with our proposed method.

2 Related Work

2.1 Single Image Super-resolution

Since the first Convolutional Neural Network (CNN) based SR network [4], a plethora of subsequent work [5, 6, 8, 9, 18, 19] have archived promising reconstruction performance on images downsampled with bicubic interpolation. Furthermore, Generative Adversarial Networks (GANs) have been used to push the SR networks to introduce realistic textures for more visually pleasing results [7, 20, 21]. However, due to the simplistic bicubic downsampling model, the classic SR methods do not generalize well to real-world scenarios [22–24]. As such, the practical applications of such methods are limited when the LR images contain complex non-uniform degradations, such as noise, blur, and compression artifacts. An overview of classic and deep-learning-based SR methods can be found in [25, 26].

2.2 Classic Blind Super-Resolution

Classic blind SR assumes that the blur kernel for the LR image is unavailable [27]. As such blind SR methods aim to enhance images beyond the bicubic degradation scenario, by including estimated blur kernel information either as a pre-processing step [28–31], or as part of the SR pipeline [32, 33].

2.3 Real-World Super-resolution

RWSR is a more practical version of blind SR, where the goal is to handle the many complex degradation types, and combinations hereof, present in real-world images. To address this, recent SoTA approaches rely on elaborate degradation models that introduce random combinations of blur and noise types, down-sampling operations, and JPEG compression artifacts [2, 3]. Other works try to estimate the average degradation in the input image and adapt the features in the SR network accordingly [1, 11, 34, 35]. FeMaSR [36] formulates the SR problem as a feature matching problem between LR features and distortion-free HR priors. Other approaches to solving the RWSR problems include [13–15, 37] that collect paired real LR and HR images for supervised learning. However, except for [37], which collect LR images with

severe under-exposure, existing SR datasets do not include challenging scenarios such as LR images with high levels of noise. Most closely related to our work, MANet [38] and KOALANet [39] perform feature modulation based on spatially varying blur kernel estimations. However, a clear distinction with our approach is the broader and more realistic degradation space, which we estimate and adapt towards on a pixel-wise level.

3 Method

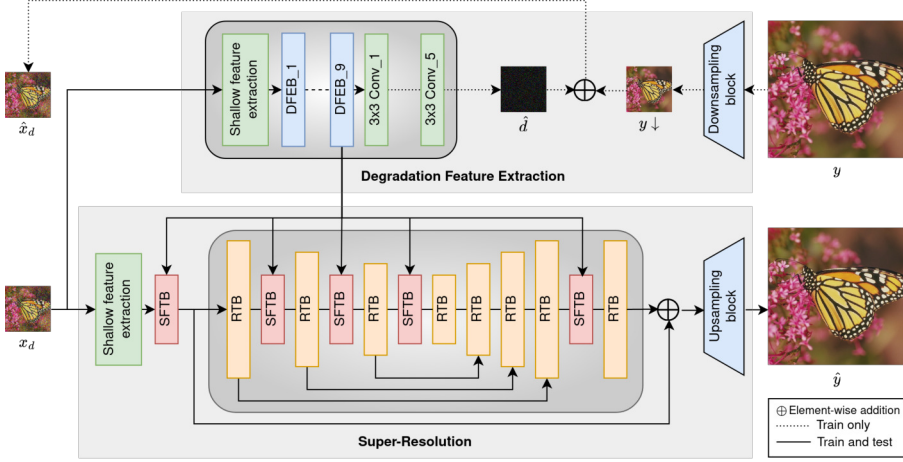


Fig. D.2: An overview of our proposed Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR). We design a Transformer-based RWSR model on the basis of Restormer Transformer Blocks (RTBs), capable of adapting the image reconstruction process based on pixel-wise degradation features via Spatial Feature Transformation Blocks (SFTBs). A supervised degradation estimation model with Degradation Feature Extraction Blocks (DFEBs) learns to separate image degradations from content, for the purpose of providing degradation features for conditioning the SR model.

We focus on the challenging task of SR of real-world LR images with complex and non-uniformly distributed degradations, a setting where current SoTA most often fails as seen in Figure D.1. Based on this observation, we design a framework to handle images with spatially variant degradations which include both pixel-wise degradation modeling, estimation and adaptation. An overview of our proposed method is presented in Figure D.2. It consists of a SR base network with Restormer Transformer Blocks (RTBs) and Spatial Feature Transformation Blocks (SFTBs), a supervised degradation estimation network with Degradation Feature Extraction Block (DFEB), and a degradation model for synthesizing LR training images with spatially variant degradations. The core novelty of our work is that the SR model is conditioned on pixel-wise

degradation features provided by the degradation estimation network for improved refinement of location-specific degradations.

3.1 Spatially Variant Degradation Model

The classic degradation pipeline for creating realistic LR/HR image pairs [40] involve convolution with a blur kernel k on the HR image y , followed by downsampling with scale factor s , and lastly degradation by additive noise to produce the degraded LR image x_d . The pipeline is formally described in Equation D.1.

$$x = (y \otimes k) \downarrow_s + n \quad (\text{D.1})$$

More elaborate and high-order degradation models for synthesis of low-quality LR images has recently been proposed by Zhang *et al.* [2], and Wang *et al.* [3] which introduce diverse combinations of degradations by a random shuffling strategy. However, we argue that a fundamental limitation of both models is the use of spatially uniform degradations, which we hypothesize limits the generalization performance to real images. Thus, we propose a novel degradation pipeline where the noise strength varies spatially across the image. This better resembles the distribution of noise in real images, which varies naturally as a result of different SNR levels [41, 42] (See also Figure D.7). More specifically, we propose to synthesize LR images with spatially varying noise with the concept of mask blending. First, we generate a mask m of the same spatial size as the LR image x , which contains either a randomly shaped and oriented gradient mask, or a mask based on the image brightness level.

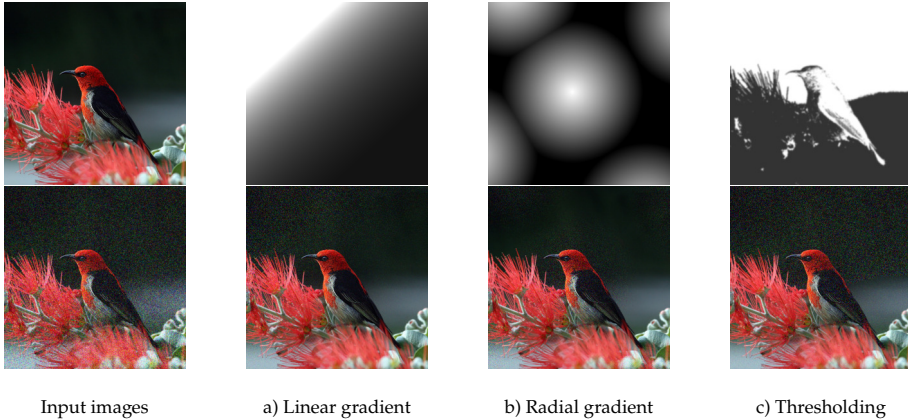


Fig. D.3: Examples of an LR image degraded by our proposed spatially variant degradation framework. Top left: clean input image. Bottom left: Input image corrupted by uniform noise. a-b: Examples of the different masks used in our framework (top row) and the corresponding output images after blending (bottom).

Next, we generate a noisy image x_n by adding spatially invariant Gaussian or Poisson noise to x . Then x and x_n are blended according to the varying intensity levels defined in the mask, to form the degraded image x_d with spatially varying noise, formally:

$$x_d = (1 - m) * x + x_n * m \quad (\text{D.2})$$

Examples of different masks and the resulting noisy images can be seen in Figure D.3. More details about the mask generation are given in the supplementary material.

3.2 Pixel-Wise Degradation Estimation

Most existing degradation estimation methods only provide a global average estimate of the degradations in the input image [1, 35]. For more fine-grained control of the reconstruction of local degradations, we propose to estimate the degradation on a pixel level. However, complex combinations of different degradations are difficult to quantify and label for supervised learning, and unsupervised learning requires elaborate frameworks with large batch sizes. As such, we propose to estimate the degradations by learning to extract them directly from a degraded image. More specifically, as shown in Figure D.2, the degradation feature extraction network D takes as input an LR image x_d , which is a degraded version of y with spatially variant degradations. In D , shallow features are first extracted by a 7×7 convolutional layer. Next, these features are further processed by 9 DFEBs to extract spatially variant degradation features. Lastly, the deep degradation features are mapped to 3-channels by four 3×3 convolutional layers to form d , which are combined with a bicubically downsampled version of y by element-wise addition to produce \hat{x}_d . The design of the DFEBs, illustrated in Figure D.4, combines a gating mechanism and depth-wise convolutions for efficient extraction of local degradation information [43]. In each DFEB, information is first processed by one 3×3 convolutional layer with LeakyReLU followed by two parallel paths through depth-wise convolutional layers, where one is activated with a ReLU non-linearity. Lastly, the two paths are combined by taking the element-wise product followed by a 1×1 convolutional layer. An additive skip connection is used to allow direct information flow from the initial convolutional layer. D is optimized by the loss between \hat{x}_d and x_d . To encourage images with similar structure and frequency distributions we use a combination of SSIM [44] and focal frequency loss [45]. The whole degradation feature extraction model has 4.6M parameters and moderate receptive field of 51×51 . During inference, we extract degradation features from the 9th DFEB for conditioning of the SR network.

3. Method

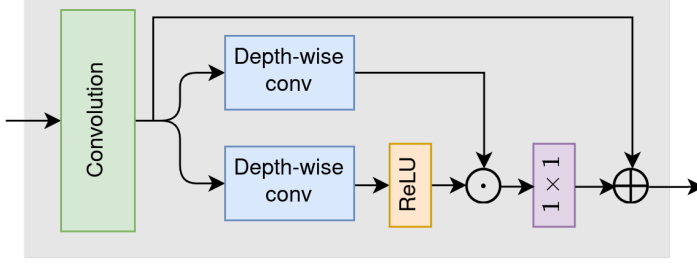


Fig. D.4: Details of the proposed Degradation Feature Extraction Block (DFEB).

3.3 Pixel-Wise Feature Modulation

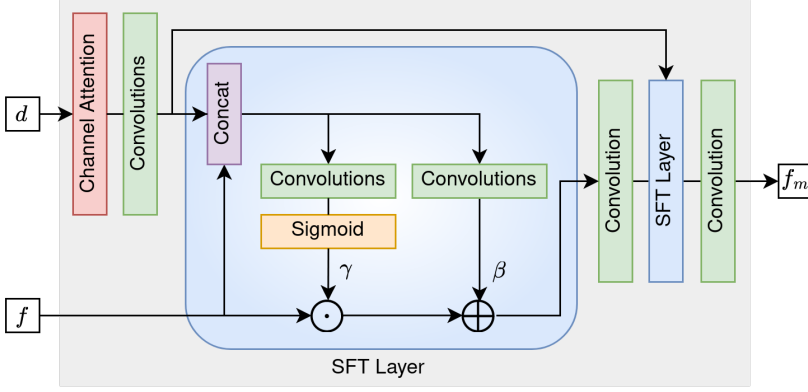


Fig. D.5: Details of the proposed Spatial Feature Transformation Block (SFTB) for adaptive conditional feature-wise and spatial-wise transformation.

To condition the SR model on the pixel-wise degradations estimated by the degradation feature extraction network, we design a feature modulation block that transforms the deep spatial features of the SR network adaptively and individually for each pixel accordingly. As visualized in Figure D.5, the Spatial Feature Transformation Block (SFTB) takes a degradation feature map d and an image feature map f of the same spatial dimensions as input. First, channel-wise attention is applied to d , followed by two convolutional layers with LeakyReLU to reduce the channel dimension from 256 to the same dimension as the feature maps in the SR network. As each SFTB shares the same degradation map, the channel attention serves to emphasize the most relevant degradation features for each part of the SR network. Next, feature transformation is performed by two Spatial Feature Transformation (SFT)-layers [46], each followed by convolutional layers, which learn parameters for a spatially affine transformation of each feature map individually. Formally,

feature maps f are conditioned on the degradation map d by a scaling and shifting operation:

$$SFT(f, d) = \gamma \odot f + \beta \quad (\text{D.3})$$

where γ and β are the scaling and shifting parameters and \odot represents the element-wise addition operation. To avoid mixing spatially adjacent degradations, the filter size of all convolutional layers in the feature transformation block are 1×1 . Furthermore, multiple separate SFTB are inserted in the SR backbone model, as the deep features propagating through the network have different sensitivity to the degradations for each level in the network.

4 SVSR Dataset

In this section, we present the data collection method for our Spatially Variant Super-Resolution (SVSR) benchmark dataset along with an analysis of the characteristics of the images. The purpose of this novel dataset is to advance the research in RWSR by enabling evaluation on real LR images with challenging and spatially variant degradations. The dataset will be released publicly upon publication.

4.1 Data Collection

The goal of the data collection is to acquire high-quality HR reference images and corresponding low-quality and LR images. For this dataset, we focus on corrupting the LR images by the noise naturally occurring in the imaging process of a digital camera. We aim for images with diverse content and static scenes, which we collect both in- and outdoors. To capture such image pairs, we use three different Canon Digital single-lens reflex (DSLR) cameras, two different zoom lenses, and three different aperture values. This ensures more diverse degradations, as the noise characteristics and point-spread-function vary between the different cameras, lenses, and aperture settings. Additional details about the cameras, setup, and examples are given in the supplementary material. The scale difference is obtained by changing the focal length of the zoom lens, by which we collect image pairs of both $\times 2$ and $\times 4$ scale difference. To obtain varying degrees of noise, we capture multiple images of the same static scene using aperture priority and changing the camera’s ISO setting. Specifically, at small ISO values (low signal gain) the camera will produce the most noise-free images, while at larger ISO values, and appropriately shorter exposure times, the images will contain more noise due to the lower signal-to-noise ratio. As such, we capture the clean images at the cameras native ISO setting (ISO100), while the noisy images are captured at incrementally higher ISO levels. The ISO level for the noisy images ranges from ISO1600, where all three cameras start to introduce visible noise, up to the maximum ISO setting

4. SVSR Dataset

for each camera. The dataset contains 978 images in total. There are 141 noise-free images for each HR scale level, and the LR scale level, while the remaining 555 images are noisy LR counterparts. A breakdown of the dataset can be seen in Table D.1. Note that due to different technologies, images captured at the same ISO setting by different cameras do not necessarily contain similar noise levels and types.

Table D.1: Overview of the different combinations of camera types and ISO settings and the resulting number of degraded LR images in the SVSR benchmarking dataset.

Camera	ISO1600	ISO3200	ISO6400	ISO12800	ISO25600	ISO51200	ISO65535
Canon 6D	✓	✓	✓	✓	✓	✓	✓
Canon 600D	✓	✓	✓	✗	✗	✗	✗
Canon 1Ds Mark II	✓	✓	✗	✗	✗	✗	✗
Total noisy LR images	141	141	93	45	45	45	45

4.2 Data Pre-processing

Even though the image collection is done with the camera mounted on a tripod and using a remote trigger, misalignment between the LR and HR image pairs can still occur, as the different focal lengths distort the image differently. To mitigate this, we design a pre-processing pipeline. First, the lens distortion is removed using Adobe Lightroom [47], followed by center cropping to keep only the sharpest part of the images. Next, we obtain pixel-wise registration of LR and HR images using a luminance-aware iterative algorithm [13], which we empirically found to be more accurate for the highly noisy images, compared to keypoint-based algorithms. To maintain the scale difference between the LR and HR images, we perform the alignment in LR space. Finally, all image pairs are examined, and ones with misalignment, out-of-focus or other unwanted defects are discarded. The resulting image pairs have a resolution of 640×640 , 1280×1280 , and 2560×2560 px for the $\times 1$, 2, and 4 scale factors, respectively.

4.3 Data Analysis

To demonstrate the spatially variant distribution of noise in the dataset, we visualize the color-channel average absolute distance between LR images of different ISO levels in Figure D.7. As seen, a larger degree of noise is present in the darker regions of the image, compared to lighter regions. Furthermore, to quantify the effect of varying ISO levels on the image quality, we compare clean and noisy images at LR scale for the different ISO values. In Table D.2 we present the average standard deviation of the noise, and the resulting change in image quality as the ISO increases. As seen, high ISO settings result in higher noise contributions, which translates to accordingly lower image quality, e.g.

the Peak Signal-to-Noise Ratio (PSNR) for the highest ISO setting is 12.53dB lower than for ISO1600. Examples of the different noise levels can be seen in Figure D.6, respectively.

Table D.2: Overview of the std. deviation σ of the noise at the different ISO levels in the SVSR benchmarking dataset, and how it affects image quality at LR scale.

ISO	σ	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow
100	0.0	∞	1.0	0.0	0.0
1600	4.93	34.32	0.9041	0.0305	0.0780
3200	6.43	32.06	0.8516	0.0711	0.1203
6400	8.06	30.18	0.8318	0.1061	0.1543
12800	9.38	28.77	0.7813	0.1403	0.1670
25600	12.31	26.37	0.6420	0.2693	0.2232
51200	15.15	24.58	0.5649	0.3364	0.2525
65535	20.87	21.79	0.4239	0.4562	0.3015

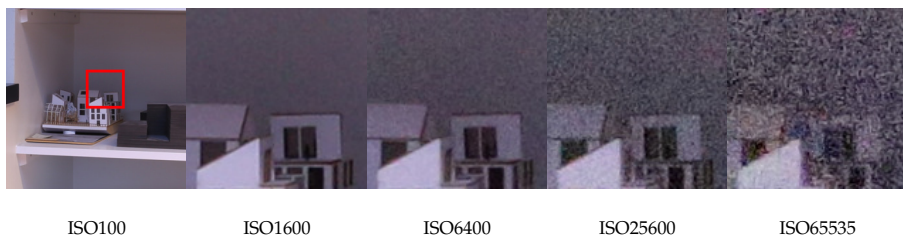


Fig. D.6: Visual examples from the SVSR benchmarking dataset illustrating how the noise level changes at different ISO settings for images captured with the Canon EOS 6D camera.

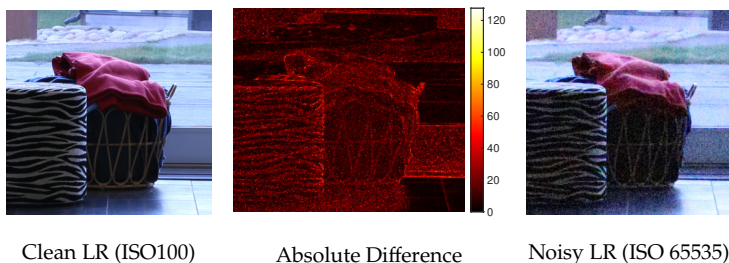


Fig. D.7: Visualization of the color-channel average absolute distance in LR space between a noisy and clean image pair from the SVSR dataset. As seen, more noise is present in the darker regions of the noisy image.

5 Experiments and Analysis

5.1 Experimental Setup

Datasets: Following recent practice in SR research [2, 3, 7, 11], we use the DIV2K [48] and Flickr2K [6] dataset for training. For evaluation on images with synthetic degradations we use Set14 [49], BSD100 [50] and Urban100 [51] which we corrupt by additive Gaussian noise with zero mean and standard deviation $\sigma = 15, 25, 50$, respectively. For evaluation on real-world degraded LR images, we use the SVSR dataset. In both cases, we experiment with $\times 4$ upsampling as commonly used in the SR literature.

Implementation Details: We use our proposed spatially variant noise degradation model together with the degradation pipeline from [2] by replacing the degradation with uniform Gaussian noise with spatially variant Gaussian and Poisson noise. Following [52], we set the noise standard deviation to [1,50] and scale to [2,4] for Gaussian and Poisson noise, respectively. The remaining steps in the degradation pipeline includes Gaussian blur, downsampling and JPEG compression noise, with the same hyperparameters as defined in [2] for comparability. As such, any performance improvements related to the degradation modeling is solely due to the introduction of spatially variant noise. We perform our experiments on a Transformer based image reconstruction backbone. Specifically, we adopt the U-Net shaped Restormer transformer network [43], where we add SFTBs for each encoder level, and before the final refinement block. We use average pooling of the degradation maps to match the spatial dimensions of the feature maps at the different encoder levels. $\times 4$ upsampling is done as final step by nearest-neighbour interpolation + convolutional layers, as commonly used in the SR literature [2, 7, 10]. Otherwise, the architecture follows the original implementation. We train our proposed degradation estimation and SR network jointly for 1M iterations with a batch size of 16 using the ADAM [53] optimizer, a learning rate of 2×10^{-4} , LR patch sizes of 64×64 , and L1-loss. Note that we do not focus on finding the optimal architecture, or training hyperparameters, but rather on showing the importance of handling the phenomenon of spatially variant degradations. As such, the performance of our proposed method can likely be further improved.

Evaluation Metrics: We evaluate the reconstruction performance using two hand-crafted (PSNR, SSIM [44]), and two SoTA DNN-based (LPIPS [54], DISTS [55]) Full-Reference Image Quality Assessment (FR-IQA) metrics. PSNR reports the image fidelity as a measure of the peak pixel-wise error between the prediction and target, while SSIM, LPIPS, and DISTS are more focused on the perceived image quality [56].

5.2 Comparison with State-of-The-Art Methods

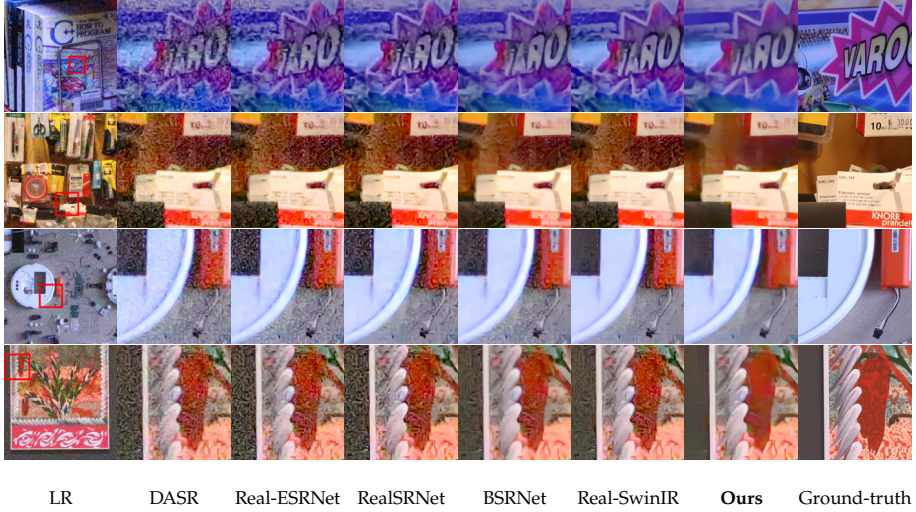


Fig. D.8: Visual comparison of the reconstruction performance on the SVSR dataset. In comparison to the SoTA approaches, our PDA-RWSR produces more visually faithful results with less artifacts.

Table D.3: Average PSNR(dB) results of state-of-the-art methods for $\times 4$ SR on synthetic noisy LR images. DN and SR indicates if the method has denoising and/or super-resolution capabilities, respectively. σ indicates the noise level.

DN	SR	Method	Set14 [49]			BSD100 [50]			Urban100 [51]		
			$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$
✓	✓	FeMaSR [36]	21.89	21.41	17.17	21.80	21.40	17.81	20.00	19.67	17.19
✓	✓	PDM-SR [57]	19.83	17.27	14.19	19.36	16.87	14.35	18.36	16.69	14.35
✓	✓	DASR [1]	24.64	23.69	20.80	24.49	23.69	21.15	22.59	21.92	19.51
✓	✓	RRDB [7]	19.84	16.48	11.75	19.81	16.42	11.74	18.96	15.98	11.64
✓	✓	DAN [33]	20.98	18.07	13.94	20.73	17.95	13.84	19.79	17.26	13.53
✓	✓	DASR [35]	23.26	21.73	17.97	23.14	21.84	18.23	21.26	20.14	17.25
✓	×	Bicubic	22.05	19.76	15.57	22.08	19.76	15.54	20.22	18.56	15.05
✓	✓	Real-ESRNet [3]	23.93	22.74	20.52	23.90	22.97	20.99	22.06	21.24	19.38
✓	×	3×3 Median + Bicubic	20.38	19.89	18.62	21.56	21.05	19.72	19.13	18.79	17.86
✓	✓	MM-RealSRNet [11]	23.41	22.69	21.03	23.68	23.04	21.56	21.38	20.90	19.62
✓	✓	BSRNet [2]	22.08	19.58	15.96	22.14	19.66	15.91	20.81	18.98	15.68
✓	✓	Real-SwinIR-L [10]	23.61	22.12	18.19	23.75	22.48	18.36	21.96	20.91	17.59
✓	✓	Ours	24.07	23.12	21.30	24.10	23.29	21.84	22.04	21.41	19.99

We compare our method with recent SoTA real-world SR methods. Specifically, we include one codebook based method (FeMaSR [36]), three degradation estimation and adaptation-based methods (DASR [1], DASR [35], DAN [33]), five methods relying on elaborate degradation modeling (Real-ESRNet [3], MM-RealSRNet [11], BSRNet [2], PDM-SR [57]) and Transformers (SwinIR [10]), and for completeness, one method trained on bicubically downsampled

5. Experiments and Analysis

Table D.4: Quantitative comparison of state-of-the-art methods for $\times 4$ SR on real noisy LR images from the SVSR benchmarking dataset. DN and SR indicate if the method has denoising and/or super-resolution capabilities, respectively.

DN	SR	Method	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow	DISTS \downarrow
✓	✓	FeMaSR [36]	0.5914	22.87	0.2772	0.1557
✓	✓	PDM-SR [57]	0.6685	23.82	0.3047	0.1805
✓	✓	DASR [1]	0.7036	24.64	0.2966	0.1741
✗	✓	RRDB [7]	0.7073	24.69	0.3011	0.1745
✓	✓	DAN [33]	0.7085	24.69	0.2997	0.1741
✓	✓	DASR [35]	0.7092	24.53	0.2478	0.1577
✗	✗	Bicubic	0.7282	24.84	0.3093	0.1717
✓	✓	Real-ESRNet [3]	0.7650	24.32	0.2063	0.1407
✓	✗	3×3 Median+Bic.	0.7690	25.08	0.2953	0.1687
✓	✓	MM-RealSRNet [11]	0.7708	24.26	0.2071	0.1501
✓	✓	BSRNet [2]	0.7844	25.13	0.2067	0.1401
✓	✓	Real-SwinIR-L [10]	0.7853	25.01	0.1956	0.1442
✓	✓	Ours	0.7943	25.16	0.1916	0.1374

images (RRDBNet [3]). For reference, we also include a filter-based method *i.e.* 3×3 Median filter followed by Bicubic upsampling. For all DNN-based methods, we use the pre-trained weights provided by the authors, for enhancement of real images and optimized for PSNR rather than perceptual quality, since our goal is to restore the original image with the highest possible fidelity.

Comparison on Synthetic Data: Table D.3 shows the results on synthetically degraded LR images. In this experiment, where the degradations are less complex, and uniformly distributed, the global average estimation and adaptation method of DASR [1] results in the best performance (noise levels 15 and 25), while our method performs second best. However, when the noise is stronger (noise level 50) our method outperforms all the competing methods.

Comparison on Real Data: Table D.4 shows the results on real LR images with complex degradations. On contrary to the experiments on synthetic data, the SVSR dataset pose a more challenging reconstruction task, where the assumption of spatially invariant Gaussian noise employed by most of the SoTA methods will not hold. As such, the global degradation estimation-based methods (DASR [1], DASR [35], DAN [33]) cannot handle such real-world scenarios, resulting in low performance based on all Image Quality Assessment (IQA) metrics. Furthermore, while methods based on elaborate degradation models (Real-ESRNet [3], MM-RealSRNet [11], BSRNet [2], PDM-SR [57], Swin-IR [10]) are trained on more complex degradations, their reconstruction quality

is very inconsistent due to the spatially variant noise in the SVSR dataset. This can be seen visually in Figure D.8, and from the plot in Figure D.9 where their performance drops sharply as the ISO level increases. On the contrary, our proposed PDA-RWSR performs better and more consistently across the range. This is also reflected in Figure D.8, where the reconstructions by our methods are more faithful with fewer artifacts, proving the superiority of PDA-RWSR for dealing with real-world degradations.

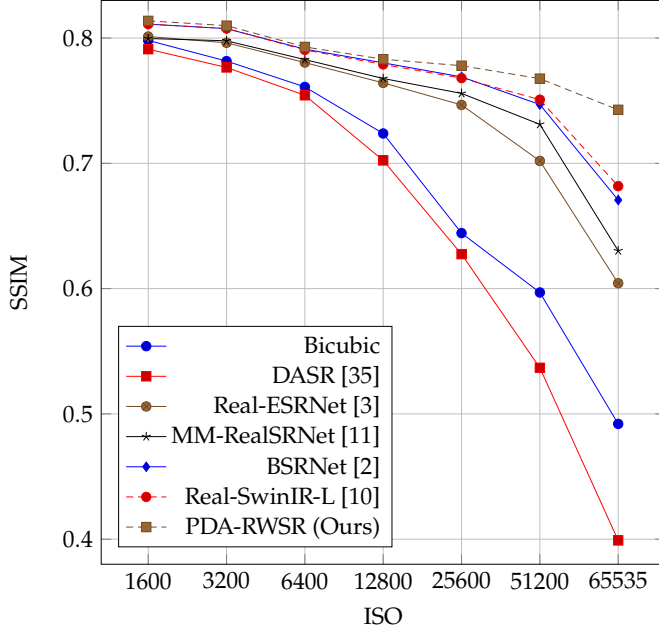


Fig. D.9: Plot of how the performance (SSIM) of SoTA methods decrease as the ISO (noise levels) in the SVSR benchmarking dataset increases. On the contrary, our PDA-RWSR has a more consistent performance across the range.

5.3 Ablation Studies

In this section, we empirically show the importance of our main technical contributions. As seen in Table D.5 incorporating our proposed degradation model with spatially variant noise (C) results in 0.09dB higher PSNR compared to using the degradation model from BSRGAN [2] (B). Our proposed per-pixel based degradation feature extraction and adaptation method (D) further improves the performance, although with the cost of additional computations.

6. Conclusion

Table D.5: Comparison of different model types. DM, and FM denotes the degradation model, and whether the model uses feature modulation, respectively. Giga Multiply-Accumulates per Second (GMACs) are computed for an input image of 64×64 pixels.

Name	DM	FM	Params $\times 10^6$	GMACs	PSNR \uparrow
A	Bicubic	\times	26.2	12.1	24.68
B	BSRGAN	\times	26.2	12.1	25.03
C	Ours	\times	26.2	12.1	25.12
D	Ours	\checkmark	28.4	51.4	25.16

6 Conclusion

In this paper, we take a step towards SR of real images with complex and spatially varying degradations. Specifically, we propose to adapt the SR reconstruction process on pixel-wise degradations. This is made possible by a novel pixel-wise degradation feature extraction network that is used to condition the SR backbone model by pixel-wise modulation blocks, and a new degradation pipeline capable of introducing spatially variant degradations to the LR training images. We further propose a new SR benchmarking dataset that challenges all the existing RWSR approaches. Our experiments on both synthetic and real LR images demonstrate that our proposed PDA-RWSR performs favorably against the current SoTA methods.

References

- [1] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, “Unsupervised degradation representation learning for blind super-resolution,” in *CVPR*, 2021.
- [2] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, “Designing a practical degradation model for deep blind image super-resolution,” in *IEEE International Conference on Computer Vision*, 2021, pp. 4791–4800.
- [3] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-esrgan: Training real-world blind super-resolution with pure synthetic data,” in *International Conference on Computer Vision Workshops (ICCVW)*.
- [4] C. Dong, C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [5] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.
- [6] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.

References

- [7] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.
- [8] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 294–310.
- [9] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 11 065–11 074.
- [10] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *IEEE International Conference on Computer Vision Workshops*, 2021.
- [11] C. Mou, Y. Wu, X. Wang, C. Dong, J. Zhang, and Y. Shan, "Metric learning based interactive modulation for real-world super-resolution," in *European Conference on Computer Vision (ECCV)*.
- [12] J. Zhang, J. Pan, J. S. J. Ren, Y. Song, L. Bao, R. W. H. Lau, and M. Yang, "Dynamic scene deblurring using spatially variant recurrent neural networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 2521–2529. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Dynamic_Scene_Deblurring_CVPR_2018_paper.html
- [13] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3086–3095.
- [14] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2020.
- [15] X. Zhang, Q. Chen, R. Ng, and V. Koltun, "Zoom to learn, learn to zoom," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 3762–3770. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Zoom_to_Learn_Learn_to_Zoom_CVPR_2019_paper.html
- [16] T. Plötz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2750–2759. [Online]. Available: <https://doi.org/10.1109/CVPR.2017.294>
- [17] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 1692–1700.

References

- [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Abdelhamed_A_High-Quality_Denoising_CVPR_2018_paper.html
- [18] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 1664–1673. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Haris_Deep_Back-Projection_Networks_CVPR_2018_paper.html
- [19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 2472–2481. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Residual_Dense_Network_CVPR_2018_paper.html
- [20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [21] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3096–3105.
- [22] A. Lugmayr, M. Danelljan, R. Timofte, M. Fritsche, S. Gu, K. Purohit, P. Kandula, M. Suin, A. N. Rajagoapalan, N. H. Joon, Y. S. Won, G. Kim, D. Kwon, C. Hsu, C. Lin, Y. Huang, X. Sun, W. Lu, J. Li, X. Gao, S. Bell-Kligler, A. Shocher, and M. Irani, "Aim 2019 challenge on real-world image super-resolution: Methods and results," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3575–3583.
- [23] P. W. et al., "AIM 2020 challenge on real image super-resolution: Methods and results," in *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, ser. Lecture Notes in Computer Science, A. Bartoli and A. Fusiello, Eds., vol. 12537. Springer, 2020, pp. 392–422. [Online]. Available: https://doi.org/10.1007/978-3-030-67070-2_24
- [24] A. Lugmayr et al., "Ntire 2020 challenge on real-world image super-resolution: Methods and results," *CVPR Workshops*, 2020.
- [25] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vision Appl.*, vol. 25, no. 6, pp. 1423–1468, Aug. 2014.
- [26] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.
- [27] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 945–952. [Online]. Available: <https://doi.org/10.1109/ICCV.2013.121>

References

- [28] M. I. Assaf Shocher, Nadav Cohen, ““zero-shot” super-resolution using deep internal learning,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [29] J. W. Soh, S. Cho, and N. I. Cho, “Meta-transfer learning for zero-shot super-resolution,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [30] K. Zhang, W. Zuo, and L. Zhang, “Deep plug-and-play super-resolution for arbitrary blur kernels,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1671–1681.
- [31] S. Bell-Kligler, A. Shocher, and M. Irani, “Blind super-resolution kernel estimation using an internal-gan,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 284–293.
- [32] J. Gu, H. Lu, W. Zuo, and C. Dong, “Blind super-resolution with iterative kernel correction,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [33] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, “Unfolding the alternating optimization for blind super resolution,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.
- [34] Y. Zhou, C. Lin, D. Luo, Y. Liu, Y. Tai, C. Wang, and M. Chen, “Joint learning content and degradation aware feature for blind super-resolution,” in *MM ’22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, Eds. ACM, 2022, pp. 2606–2616. [Online]. Available: <https://doi.org/10.1145/3503161.3547907>
- [35] J. Liang, H. Zeng, and L. Zhang, “Efficient and degradation-adaptive network for real-world image super-resolution,” in *European Conference on Computer Vision*, 2022.
- [36] C. Chen, X. Shi, Y. Qin, X. Li, X. Han, T. Yang, and S. Guo, “Real-world blind super-resolution via feature matching with implicit high-resolution priors,” 2022.
- [37] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, “RELLISUR: A real low-light image super-resolution dataset,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/7ef605fc8dba5425d6965fbd4c8fbe1f-Paper-round2.pdf>
- [38] J. Liang, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Mutual affine network for spatially variant kernel estimation in blind image super-resolution,” in *IEEE International Conference on Computer Vision*, 2021.
- [39] S. Y. Kim, H. Sim, and M. Kim, “Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 10 611–10 620.

References

- [40] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1646–1658, 1997. [Online]. Available: <https://doi.org/10.1109/83.650118>
- [41] M. E. Helou, R. Zhou, and S. Süsstrunk, "Stochastic frequency masking to improve super-resolution and denoising networks," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12361. Springer, 2020, pp. 749–766. [Online]. Available: https://doi.org/10.1007/978-3-030-58517-4_44
- [42] G. Healey and R. Kondepudy, "Radiometric CCD camera calibration and noise estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 3, pp. 267–276, 1994. [Online]. Available: <https://doi.org/10.1109/34.276126>
- [43] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," *CoRR*, vol. abs/2111.09881, 2021. [Online]. Available: <https://arxiv.org/abs/2111.09881>
- [44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004. [Online]. Available: <https://doi.org/10.1109/TIP.2003.819861>
- [45] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *ICCV*, 2021.
- [46] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 606–615. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Recovering_Realistic_Texture_CVPR_2018_paper.html
- [47] A. L. Classic, *version 10.0*. Adobe Inc., 2020.
- [48] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [49] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 711–730.
- [50] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 416–423 vol.2.
- [51] J. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5197–5206.

References

- [52] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, R. Timofte, and L. Van Gool, "Practical blind denoising via swin-conv-unet and data synthesis," *arXiv preprint*, 2022.
- [53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
- [54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 586–595.
- [55] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *CoRR*, vol. abs/2004.07728, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07728>
- [56] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 334–355.
- [57] Z. Luo, Y. Huang, , S. Li, L. Wang, and T. Tan, "Learning the degradation distribution for blind image super-resolution," in *CVPR*, 2022.

Paper E

RELLISUR: A Real Low-Light Image Super-Resolution Dataset

Andreas Aakerberg, Kamal Nasrollahi, and Thomas B Moeslund

The paper has been published in the
*Proceedings of the Neural Information Processing Systems Track on Datasets and
Benchmarks 1 (NeurIPS), 2021.*

© 2021 by the authors.

The layout has been revised.

Abstract

In this paper, we introduce RELLISUR, a novel dataset of real low-light low-resolution images paired with normal-light high-resolution reference image counterparts. With this dataset, we seek to fill the gap between low-light image enhancement and low-resolution image enhancement (Super-Resolution (SR)) which is currently only being addressed separately in the literature, even though the visibility of real-world images are often limited by both low-light and low-resolution. Part of the reason for this, is the lack of a large-scale dataset. To this end, we release a dataset with 12750 paired images of different resolutions and degrees of low-light illumination, to facilitate learning of deep-learning based models that can perform a direct mapping from degraded images with low visibility to sharp and detail rich images of high resolution. Additionally, we provide a benchmark of the existing methods for separate Low-Light Enhancement (LLE) and SR on the proposed dataset along with experiments with joint LLE and SR. The latter shows that joint processing results in more accurate reconstructions with better perceptual quality compared to sequential processing of the images. With this, we confirm that the new RELLISUR dataset can be useful for future machine learning research aimed at solving simultaneous image LLE and SR. The dataset is available at: <https://doi.org/10.5281/zenodo.5234969>.

1 Introduction

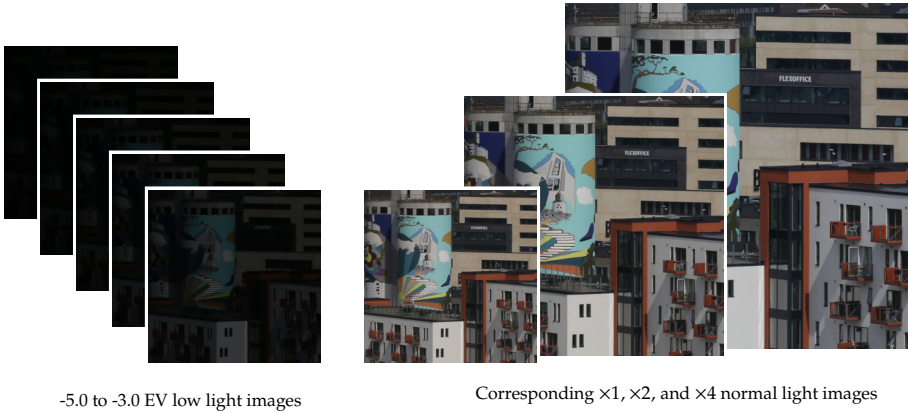


Fig. E.1: Example of a sequence of aligned images with different exposure (left) and scale levels (right) from the Real Low-Light Image Super-Resolution (RELLISUR) dataset.

Digital images can suffer from several different degradations that reduce the visibility and level of details in the images. These degradations can occur both due to environmental factors in the scene, and limitations of the hardware.

Two common degradation types are under-exposure, due to poor illumination of the scene, and low resolution, due to the limited spatial resolution of the image sensor. However, with the recent advancements in Convolutional Neural Networks (CNNs), the performance of image processing techniques, such as Low-Light Enhancement (LLE) and image Super-Resolution (SR), that can counteract these degradations have been consistently improving.

Imaging in low-light conditions is very challenging due to the low photon count, which leads to low Signal-to-Noise Ratios (SNRs). While increasing the exposure time and ISO setting will result in brighter images, this can also introduce unwanted motion blur and noise. As such, it is difficult to capture high-quality recordings at typical video frame rates in low-light conditions without using external illumination, which is not always a possibility. Simply increasing the brightness of a Low-Light (LL) image in postprocessing, will cause the artifacts introduced by the low SNR to be amplified as well. LLE is an active research field that aims to convert degraded LL images to normally exposed high-quality images. However, this is a challenging task as not only the brightness, but also more complex degradations such as color distortion and noise needs to be considered. While the resolution of digital cameras has generally increased recently, many cameras are used in combination with lenses with a wide field of view. This leaves very few pixels to resolve objects of interest, such as faces or license plates which can be critical in forensics applications. Hence, it is often desirable to increase the resolution of the Low-Resolution (LR) images to reveal more details. Image SR aims at reconstructing a High-Resolution (HR) image from its LR counterpart. However, most SR methods are trained and evaluated on datasets with synthetically created LR images, where the degradation is assumed to be an ideal bicubic downsampling kernel. This makes these methods unsuitable for real LR images where the degradation models are much more complex [1].

Even though the visibility of real images is often degraded simultaneously by multiple factors, such as low illumination and low resolution, these problems have only been addressed separately in the literature by dedicated LLE and SR methods. However, recent studies have investigated the effect of jointly performing two image processing tasks, e.g. joint LLE and deblurring [2], joint demosaicing and SR [3], and joint denoising and SR [4]. In all of these works, it was found that the joint processing outperforms sequential processing. This is mainly due to the accumulation of errors produced by the individual methods, and the possibility of early algorithms removing information that could be valuable for subsequent processing. We believe that part of the reason for why joint LLE and SR of real images has not yet been investigated in the literature, is due to the lack of a large-scale dataset of paired Low-Light Low-Resolution (LLLR) and Normal-Light High-Resolution (NLHR) images. Hence, we argue that such a dataset is of major importance in the image processing, computer vision, and machine learning community with the advent of deep-learning

based methods which performance is highly dependant on data [5]. Furthermore, it is highly desired that such a dataset consists of real-world LLLR and NLHR image pairs, as opposed to synthetic image pairs, in order to allow algorithms to generalize to practical applications. However, constructing such a dataset is a non-trivial task as real image pairs are difficult to obtain.

In this work, we present the **REal Low-Light Image SUPER-Resolution** (RELLISUR) dataset which is the first dataset to contain real LLLR and NLHR image pairs. The dataset is made publicly available and contains a large number of in- and outdoor scenes captured by a Digital single-lens reflex (DSLR) camera. There are more than 12000 image-pairs of diverse content and degradation strength in the dataset, which is more than sufficient to train Deep Neural Networks (DNNs). Applications of the dataset include remote sensing, surveillance, and forensics among others. Figure E.1 shows an example of a sequence from RELISUR containing aligned LLLR and NLHR images of the same scene.

Our collection method is reproducible and easy to follow. We collect images of different resolutions from the same static scene by changing the focal length of a zoom lens. An increasing amount of details are obtained as the focal length is increased. Along with images of different resolution, we also collect corresponding images of different low-light levels. We obtain the low-light images by shortening the exposure time. As the changing focal lengths naturally introduce misalignment between the image pairs, mainly due to varying lens distortion, we develop an effective post processing pipeline to align the image pairs. LLE or SR are both ill-posed problems, and as such, simultaneously reconstructing images degraded by both LL and LR images is a highly challenging problem. To analyze the effectiveness of RELISUR in this regard, we train and evaluate both dedicated models for each task as well as models for joint LLE and SR. The experimental results demonstrate the value of RELISUR by showing that joint processing outperforms sequential processing. Thus, we hope that the RELISUR can help facilitate further work in joint LLE and SR.

The contributions of our work are summarized as follows:

- We present the first large-scale dataset of paired and aligned low-light/low-resolution and normal-light/high-resolution images of diverse content, which closes the gap between the LLE and SR problems.
- We provide a comprehensive benchmark of existing methods for separate image LLE and SR along with experiments on joint processing on the proposed dataset.
- We show that joint image LLE and SR leads to better results than sequential processing, which highlights the need for new machine learning methods to handle the LLLR image enhancement problem.

2 Related Work

2.1 Real-world super-resolution datasets

There exist several image datasets to facilitate training and evaluation of SR methods. These include Set5 [6], Set14 [7], BSD100 [8], and DIV2K [9] among others. However, these datasets only contain the HR image, and the corresponding LR image then has to be created synthetically. The traditional way of doing this is to downsample the HR image with bicubic interpolation. As the real-world image degradation is much more complicated, SR models trained on such data often show poor performance on real LR images due to the domain difference [1, 10]. To overcome this issue, some researchers recently started to collect real LR/HR image pairs. And overview of such datasets can be seen in Table E.1. Qu et al. used a beam splitter and two cameras to collect 31 paired LR/HR face images in an indoor lab environment [11]. The City100 dataset by Chen et al. [12] consists of 100 paired images of postcards with cityscapes, captured by DSLR and smartphone cameras. Kohler et al. relied on hardware binning to capture image-pairs of different resolution [1]. The dataset contains 5670 HR images, but the variance and application to real-world scenarios are limited as the dataset only depicts 14 different indoor lab scenes acquired in grayscale. Zhang et al. collected 500 scenes of LR/HR resolution using a DSLR camera equipped with a zoom lens, which made it possible to obtain images with varying degrees of detail [13]. Images captured with a long focal length contain finer details compared to an image of the same scene captured with a short focal length. However, the images in this dataset are not pixel-wise aligned, which complicates the learning of a mapping from LR to HR. Cai et al. [14] proposed an image registration algorithm to align 243 LR/HR pairs collected with two DSLR cameras and using different focal lengths of a zoom lens. The images in the dataset depict various outdoor scenes and objects located indoors. However, a limitation of this dataset is the number of images, as there are only 175 pairs for the $\times 4$ scale. Most recently Wei et al. proposed the DRealSR dataset [15] which contains a total of 2507 LR/HR image pairs

Table E.1: Overview of real-world super-resolution datasets of paired real LR and HR images.

Name	Year	LR/HR Pairs	Type	HR resolution	Method	Content
Qu et al. [11]	2016	31	RAW	2.3MPiX	Beam-splitter	Faces
RealSR [14]	2019	595	RGB	0.48 to 5.28MPiX	Zoom lens	In/outdoor scenes
City100 [12]	2019	100	RGB	1.06MPiX	Zoom + translation	Postcards
Super [1]	2019	5,670	Grayscale	2.2MPiX	Hardware binning	Indoor lab
SR-RAW [13]	2019	500	RAW	12MPiX	Zoom lens	In/outdoor scenes
DRealSR [15]	2020	2,507	RGB	20 to 24MPiX	Zoom lens	In/outdoor scenes
Ours	2021	2,250	RGB	0.39 to 6.25MPiX	Zoom lens	In/outdoor scenes

collected with five different cameras using different zoom-lens focal lengths. However, as all of the existing real SR dataset contains image pairs where the illumination of the HR images is consistent with that of the LR images, SR models trained on such data naturally perform poorly on low-light images.

2.2 Low/normal-light datasets

Only very few datasets of paired low/normal-light images captured in real scenes exist. The LOL dataset [16] contains 500 low/normal-light image pairs which are all downsampled to a resolution of 600×400 pixels. The images are captured both in and outdoors at daylight, and the low-light images are created by changing the ISO and exposure settings of the camera, which results in LL images with low contrast, color distortion, and sensor noise due to the low SNR. Unfortunately, the downscaling of the images reduces the natural sensor noise and changes other real-world characteristics [17], such that the images can no longer be considered real LL images. The SID dataset [18] contains 5094 short exposure, and 424 long exposure RAW image pairs of either 12 and 24 MPiX resolution. All images are captured outside at nighttime or indoors in rooms with low illumination. The normal-light images are created by capturing long exposure images of the same static scenes. However, this method leads to Normal-Light (NL) images with less vibrant colors than actual daylight images and the risk of locally overexposed areas and excessive noise. In [19] a collection of HDR images along with their SDR counterparts are presented. The HDR sequence contains both under- and over-exposed images.

All the existing LLE datasets contain LL and NL image pairs of the same spatial resolution, which means that they are not feasible to use for jointly handling the LLE and SR problem. An overview of the datasets can be seen in Table E.2.

Table E.2: Overview of low-light image datasets with LL and NL pairs.

Name	Year	GT images	LL/NL Pairs	Type	Resolution	Method
LOL [16]	2018	500	500	RGB	0.24MPiX	Normal + under-exposure
SID [18]	2018	424	5,094	RAW	12/24MPiX	Under + long-exposure
SICE [19]	2018	589	4,413	RGB	6 to 24MPiX	HDR
Ours	2021	2,250	12,750	RGB	0.39 to 6.25MPiX	Normal + under-exposure

3 RELISUR Dataset

This section introduces the RELISUR dataset. We discuss in detail the data collection process, preprocessing, statistics, and present a suggested train/validation/test split.

3.1 Collection method

The RELISUR dataset is a novel collection of image sequences containing real $\times 1$, $\times 2$, and $\times 4$ NL images, together with five real LL images. The $\times 1$ and $\times 2$ scale levels represent the LR images while the $\times 4$ scale level represents the high-resolution Ground-Truth (GT) reference images. The LL images are acquired at scale $\times 1$ and are also considered low-resolution.

The dataset is collected with a Canon EOS 6D camera equipped with a Canon 70-300mm L IS USM zoom lens. Since the size of an object depicted on the image sensor is approximately linear to the focal length [20], a doubling of the scale level can be obtained by doubling the focal length. Hence, to capture images of different scale levels, we used a focal length of 70mm, 140mm, and 280mm to capture the $\times 1$, $\times 2$ and $\times 4$ scale levels, respectively.

All normal light images are captured using auto-exposure, auto-white-balance, and auto-focus using the center focus point only. The exposure metering is set to partial metering. The ISO value is set between 100 and 400 to ensure low noise levels in the NL images. To avoid misalignment issues, we aim at capturing static scenes and minimize camera movement due to wind, which is essential when using a telephoto lens. To minimize camera shake, the camera is mounted on a sturdy tripod, and hence the lens stabilization feature is disabled. To obtain a high depth-of-field we use an f-stop setting of $f/22$. The camera is triggered remotely to avoid movement.

In photography, the Exposure value (EV) is defined as $\log_2 \frac{N^2}{t}$, where N and t are the camera lens f-stop number and exposure time in seconds, respectively. Hence, a decrease of -1.0 EV corresponds to half as long exposure time, or one-stop, in our case as the f-stop is kept fixed. To capture LLLR images with different degrees of under-exposure, we used the camera’s auto bracketing mode to obtain five successive images that are under-exposed in different levels from the auto exposure setting. We used two different ranges going from from -4.5 to -2.5 and -5.0 to -3.0 EV steps. The resulting average exposure times for both the in- and outdoor scenes can be seen in Table E.3. This wide range of under exposure levels can help to improve the generalization abilities of models trained on the RELISUR dataset.

The images in the dataset are collected in natural scenes, both in- and outdoors, and depict architecture, signs, plants, common office items, art, etc. The number of in- and outdoor scenes are nearly identical with a 49% and 51% distribution, respectively. We decided not to collect images that could enable identification of individuals, by avoiding faces, persons, license plates, or other personally identifiable information. Likewise, we avoided capturing images with content that could be considered offensive, insulting, or threatening. We have manually screened the dataset to ensure that all images apply to these requirements.

In total, the RELISUR dataset consists of 850 distinct sequences. An exam-

ple of a sequence can be seen in Figure E.1. With three different scale levels, the total number of normal light LR and HR pairs is 2550. As the five underexposed images in a sequence corresponds to the same NL reference image, the resulting number of LL / NL image pairs is 4250 for each of the three scale levels. Hence, the total number of LL / NL images pairs in the RELISUR is 12750.

3.2 Preprocessing

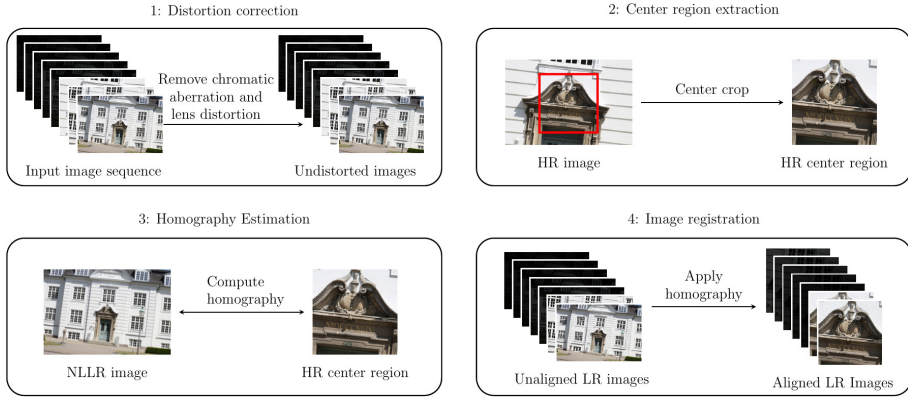


Fig. E.2: Overview of the preprocessing pipeline.

During the collection of image sequences, multiple factors can unintentionally affect the images quality negatively. First, the lens characteristics change when zooming, resulting in different levels of warping and distortion of the image. Next, external factors, such as wind, can affect the camera causing a slight shift in the scene or motion blur. To mitigate this, we apply a carefully designed preprocessing scheme to the collected images.

First, we manually screen the collected sequences and discard ones that contain images which are out-of-focus, incorrectly exposed, contain moving objects, or other undesired defects. Next, we apply lens correction in Adobe Lightroom [21] using the appropriate lens and camera profiles. This removes chromatic aberration and corrects the lens distortion. However, as the corner regions of the images are difficult to undistort, and also less sharp than the center part, we center crop the $\times 4$ NLHR reference images to the center 2500×2500 pixels. Although the images are now distortion-free, the individual images in a sequence are not guaranteed to be pixel-wise aligned due to inability to accurately adjust the zoom lens at the exact desired focal lengths. Furthermore, the optical center of the lens might shift slightly during zooming [22]. To register all images in a sequence to match the $\times 4$ NLHR reference image, we first detect and match SURF [23] features between the $\times 1$ and $\times 4$ NL images for a given

sequence. To maintain the spatial resolution difference of the three scale levels we use a downsampled version of the $\times 4$ NLHR as target. Then, we use the matched coordinates to estimate a homography using MSAC [24]. Using the translation parameters, we crop and align both the $\times 1$ LL and NL images to the $\times 4$ NLHR reference image. Lastly, we use the same method to register the $\times 2$ NL image to the $\times 4$ NLHR reference image. As such, the resolution of the $\times 1$ and $\times 2$ images become 625×625 and 1250×1250 pixels, respectively. An overview of the preprocessing pipeline can be seen in Figure E.2. One limitation of RELISUR is that the LL images are so dark that it is impossible to verify if something undesired has entered the scene, such as a bird flying by. Furthermore, changes in environmental lighting conditions can affect the brightness of the images within a sequence. Considering that this does not affect a model’s ability to learn to solve the LLE problem, we do not attempt to match the brightness levels.

Lastly, we partition the dataset into train, validation, and test splits, with a 85%/5%/10% distribution, respectively. This results in 722 train, 43 validation, and 85 test sequences. We encourage researchers to use this split to enable direct comparison with future works.

3.3 Analysis of dataset content

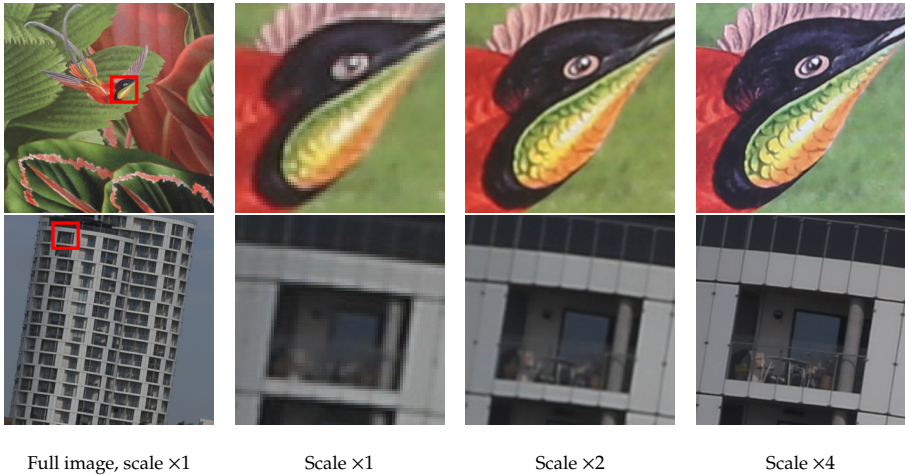


Fig. E.3: Examples of the difference in image quality between the scale levels in RELISUR. To aid visualization we show image crops.

As seen in Figure E.3, the three different scale levels of the NL images in RELISUR are all properly exposed and noise free, but have a clear difference in details and sharpness. In comparison, the LL images lack contrast and contain strong color distortion and sensor noise, as illustrated in Figure E.4.

3. RELISUR Dataset

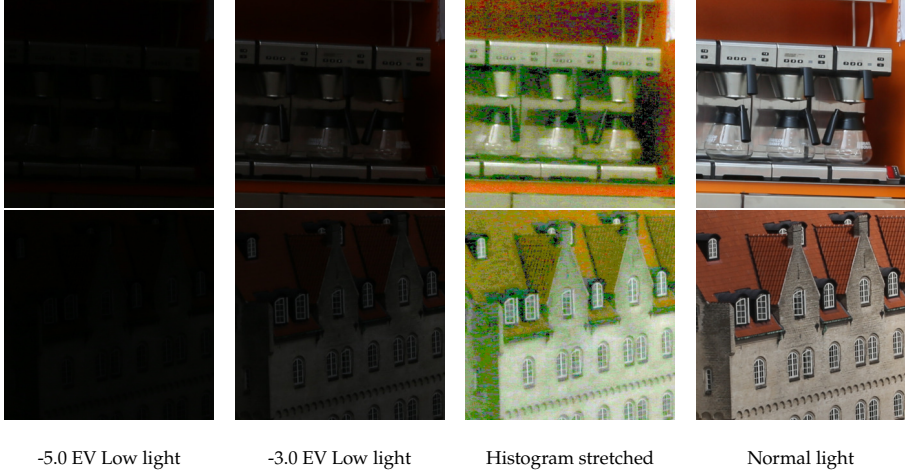


Fig. E.4: Examples of the noise and color distortion in the under-exposed images in RELISUR. To aid visualization, the -5.0 EV LL images have been histogram stretched to match the NL images

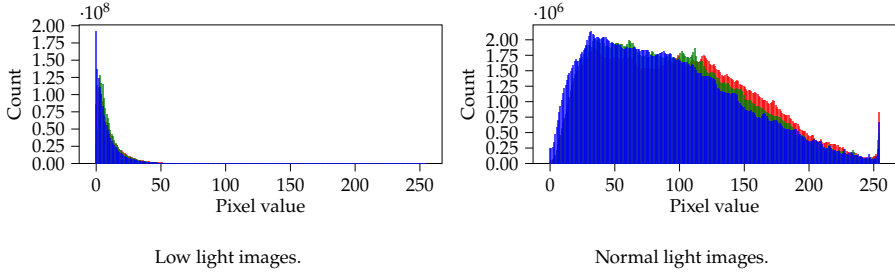


Fig. E.5: Average RGB histograms of the low light and normal light reference images in RELISUR. The horizontal axis represents the pixel value and the vertical axis the number of occurrences.

The average pixel value of the LL and NL images is shown in Figure E.5. Here it can be seen that most of the pixel values of the LL images are below 50, while the ones of the NL image are more evenly distributed across the range. This is supported by the average mean μ and standard deviation σ values computed on grayscaled versions of the images in the dataset. As seen in Table E.4, the average pixel values of the LL images in RELISUR are lower and less spread compared to the ones in the widely used LOL dataset [16], which indicates that the LLE task on RELISUR is more challenging. To quantify how the different levels of under-exposure degrades the image quality, we have computed the average Peak Signal-to-Noise Ratio (PSNR), Structural Similarity index (SSIM) [25] and LPIPS [26] quality scores for each of the different EV ranges against the properly exposed images. As seen in Table E.3, both the fidelity and perceptual quality drops significantly as the exposure time is

decreased. Furthermore, the average exposure times for indoor scenes are longer than the ones for outdoors scenes, mainly due to differences in available light.

Table E.3: Average decrease in image quality and exposure time for the different under-exposure levels in RELLISUR

Exposure	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Indoor	Outdoor
Auto	∞	∞	∞	1.722s	0.095s
-2.5 EV	10.35	0.30	0.46	0.505s	0.025s
-3.0 EV	9.40	0.22	0.57	0.212s	0.013s
-3.5 EV	8.82	0.16	0.67	0.152s	0.009s
-4.0 EV	8.42	0.11	0.76	0.107s	0.006s
-4.5 EV	8.13	0.07	0.84	0.077s	0.004s
-5.0 EV	7.87	0.05	0.89	0.031s	0.003s

Table E.4: Average mean μ and standard deviation σ values.

Name	LOL [16]	Ours
μ LL	15.48	10.59
σ LL	10.40	8.14
μ NL	116.92	96.35
σ NL	45.96	47.73

4 Experiments

We conduct several experiments on the RELLISUR dataset to evaluate its usefulness for future research on the development of machine learning models for end-to-end mapping from LLLR to NLHR. All experiments are done using the splits defined in section 3.2.

Since no publicly available methods for joint LLE and SR of real images currently exist, we first separately benchmark ten different State-of-The-Art (SoTA) LLE and SR methods by training and evaluating them on the RELLISUR dataset. Next, we select the best performing LLE and SR methods, in terms of reconstruction accuracy and perceptual quality, and combine these to sequentially process the LLLR images to obtain NLHR images. Lastly, to verify that the dataset can also be used to learn an end-to-end mapping from LLLR to NLHR, we train an SR model and an LLE model with an added upscaling module. All experiments involving SR are conducted on both scale levels in the dataset ($\times 2$ and $\times 4$).

4.1 Baseline methods for end-to-end learning

While SR models are not aimed at enhancing LL images, the ESRGAN [27] is a very capable model with more than 16 million parameters. Furthermore, this model produces HR reconstructions with the best perceptual quality of all the evaluated methods. Hence, we chose this SR model to learn the full end-to-end mapping directly from LLLR to NLHR. As LLE methods are not capable of increasing the resolution of the input images, these have to be modified in order to be able to learn the end-to-end mapping. For this we choose the MIRNet model it has the LLE performance in terms of reconstruction accuracy.

4. Experiments

To enable the MIRNet to transform LR to HR images, we add the learnable upsampling module from [28] to the end of the model. This module utilizes sub-pixel convolution [29] for efficient upsampling.

4.2 Implementation details

All supervised models have been re-trained using the hyperparameter settings described by the authors. The modified MIRNet model was trained for 1000 epochs. We used a single NVIDIA V100 card to perform the training. To evaluate the reconstruction accuracy of the different methods, we first crop 4 border pixels to avoid boundary artifacts, and calculate the average PSNR and SSIM [25] values on the test set using MATLAB [30]. While these metrics are typically used in LLE and SR research to measure the similarity to GT images, the resulting scores often correlate poorly with perceived similarity. To this end, we also include the more recent LPIPS [26] metric which has shown to correlate better with human judgment. We use the LPIPS implementation provided by the authors and used the weights from the pre-trained AlexNet [31] for evaluation.

4.3 Results

As seen in Table E.5 the best performing LLE method, according to the hand-crafted PSNR and SSIM metrics, is the MIRNet [32], while the method resulting in the best perceptual quality according to LPIPS [26] is the MBLLEN [33]. A visual comparison can be seen in Figure E.6. For the SR methods, as seen in Table E.6, the best performing models are the DBPN [34] and ESRGAN [27] in terms of fidelity and perceptual quality, respectively. A visual comparison can be seen in Figure E.7.

Regarding simultaneous LLE and SR, we see that sequential processing with the respectively best performing methods, in terms of either PSNR and LPIPS is worse than joint processing. Interestingly, the LLLR images reconstructed with the ESRGAN [27] have the best perceptual quality even though this model is not designed for LLE. At the same time the ESRGAN results in the lowest PSNR value, but this is expected due to the perception distortion tradeoff [35], since this model is optimized to produce visually pleasing images. Conversely, the MIRNet [32] model with the added upscaling module and optimized for low distortion with Charbonnier loss [36], results in the best PSNR and SSIM values. The qualitative results and examples of reconstructed images can be seen in Table F.2 and Figure E.8, respectively.

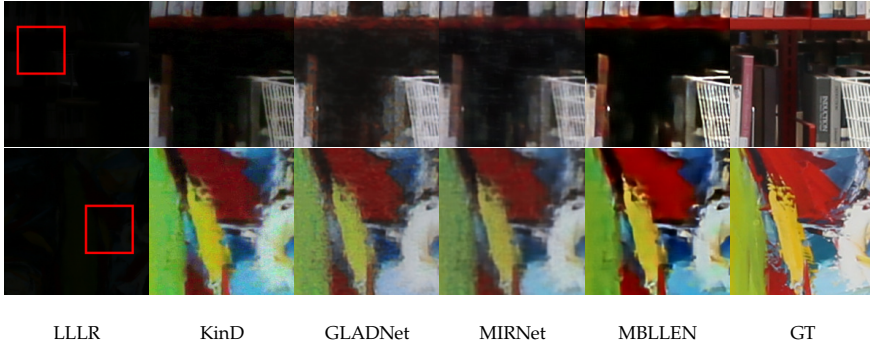


Fig. E.6: LLE results on the RELISUR test set by different methods trained on the training set.

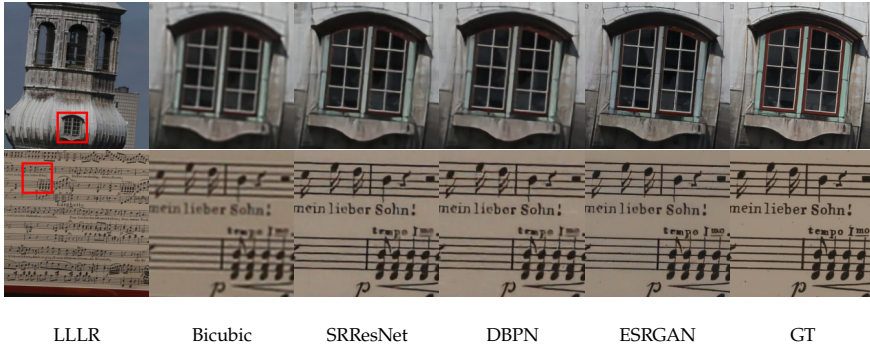


Fig. E.7: SR results (x4) on the RELISUR test set by different methods trained on the training set.

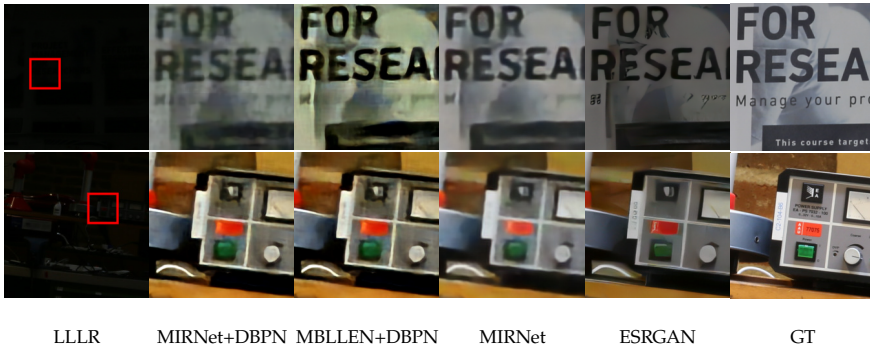


Fig. E.8: Simultaneous LLE and SR results on the RELISUR test set by different methods trained on the training set.

5. Conclusion

Table E.5: LLE results for different methods trained and tested on the RELISUR dataset.

Name	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Zero-DCE [37]	12.99	0.44	0.79
Retinex-Net [16]	15.43	0.34	0.68
LECARM [38]	10.04	0.25	0.53
RUAS [39]	11.92	0.34	0.51
LIME [40]	14.95	0.45	0.42
EnlightenGAN [41]	11.61	0.39	0.39
KinD [42]	15.84	0.49	0.33
GLADNet [43]	21.09	0.69	0.30
MIRNet [32]	21.62	0.77	0.28
MBLLEN [33]	17.52	0.60	0.23

Table E.6: SR results for different methods trained and tested on the RELISUR dataset.

Name	$\times 2$			$\times 4$		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Bicubic	28.70	0.91	0.20	23.97	0.82	0.43
SRCNN [44]	29.92	0.92	0.16	24.90	0.83	0.35
SRFBN [45]	29.78	0.92	0.16	24.77	0.84	0.33
RDN [46]	28.48	0.92	0.17	22.96	0.84	0.33
SRResNet [47]	29.82	0.92	0.15	24.52	0.84	0.32
EDSR [48]	29.69	0.92	0.16	24.06	0.85	0.32
DBPN [34]	29.99	0.92	0.15	24.98	0.84	0.30
Real-ESRGAN [49]	27.73	0.89	0.16	23.14	0.80	0.29
SRGAN [47]	29.42	0.90	0.11	24.29	0.80	0.22
ESRGAN [27]	29.79	0.91	0.10	24.71	0.80	0.21

Table E.7: Simultaneous LLE and SR results for different approaches trained and tested on the RELISUR dataset.

Type	Name	$\times 2$			$\times 4$		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Sequential	MIRNet + DBPN [32, 34]	20.73	0.73	0.49	19.85	0.74	0.58
	MIRNet + ESRGAN [27, 32]	20.67	0.72	0.47	19.81	0.71	0.56
	MBLLEN + DBPN [33, 34]	17.89	0.60	0.38	17.15	0.58	0.50
	MBLLEN + ESRGAN [27, 33]	17.74	0.56	0.40	17.03	0.50	0.52
Joint	MIRNet [32] + Upscaling module	21.33	0.75	0.41	20.62	0.75	0.53
	ESRGAN [27]	17.67	0.68	0.35	17.28	0.66	0.39

5 Conclusion

We have argued for the need for a dataset to fill the gap between LLE and SR. To this end, we have introduced the RELISUR dataset to the community, a novel

large-scale collection of paired LLLR and NLHR reference images. We offer the dataset as free and open-source with the purpose of advancing machine learning applications in the area of image processing. We also provided an extensive benchmark of the existing methods for LLE and SR, and highlighted the need for new methods to reconstruct images that are degraded by both low light and low resolution. Additionally, we have experimentally demonstrated that this dataset can be used to train deep-learning-based methods for joint LLE and SR, that outperform sequential processing. As such, we believe the RELISUR dataset will be valuable for the community.

Broader impact As this dataset contains image data that can be used to improve the performance of LLE and SR algorithms, there is a risk that malicious parties could harness this to develop more capable surveillance systems for monitoring and tracking of people. However, we have carefully screened the dataset to remove any personal information, such as persons and faces, which greatly reduce the possible negative uses of the data. On the positive side, our dataset enables reproducible research on image restoration problems which will aid in advancing these by consistent and reliable baselines.

Disclosure of Funding This research was funded by Milestone Systems A/S, Brøndby Denmark and the Independent Research Fund Denmark, under grant number 8022-00360B.

References

- [1] T. Köhler, M. Bätz, F. Naderi, A. Kaup, A. Maier, and C. Riess, “Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2944–2959, 2020. [Online]. Available: <https://doi.org/10.1109/TPAMI.2019.2917037>
- [2] Z. Liang, D. Zhang, and J. Shao, “Jointly solving deblurring and super-resolution problems with dual supervised network,” in *IEEE International Conference on Multimedia and Expo, ICME 2019, Shanghai, China, July 8-12, 2019*. IEEE, 2019, pp. 790–795. [Online]. Available: <https://doi.org/10.1109/ICME.2019.00141>
- [3] T. Klatzer, K. Hammernik, P. Knöbelreiter, and T. Pock, “Learning joint demosaicing and denoising based on sequential energy minimization,” in *2016 IEEE International Conference on Computational Photography, ICCP 2016, Evanston, IL, USA, May 13-15, 2016*. IEEE Computer Society, 2016, pp. 1–11. [Online]. Available: <https://doi.org/10.1109/ICCPHOT.2016.7492871>
- [4] R. Zhou, M. E. Helou, D. Sage, T. Laroche, A. Seitz, and S. Süsstrunk, “W2S: microscopy data with joint denoising and super-resolution for widefield to SIM mapping,” in *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, A. Bartoli and

References

- A. Fusiello, Eds., vol. 12535. Springer, 2020, pp. 474–491. [Online]. Available: https://doi.org/10.1007/978-3-030-66415-2_31
- [5] N. O. Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. A. Velasco-Hernández, L. Krpalkova, D. Riordan, and J. Walsh, “Deep learning vs. traditional computer vision,” in *Advances in Computer Vision - Proceedings of the 2019 Computer Vision Conference, CVC 2019, Las Vegas, Nevada, USA, 25-26 April 2019, Volume 1*, ser. Advances in Intelligent Systems and Computing, K. Arai and S. Kapoor, Eds., vol. 943. Springer, 2019, pp. 128–144. [Online]. Available: https://doi.org/10.1007/978-3-030-17795-9_10
- [6] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi-Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” in *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, 2012, pp. 1–10. [Online]. Available: <http://dx.doi.org/10.5244/C.26.135>
- [7] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in *Curves and Surfaces*, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 711–730.
- [8] D. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 416–423 vol.2.
- [9] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [10] A. Lugmayr et al., “Ntire 2020 challenge on real-world image super-resolution: Methods and results,” *CVPR Workshops*, 2020.
- [11] C. Qu, D. Luo, E. Monari, T. Schuchert, and J. Beyerer, “Capturing ground truth super-resolution data,” in *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*. IEEE, 2016, pp. 2812–2816. [Online]. Available: <https://doi.org/10.1109/ICIP.2016.7532872>
- [12] C. Chen, Z. Xiong, X. Tian, Z. Zha, and F. Wu, “Camera lens super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1652–1660. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Camera_Lens_Super-Resolution_CVPR_2019_paper.html
- [13] X. Zhang, Q. Chen, R. Ng, and V. Koltun, “Zoom to learn, learn to zoom,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 3762–3770. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Zoom_to_Learn_Learn_to_Zoom_CVPR_2019_paper.html
- [14] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, “Toward real-world single image super-resolution: A new benchmark and a new model,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3086–3095.

References

- [15] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2020.
- [16] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 155. [Online]. Available: <http://bmvc2018.org/contents/papers/0451.pdf>
- [17] A. Lugmayr, M. Danelljan, and R. Timofte, "Unsupervised learning for real-world super-resolution," in *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE, 2019, pp. 3408–3416. [Online]. Available: <https://doi.org/10.1109/ICCVW.2019.00423>
- [18] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 3291–3300. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Learning_to_See_CVPR_2018_paper.html
- [19] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.
- [20] E. Hecht, "Optics," Addison-Wesley, 1987.
- [21] A. L. Classic, version 10.0). Adobe Inc., 2020.
- [22] R. G. Willson and S. A. Shafer, "What is the center of the image?" in *Conference on Computer Vision and Pattern Recognition, CVPR 1993, 15-17 June, 1993, New York, NY, USA*. IEEE, 1993, pp. 670–671. [Online]. Available: <https://doi.org/10.1109/CVPR.1993.341035>
- [23] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.
- [24] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Underst.*, vol. 78, no. 1, pp. 138–156, 2000. [Online]. Available: <https://doi.org/10.1006/cviu.1999.0832>
- [25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004. [Online]. Available: <https://doi.org/10.1109/TIP.2003.819861>
- [26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 586–595.
- [27] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops, L. Leal-Taixé and S. Roth, Eds.* Cham: Springer International Publishing, 2019, pp. 63–79.

References

- [28] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 294–310.
- [29] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 1874–1883. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.207>
- [30] MATLAB, *version 9.8.0 (R2020a)*. Natick, Massachusetts: The MathWorks Inc., 2020.
- [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
- [32] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, M. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12370. Springer, 2020, pp. 492–511. [Online]. Available: https://doi.org/10.1007/978-3-030-58595-2_30
- [33] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: low-light image/video enhancement using cnns," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 220. [Online]. Available: <http://bmvc2018.org/contents/papers/0700.pdf>
- [34] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 1664–1673. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Haris_Deep_Back-Projection_Networks_CVPR_2018_paper.html
- [35] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 6228–6237. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html
- [36] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proceedings 1994 International Conference on Image Processing, Austin, Texas, USA, November 13-16, 1994*. IEEE Computer Society, 1994, pp. 168–172. [Online]. Available: <https://doi.org/10.1109/ICIP.1994.413553>
- [37] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *2020 IEEE/CVF*

References

- Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020.* IEEE, 2020, pp. 1777–1786. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00185>
- [38] Y. Ren, Z. Ying, T. H. Li, and G. Li, “LECARM: low-light image enhancement using the camera response model,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 968–981, 2019. [Online]. Available: <https://doi.org/10.1109/TCSVT.2018.2828141>
- [39] L. Risheng, M. Long, Z. Jiaao, F. Xin, and L. Zhongxuan, “Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- [40] X. Guo, Y. Li, and H. Ling, “LIME: low-light image enhancement via illumination map estimation,” *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, 2017. [Online]. Available: <https://doi.org/10.1109/TIP.2016.2639450>
- [41] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, “Enlightengan: Deep light enhancement without paired supervision,” *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021. [Online]. Available: <https://doi.org/10.1109/TIP.2021.3051462>
- [42] Y. Zhang, J. Zhang, and X. Guo, “Kindling the darkness: A practical low-light image enhancer,” in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. ACM, 2019, pp. 1632–1640. [Online]. Available: <https://doi.org/10.1145/3343031.3350926>
- [43] W. Wang, C. Wei, W. Yang, and J. Liu, “Gladnet: Low-light enhancement network with global awareness,” in *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018.* IEEE Computer Society, 2018, pp. 751–755. [Online]. Available: <https://doi.org/10.1109/FG.2018.00118>
- [44] C. Dong, C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [45] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, “Feedback network for image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019.* Computer Vision Foundation / IEEE, 2019, pp. 3867–3876. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Li_Feedback_Network_for_Image_Super-Resolution_CVPR_2019_paper.html
- [46] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018.* IEEE Computer Society, 2018, pp. 2472–2481. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Residual_Dense_Network_CVPR_2018_paper.html
- [47] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-

References

- resolution using a generative adversarial network,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [48] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [49] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-esrgan: Training real-world blind super-resolution with pure synthetic data,” in *International Conference on Computer Vision Workshops (ICCVW)*.

References

Paper F

RELIEF: Joint Low-Light Image Enhancement and Super-Resolution with Transformers

Andreas Aakerberg, Kamal Nasrollahi, and Thomas B Moeslund

The paper has been published in the
Scandinavian Conference on Image Analysis (SCIA), Lecture Notes in Computer
Science, vol. 13885. Springer, pp. 157-173, 2023.

© 2023 by the authors.

The layout has been revised.

Abstract

The goal of Single-Image Super-Resolution (SISR) is to reconstruct a High-Resolution (HR) version of a degraded Low-Resolution (LR) image. Existing Super-Resolution (SR) methods mostly assume that the LR image is a result of blurring and down-sampling the HR image, while in reality LR images are often degraded by additional factors such as low-light, low-contrast, noise, and color distortion. Due to this, current State-of-The-Art (SoTA) SR methods cannot reconstruct real low-light low-resolution images, and a straightforward strategy is, therefore, to first perform Low-Light Enhancement (LLE), followed by SR, using dedicated methods for each task. Unfortunately, this approach leads to poor performance, which motivates us to propose a method for joint LLE and SR. However, since LLE and SR are both ill-posed and ill-conditioned inverse problems, the joint reconstruction task becomes highly challenging, which calls for efficient ways to leverage as much as possible of the available information in the degraded image during reconstruction. In this paper, we propose Resolution and Light Enhancement Transformer (RELIEF), a novel Transformer-based multi-scale hierarchical encoder-decoder network with efficient cross-shaped attention mechanisms that can extract informative features from large training patches due to its strong long-range dependency modeling capabilities. This in turn leads to significant improvements in reconstruction performance on real Low-Light Low-Resolution (LLLR) images. We evaluate our method on two publicly available datasets and present SoTA results on both.

1 Introduction

Single-Image Super-Resolution (SISR) aims at increasing the spatial resolution and produce High-Resolution (HR) details given a Low-Resolution (LR) input image. Due to the many practical applications of enhancing details in images, Super-Resolution (SR) has been an active research field for decades. However, current State-of-The-Art (SoTA) SR methods are trained on well-illuminated images and they are therefore not suitable for reconstruction of real LR images captured in poor lighting conditions, e.g., by surveillance or remote sensing cameras. The conventional strategy is therefore to correct the exposure level with dedicated Low-Light Enhancement (LLE) algorithms before super-resolving the image. However, this sequential processing scheme leads to poor reconstruction accuracy mainly due to error accumulation. On the other hand, it has been shown that joint processing e.g. joint SR and denoising [1], SR and demosaicing [2], and SR and deblurring [3] leads to superior performance, compared to sequential processing. This motivates us to jointly handle the LLE and SR reconstruction problem.

Current SoTA SR methods are based on Convolutional Neural Networks (CNNs) which are typically trained on LR patches with a dimension of 64×64



Fig. F.1: Our proposed Resolution and Light Enhancement Transformer (RELIEF) can produce high-quality images from real low-light low-resolution inputs with severe noise and color distortions.

pixels and their corresponding HR patch, typically of $\times 2$, $\times 3$, or $\times 4$ times larger scale. As reconstruction of HR details are mostly a local problem, i.e. distant neighbor pixels provides little information regarding the reconstruction of the local pixel, SR models do not benefit much from using larger training patches [4, 5]. However, for the problem of LLE, the use of more global contextual information can provide valuable cues about the light enhancement level of specific pixels, see Figure F.2. Yet this has not been explored in the literature, which can possibly be explained by the ineffective long-range dependency modeling capabilities of CNNs, which limits their ability to benefit from more global contextual information.

In this paper, we propose to use Transformers to effectively utilize additional global contextual information for reconstruction of Low-Light Low-Resolution (LLLR) images, as Transformers have recently shown impressive performance on both high- and low-level vision tasks due to their high capability in modeling long-range dependencies.

A key component in Transformers is the self-attention mechanism, but due to high computational cost, and memory requirements, it is not feasible to apply full self-attention on larger images. Attempts to mitigate this problem have been made by either limiting the attention to fine-grained local self-attention [7, 8] or coarse-grained global self-attention [9, 10]. However, most approaches hinder the modeling capability of the original self-attention mechanism. Another challenge with vanilla Transformers is the lack of locality mechanisms which is essential for vision tasks. To this end, we propose a novel efficient Transformer block, Enhanced Cross-Shaped Window (ECSWin) that utilizes cross-shaped attention windows [11] together with locality enhancement in the positional encoding and feed-forward network to effectively capture long-range pixel dependencies while also leveraging local context. Our ECSWin Transformer block is used in a novel multi-scale hierarchical network, RELIEF, to perform reconstruction of real LLLR images. RELIEF can benefit from large training patch sizes due to its efficient local and global self-attention mechanism, which is applied in multiple encoder-decoders with skip-connections at different scales to aid the reconstruction process. As a result, our RELIEF is capable of achieving better visual quality and more accurate reconstructions that can help reveal information previously hidden in the LLLR images (See Figure F.1). We conduct experiments on the RELLISUR [6] and SICE [12] datasets and our empirical results show that RELIEF brings significant performance improvements over existing methods.

The contributions of our work are twofold:

- We propose a novel Transformer-based multi-scale hierarchical encoder-decoder network with an efficient cross-shaped attention mechanism for accurate reconstruction of real low-light low-resolution images. To our knowledge, RELIEF is the first method for joint LLE and SR of real LLLR images.
- We demonstrate that increased use of global information, obtained by efficient global self-attention and large training patches results in significant performance improvements on two benchmark datasets.



Fig. F.2: Example of different training patch sizes on an image from [6] (Blue: 64×64 pixels, red: 256×256 pixels). With its long-range modelling capabilities, Resolution and Light Enhancement Transformer (RELIEF) are able to utilize the information available in the larger patch, which leads to more accurate reconstructions.

2 Background

2.1 Low-light image enhancement

Low-light image enhancement has been an active research topic in the past several years resulting in a large number of methods for enhancing the light level of images. Early attempts at LLE relied on histogram equalization [13, 14], illumination map estimation [15], and Retinex theory [16, 17] to correct the image illumination. However, as these methods fail to consider the inherent noise in the Low-Light (LL) images, the reconstruction results are often unsatisfactory. Recently, deep-learning has been utilized to learn an end-to-end mapping between LL and Normal-Light (NL) images [18]. The Retinex theory was further explored in combination with deep learning in [19, 20], where a CNNs were used to learn decomposition and illumination enhancement, and most recently, a self-reinforced Retinex projection model was proposed in [21]. Furthermore, Generative Adversarial Networks (GANs) [22, 23] have also been applied to the LL image enhancement problem. Nonetheless, LLE methods do not increase the spatial resolution of the images, but mainly aim at correcting the brightness level. As such, these methods only recover limited additional details in the image.

2.2 Image Super-resolution

Like LLE, image super-resolution is one of the fundamental low-level computer vision problems [24]. From the first CNN based SR network [25], researchers have improved the reconstruction performance of the SR models by extending the network depth [26], utilizing residual learning [27, 28], applying dense connections [5, 29], and attention mechanisms [30]. Research has also been focusing on improving the perceptual quality, and not only the reconstruction accuracy, by the use of feature losses [31, 32] and GANs [5, 27, 33]. However, most approaches assume that the LR images are created by an ideal bicubic downsampling kernel, which is an oversimplification of the real-world situation [34]. Furthermore, real-world images are often degraded by additional factors besides just downsampling, e.g. blur, low-contrast, color-distortion, noise, and low-light to name a few. To remedy this, a research direction focused on SR methods that can handle more diverse degradations has emerged. These methods often improve upon classical SR methods by extending the degradation model to include more diverse degradations e.g. Gaussian noise, blur, and compression artifacts in the LR training images [35–37]. Yet, only very few works in the literature consider LR images degraded by low-light. Some of the most closely related works to our goal of SR of real natural LLLR RGB images are [38–40], which address the problem within different image-specific domains. In [40], a GAN-based method for reconstruction of synthetic

3. Method

LLLR face images is presented. In [38] a dedicated method for SR of LL Near-Infrared (NIR) images is presented, while a method for SR of LL images captured by intensified charge-coupled devices is presented in [39]. Guo *et al.* [40] experiments on synthetic LLLR face images created by gamma correction and downsampling which is another oversimplification of the complex degradation in real LLLR images. Furthermore, the method is only applicable to face images with a fixed size of 32×32 pixels. Likewise, since the method proposed by Han *et al.* [38] relies on paired NIR and visible images to enhance the NIR images, the method does not apply to SR or RGB images. The latter also applies to the method proposed by Ying *et al.* [39] since it only applies to image sensors with a proximity-focused image intensifier and requires a photon image. Therefore, as discussed above, no existing SR model has been developed for reconstructing real LLLR RGB images. Hence, with RELIEF we provide the first method to enhance the visibility, quality, and details of such images.

2.3 Vision Transformer

The Transformer was initially developed for natural language processing [41], but recently Transformers has also achieved great success in high-level vision tasks such as object detection [7, 42, 43], human pose estimation [44, 45] and semantic segmentation [7, 46, 47]. Different from CNNs, most vision Transformers decompose an image into a sequence of patches and learn long-range dependencies between each patch. Due to their promising performance, Transformers have also been studied for different low-level vision problems [48–50]. However, their potential in joint LLE and SR has not been explored in the literature. As such, we design a novel Transformer based network that proves to be highly effective for the task of reconstructing real LLLR images by joint LLE and SR.

3 Method

In this section, we describe the proposed RELIEF for joint LLE and SR starting with an overview of the overall pipeline, followed by descriptions of the individual components. Figure F.3 shows the architecture of RELIEF which is designed as a U-shaped [51] multi-scale hierarchical Transformer network.

3.1 Overall pipeline

Given an LLLR image $I_{LLLR} \in \mathbb{R}^{H \times W \times 3}$, where W and H are the width and height, respectively, our goal is to restore its Normal-Light High-Resolution (NLHR) version I_{NLHR} . To accomplish this, RELIEF first extracts low-level

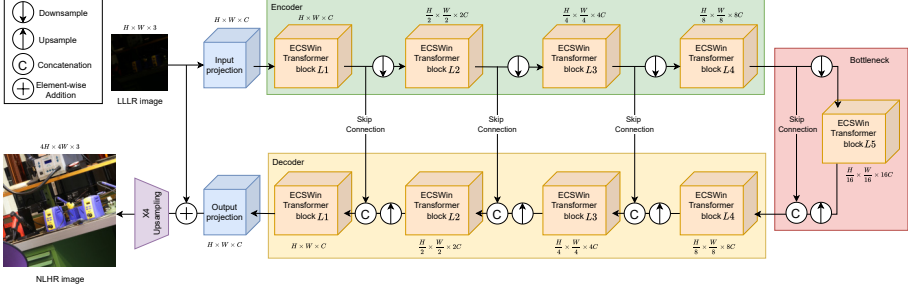


Fig. F.3: The architecture of our RELIEF for joint LLE and SR. RELIEF consists of multiple ECSWin Transformer blocks organized in a U-shaped multi-scale hierarchical network with skip-connections.

features $F_0 \in \mathbb{R}^{H \times W \times C}$, where C is the number of channels, from I_{LLR} . F_0 is obtained by a 3×3 convolutional layer with LeakyReLU. Next, deep features F_d are extracted from the low-level features F_0 in K symmetrical encoder-decoder levels. Each level contains multiple ECSWin Transformer blocks with large attention areas to capture long-range dependencies. After each encoder level, the features are reshaped to 2D feature maps and downsampled, while the number of channels is increased. We perform this operation using a 4×4 convolutional operation with stride 2. We use $K = 4$ encoder levels and as such the latent feature output at the last encoder stage is $F_l \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 8C}$ given an input feature map $F_0 \in \mathbb{R}^{H \times W \times C}$. Next, to capture even longer dependencies, we incorporate a bottleneck stage between the encoder and decoder at the lowest level. The output from the bottleneck stage is processed by a 2×2 transposed convolution operation with stride 2 to upsample the size of the latent features and reduce the channel number before entering the first decoder level. To improve the reconstruction process, skip connections are used to concatenate encoder and decoder features resulting in feature maps with twice the amount of channels. After each decoder Transformer block, the features are upsampled with a transposed convolution operation similar to the one used after the bottleneck stage. Then, at the last decoder level, the deep features F_d is reshaped using a 3×3 convolutional layer to obtain a residual image $I_R \in \mathbb{R}^{H \times W \times 3}$. Finally, the reconstructed HR and light-enhanced image is obtained as $\hat{I}_{NLHR} = (I_{LLR} + I_R) \uparrow_s$, where s is the scaling factor of the upsampling operation. The latter is performed with pixel-shuffle [52] and 3×3 convolutional operations. We optimize RELIEF with L_1 pixel loss.

3.2 ECSWin Self-Attention Transformer Block

The computational complexity of the original full self-attention mechanism grows quadratically with the input size and is therefore not feasible to use in

combination with large training image patches. Several works have tried to reduce the computational complexity by shifted [7], halo [53], and focal [54] windows to perform self-attention. However, for most methods, the effective receptive field grows slowly, which hinders the long-range modeling capability. To reduce the computational burden, while maintaining strong long-range modeling capability, we use a Cross-Shaped Window (CSWin) attention mechanism [11]. With CSWin, self-attention is calculated in horizontal and vertical stripes by splitting the multi-heads into parallel groups to achieve efficient global self-attention. We gradually increase the widths of the stripes throughout the depth of the network to further enlarge the attention area and limit the computational cost. To improve the use of local contextual information we combine the CSWin self-attention mechanism, with Locally-enhanced Feed-Forward (LeFF) and Locally-enhanced Positional Encoding (LePE) and form our ECSWin Transformer block. The different components will be described in detail in the following sections.

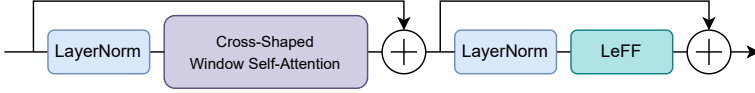


Fig. F.4: Illustration of our ECSWin Self-Attention Transformer block.

As illustrated in Figure F.4, each ECSWin Transformer block is composed of layer normalization (LN) layers [55], a CSWin self-attention module, residual connections and the LeFF layer. More formally, the ECSWin Transformer block can be defined as:

$$\begin{aligned}\hat{X}^l &= \text{CSWin-Attention} \left(\text{LN} \left(X^{l-1} \right) \right) + X^{l-1}, \\ X^l &= \text{LeFF} \left(\text{LN} \left(\hat{X}^l \right) \right) + \hat{X}^l,\end{aligned}\tag{F.1}$$

where LN represents the layer normalization [55], and \hat{X}^l and X^l are the outputs of the CSWin and LeFF modules, respectively. We design our RELIEF architecture to contain multiple CSWin Transformer blocks at each encoder-decoder level. Next, we describe the locally-enhanced feed-forward network and positional encoding in ECSWin.

3.3 Locally-enhanced Feed-Forward Network

To better utilize local context, which is essential in image restoration, we exchange the Multi-Layer Perceptron (MLP) based feed-forward network used in the vanilla Transformer block with a LeFF layer [56]. In the LeFF layer, the feature dimension of the tokens is increased with a linear projection layer and hereafter reshaped to 2D feature maps. Next, a 3×3 depth-wise convolutional

operation is applied to the reshaped feature maps. Lastly, the feature maps are flattened to tokens, and the channels are reduced with a linear layer such that the dimension of the enhanced tokens matches the dimension of the input. A Gaussian Error Linear Unit (GELU) [57] activation function is used after each linear and convolutional layer.

3.4 Locally-enhanced positional encoding

As the self-attention mechanism inherently ignores positional information in the 2D image space, we use positional encoding to add such information back. Different from the typical encoding mechanisms Absolute Positional Encoding (APE) [41], Relative Positional Encoding (RPE) [58], and Conditional Positional Encoding (CPE) [59] that adds positional information into the input tokens before the Transformer Blocks, we use LePE [11], implemented with a depth-wise convolution operator [60], to incorporate positional information within each Transformer block. Hence, the self-attention computation is formulated as:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V + \text{DWC}(V) \quad (\text{F.2})$$

where d_k is the dimension of the queries and keys and DWC is the depth-wise convolution operator. As seen in Figure F.5, LePE operates in parallel directly on V from the query (Q), key (K), and value (V) pairs obtained by a linear transformation of the input X .

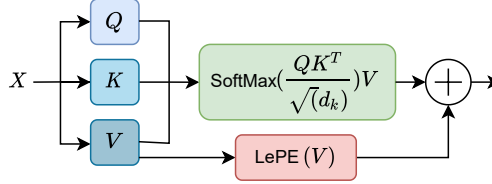


Fig. F.5: Illustration of the LePE positional encoding mechanism.

4 Experiments and Analysis

4.1 Datasets

RELLISUR

The recent RELISUR dataset [6], is the only publicly available dataset of real degraded LLLR images and their high-quality NLHR counterparts. The RELISUR dataset contains 850 distinct sequences of LLLR images, with five

4. Experiments and Analysis

different degrees of under-exposure in each sequence, paired with NLHR images of three different scale levels. In our work, we experiment with $\times 4$ upscaling which is the most challenging scale factor in the dataset. We follow the pre-defined split from [6], and as such the number of train, val, and test images are 3610, 215, and 425, respectively.

SICE

SICE [12] is a dataset of 589 various scenes captured at different exposure levels, ranging from under to overexposed including a correctly exposed Ground-Truth (GT) image. We follow the train test split defined in [12], resulting in 58 test and 531 train images. We use the GT normal-light images as is, but use only the darkest exposure of each scene as the LL image during both training and testing. We synthetically create degraded LR versions of the LL images to obtain paired degraded LLLR and clean NLHR images. The LL images is degraded by convolving with an 11×11 Gaussian blur kernel with a standard deviation of 1.5 before downsampling with factor $\times 4$. Next, we model sensor noise by adding Gaussian noise with zero mean and a standard deviation of 8. Finally, we store the images in JPEG format with a quality setting of 70 to add compression artifacts. A total of 8 images, which resolutions are less than 256×256 pixels after the downsampling, are discarded from the training set. Evaluation is performed on 256×256 center crops.

4.2 Evaluation metrics

We use two hand-crafted (PSNR, SSIM [61]) and one learning-based (DISTS [62]) Full-Reference Image Quality Assessment (FR-IQA) metrics for our quantitative comparisons. PSNR is a measure of the peak error between the reconstructed image and the GT, while SSIM is more focused on visible structure and texture differences. However, none of these metrics correlates well with the perceived image quality [63]. To this end, we use DISTS [62] which better captures the perceptual image quality as judged by human observers. Moreover, DISTS is also robust to mild geometric transformations. For all metrics, we report scores computed on the RGB channels.

4.3 Implementation details

Our RELIEF model is trained from scratch for 5×10^5 iterations with a batch size of 16 using L_1 loss. We use the ADAM optimizer [64] with a learning rate of $2e-4$ which we decrease with a factor 0.5 at 2×10^5 , 4×10^5 and 4.5×10^5 . For data augmentation, we perform rotation and horizontal and vertical flips. We use 4 encoder-decoder levels in our RELIEF implementation, with two ECSWin Transformer blocks at each level, including the bottleneck. The number of

attention heads and dimensions of the stripe widths in the encoder are set to [4,8,16,32] and [1,2,8,8], respectively, which are mirrored in the decoder. In the bottleneck, 32 heads and a stripe width of 8 are used. We use channel dimension $C = 48$ for the first encoder level in all experiments. As such, the resulting number of feature channels from level-1 to level-5 becomes [48,96,192,384,768].

4.4 Comparison with existing methods

To the best of our knowledge, no existing method in the literature can handle reconstruction of real LLLR RGB images. To this end, we compare our proposed method against dedicated methods for LLE, SR, and general image restoration. MIRNet [65] and ESRGAN [5] are SoTA methods for LLE and SR, respectively. To enable upsampling together with LLE we append a Pixel-shuffle [52] layer to MIRNet. As the VGG-discriminator in ESRGAN [5] is not compatible with large training patches, we use the patch discriminator from [66] instead. SwinIR [48] is a SoTA Transformer based method for general image restoration e.g. SR, JPEG compression artifact reduction, and denoising. We use the real-world SR configuration¹ and Pixel-shuffle upsampling for SwinIR. We re-train all competing methods using the same training hyper-parameters used for our RELIEF for a fair comparison. We use a LR training patch size of 256×256 pixels for all methods, although the performance of RELIEF can be further improved by using an even larger training patch size as shown in Section 4.6. MIRNet and SwinIR are optimized with L1 loss, while ESRGAN is optimized with a combination of L1, perceptual and adversarial loss as proposed by the authors. We emphasize that none of the above-mentioned exiting methods are designed for joint LLE and SR, but once trained on such data they can still serve as baselines against our proposed method.

Table F.1: Overview of different models and the number of parameters $\times 10^6$ and Giga Multiply-Accumulates per Second (GMACs).

Model	Parameters	GMACs
MIRNet [65] w. Pixel-shuffle [52]	31.8	51.0
ESRGAN [5]	23.2	100.0
SwinIR [48]	11.6	47.2
RELIEF	46.3	5.7

As seen in Table F.1, our RELIEF has the highest number of parameters, but a significantly lower computational burden than any of the compared methods, e.g. 5.7 vs. 47.2 GMACs for SwinIR [48]. However, as proved by empirical

¹https://github.com/cszn/KAIR/blob/master/options/swinir/train_swinir_sr_realworld_x4_psnr.json

4. Experiments and Analysis

evidence in Section 4.6, we can obtain comparable performance with a RELIEF variant with less than half the parameters. The time it takes to process an input image of 256×256 pixels with RELIEF using a RTX 3090 GPU is ≈ 59 ms.

4.5 Results

Table F.2: Quantitative comparison of state-of-the-art methods for joint LLE and $\times 4$ SR on the RELISUR and SICE datasets. Our RELIEF sets state-of-the-art results on both datasets.

Method	RELLISUR [6]			SICE [12]		
	PSNR \uparrow	SSIM \uparrow	DISTS [62] \downarrow	PSNR \uparrow	SSIM \uparrow	DISTS [62] \downarrow
MIRNet + ESRGAN [5, 65]	19.81	0.7100	0.2017	-	-	-
MIRNet [65] w. Pixel-shuffle [52]	21.04	0.7619	0.1609	18.02	0.6760	0.2749
ESRGAN [5]	17.49	0.6724	0.1518	16.44	0.6271	0.2611
SwinIR [48]	18.99	0.7478	0.1705	17.66	0.6867	0.2753
RELIEF	21.32	0.7686	0.1364	18.80	0.6980	0.2606

Quantitative results

As seen in Table F.2, sequential processing with MIRNet followed by ESRGAN, performs worse than the jointly trained MIRNet with PixelShuffle upsampling. The best performance is obtained by RELIEF which obtains gains in PSNR of 0.28 and 0.78dB on the RELISUR and SICE datasets, respectively. Similarly, our RELIEF also achieves the best perceptual quality, according to the DISTS [62] metric, even though our method is not optimized with perceptual losses like ESRGAN.

Qualitative results

We show visual comparisons of different methods on both the RELISUR and SICE datasets in Figure F.6 and Figure F.7. As seen, our RELIEF also shows its clear advantages against the other methods, by producing the most visually pleasing reconstructions with the lowest amount of artifacts. In the RELISUR dataset, there are severe noise and color distortions hiding in the extremely low-light low-resolution images, which methods like MIRNet and ESRGAN struggle to remove. In comparison, SwinIR produces fewer artifacts, but our RELIEF reconstructs images with the most accurate colors and the least artifacts while preserving most of the structural content. This is especially noticeable in Figure F.6 second and third row, where our method is the only one that manages to reconstruct a uniform and clean background as intended, without compromising edges and fine details. The same trend can be observed with the visual results from the SICE dataset, where images produced by MIRNet and ESRGAN contain severe visual defects, while our method is more faithful



Fig. F.6: Visual comparison for joint LLE and $\times 4$ SR on the RELLISUR [6] dataset. Compared to the other approaches, our RELIEF produces more visually faithful results with less artifacts.

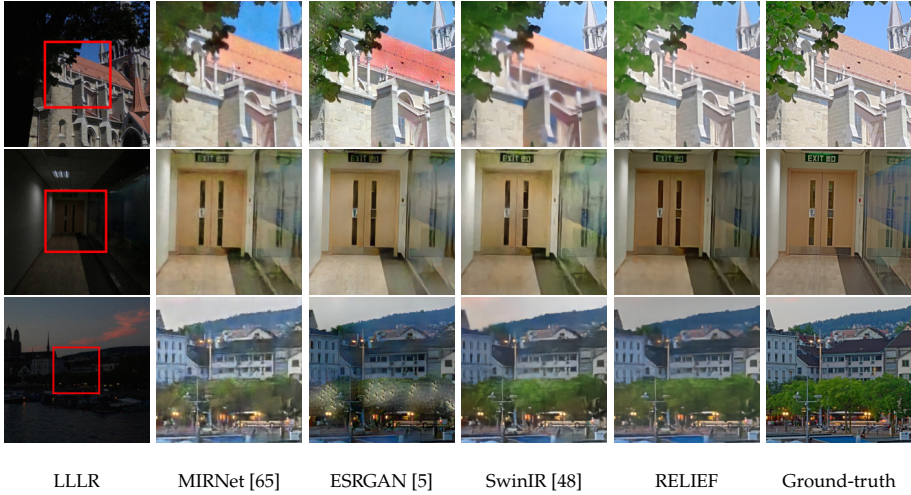


Fig. F.7: Visual comparison for joint LLE and $\times 4$ SR on the SICE [12] dataset. Our RELIEF is better at restoring the correct colors and removing undesirable artifacts.

to the ground-truth. The main difference between SwinIR and our method is that the reconstructions produced by our method appear much sharper and with less color distortions.

4.6 Ablation Studies

In this section, we investigate the effectiveness and necessity of the components in RELIEF. All evaluations are conducted on RELISUR [6] using a LR training patch size of 64×64 and a channel dimension $C = 48$, unless otherwise stated.

Training patch size

We study how the size of the LR training patch affects the reconstruction performance. As seen in Table F.3, larger patch sizes result in increased reconstruction accuracy. A significant improvement in reconstruction accuracy of 2.24dB is obtained by using a patch size of 384×384 pixels instead of 64×64 pixels. The improvement can also be confirmed visually from the results shown in Figure F.8, where it can be seen that a larger patch size contributes to more details in the reconstructions (Figure F.8, top row), while also ensuring that smooth regions appear more uniform and with fewer artifacts (Figure F.8, bottom row). Based on this we conclude that our RELIEF is effective in terms of leveraging more global contextual information for joint LLE and SR.

Table F.3: Ablation on training patch sizes.

LR patch size	PSNR \uparrow	SSIM \uparrow	DISTS [62] \downarrow
64×64	19.78	0.7430	0.1917
128×128	20.25	0.7491	0.1716
256×256	21.32	0.7686	0.1364
384×384	22.02	0.7790	0.1268

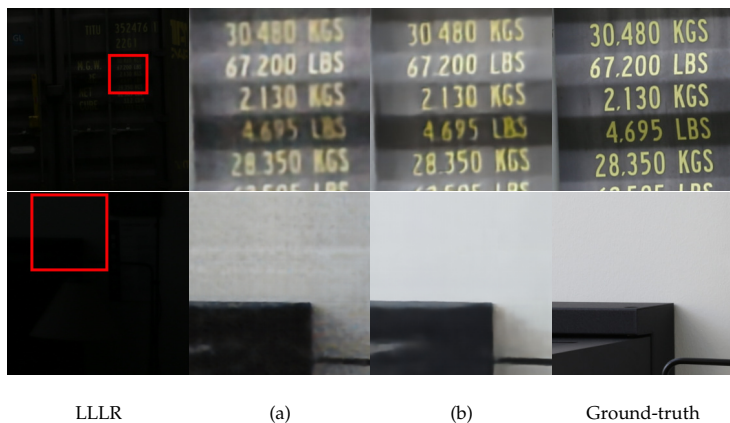


Fig. F.8: Visual effect of training RELIEF with different LR patch sizes, (a) 64×64 pixels and (b) 256×256 pixels (samples from the RELISUR [6] dataset).

Impact of skip connections and bottleneck layer

Table F.4 shows three variants of our network: RELIEF, RELIEF without skip-connections, and RELIEF without the bottleneck layer. From the table, it can be seen that the skip connections and bottleneck layer are both important components as the PSNR drops by 0.64 and 0.58dB by removal of these network components, respectively.

Table F.4: Ablation on different network designs.

Design	w/o skip-conn.	w/o bottleneck	RELIEF
PSNR \uparrow	19.14	19.20	19.78

Attention and locality

We compare different multi-headed self-attention mechanisms, feed-forward networks, and positional-encoding mechanisms for the Transformer blocks in RELIEF to show the effect on the reconstruction performance. As seen in Table F.5, the best performing configuration with cross-shaped window attention, and enhanced locality in the feed-forward network and positional-embedding yields 0.97dB improvement over the configuration with shifted-window attention [7], MLP feed-forward network and relative-positional encoding [58] without locality enhancement. Compared to CSWin, our ECSWin block with locality enhanced feed-forward network results in 0.15dB PSNR gain.

Table F.5: Ablation on different multi-headed self-attention mechanisms, feed-forward networks, and positional-encoding mechanisms. \dagger is the result of our ECSWin Transformer block.

MSA	FFN		PE		PSNR \uparrow
	MLP	LeFF [56]	RPE [58]	LePE [11]	
Swin	✓	-	✓	-	18.81
CSWin	✓	-	-	✓	19.49
	-	✓	-	-	19.63
	-	✓	-	✓	19.78 \dagger

Model parameters

We experiment with different amounts of model parameters to find a trade-off between accuracy and complexity by varying the channel number C . As shown in Table F.6, we design three variants of RELIEF: RELIEF_S, RELIEF_M, and

RELIEF_L. We observe that the PSNR is correlated with the number parameters, but also that the parameters and GMACs grow quadratically. We choose a channel number of 48 to balance performance and model size.

Table F.6: Comparison of different channel dimensions and the resulting number of model parameters, GMACs and reconstruction accuracy.

Model	C	Parameters $\times 10^6$	GMACs	PSNR \uparrow
RELIEF_S	32	20.6	2.59	19.68
RELIEF_M	48	46.3	5.74	19.78
RELIEF_L	64	82.1	10.13	19.80

5 Conclusion

In this paper, we introduced RELIEF, a novel U-shaped multi-scale hierarchical Transformer network for joint LLE and SR of real LLLR images. With its efficient ECSWin Transformer blocks, capable of capturing long-range dependencies and utilizing local context, RELIEF can benefit from large training patches which leads to better reconstruction performance. As such, RELIEF is capable of revealing details previously hidden in the dark while also removing undesired artifacts. Experimental results on two benchmark datasets show that RELIEF outperforms the state-of-the-art methods in terms of both reconstruction accuracy and visual quality. In the future, we plan to explore our RELIEF architecture for other image reconstruction tasks.

References

- [1] R. Zhou, M. E. Helou, D. Sage, T. Laroche, A. Seitz, and S. Süsstrunk, “W2S: microscopy data with joint denoising and super-resolution for widefield to SIM mapping,” in *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, A. Bartoli and A. Fusiello, Eds., vol. 12535. Springer, 2020, pp. 474–491. [Online]. Available: https://doi.org/10.1007/978-3-030-66415-2_31
- [2] T. Klatzer, K. Hammernik, P. Knöbelreiter, and T. Pock, “Learning joint demosaicing and denoising based on sequential energy minimization,” in *2016 IEEE International Conference on Computational Photography, ICCP 2016, Evanston, IL, USA, May 13-15, 2016*. IEEE Computer Society, 2016, pp. 1–11. [Online]. Available: <https://doi.org/10.1109/ICCPHOT.2016.7492871>
- [3] Z. Liang, D. Zhang, and J. Shao, “Jointly solving deblurring and super-resolution problems with dual supervised network,” in *IEEE International Conference on*

References

- Multimedia and Expo, ICME 2019, Shanghai, China, July 8-12, 2019.* IEEE, 2019, pp. 790–795. [Online]. Available: <https://doi.org/10.1109/ICME.2019.00141>
- [4] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, “EnhanceNet: Single image super-resolution through automated texture synthesis,” in *Proceedings IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ, USA: IEEE, Oct. 2017, pp. 4501–4510. [Online]. Available: http://openaccess.thecvf.com/content_ICCV_2017/papers/Sajjadi_EnhanceNet_Single_Image_ICCV_2017_paper.pdf
- [5] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, “Esr-gan: Enhanced super-resolution generative adversarial networks,” in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.
- [6] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, “RELLISUR: A real low-light image super-resolution dataset,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/7ef605fc8dba5425d6965fbd4c8fbe1f-Paper-round2.pdf>
- [7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: <https://arxiv.org/abs/2103.14030>
- [8] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, “Multi-scale vision longformer: A new vision transformer for high-resolution image encoding,” *CoRR*, vol. abs/2103.15358, 2021. [Online]. Available: <https://arxiv.org/abs/2103.15358>
- [9] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” *CoRR*, vol. abs/2102.12122, 2021. [Online]. Available: <https://arxiv.org/abs/2102.12122>
- [10] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, “Cvt: Introducing convolutions to vision transformers,” *CoRR*, vol. abs/2103.15808, 2021. [Online]. Available: <https://arxiv.org/abs/2103.15808>
- [11] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, “Cswin transformer: A general vision transformer backbone with cross-shaped windows,” *CoRR*, vol. abs/2107.00652, 2021. [Online]. Available: <https://arxiv.org/abs/2107.00652>
- [12] J. Cai, S. Gu, and L. Zhang, “Learning a deep single image contrast enhancer from multi-exposure images,” *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.
- [13] J. A. Stark, “Adaptive image contrast enhancement using generalizations of histogram equalization,” *IEEE Trans. Image Process.*, vol. 9, no. 5, pp. 889–896, 2000. [Online]. Available: <https://doi.org/10.1109/83.841534>
- [14] D. Coltuc, P. Bolon, and J. Chassery, “Exact histogram specification,” *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1143–1152, 2006. [Online]. Available: <https://doi.org/10.1109/TIP.2005.864170>

References

- [15] X. Guo, Y. Li, and H. Ling, "LIME: low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, 2017. [Online]. Available: <https://doi.org/10.1109/TIP.2016.2639450>
- [16] S. Wang, J. Zheng, H. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, 2013. [Online]. Available: <https://doi.org/10.1109/TIP.2013.2261309>
- [17] X. Fu, Y. Liao, D. Zeng, Y. Huang, X. S. Zhang, and X. Ding, "A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4965–4977, 2015. [Online]. Available: <https://doi.org/10.1109/TIP.2015.2474701>
- [18] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, 2017. [Online]. Available: <https://doi.org/10.1016/j.patcog.2016.06.008>
- [19] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 155. [Online]. Available: <http://bmvc2018.org/contents/papers/0451.pdf>
- [20] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. ACM, 2019, pp. 1632–1640. [Online]. Available: <https://doi.org/10.1145/3343031.3350926>
- [21] L. Ma, R. Liu, Y. Wang, X. Fan, and Z. Luo, "Low-light image enhancement via self-reinforced retinex projection model," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.
- [22] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021. [Online]. Available: <https://doi.org/10.1109/TIP.2021.3051462>
- [23] Y. Chen, Y. Wang, M. Kao, and Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6306–6314. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Deep_Photo_Enhancer_CVPR_2018_paper.html
- [24] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vision Appl.*, vol. 25, no. 6, pp. 1423–1468, Aug. 2014.
- [25] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [26] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.

References

- [27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [28] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.
- [29] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 2472–2481. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Residual_Dense_Network_CVPR_2018_paper.html
- [30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 294–310.
- [31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 694–711. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_43
- [32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 586–595.
- [33] C. Ma, B. Yan, W. Tan, and X. Jiang, "Perception-oriented stereo image super-resolution," in *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metze, and B. Prabhakaran, Eds. ACM, 2021, pp. 2420–2428. [Online]. Available: <https://doi.org/10.1145/3474085.3475408>
- [34] A. Lugmayr et al., "Ntire 2020 challenge on real-world image super-resolution: Methods and results," *CVPR Workshops*, 2020.
- [35] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *IEEE International Conference on Computer Vision*, 2021, pp. 4791–4800.
- [36] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *International Conference on Computer Vision Workshops (ICCVW)*.
- [37] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Learning the degradation distribution for blind image super-resolution," *CVPR*, 2022.
- [38] T. Y. Han, Y. J. Kim, and B. C. Song, "Convolutional neural network-based infrared image super resolution under low light environment," in *25th*

References

- European Signal Processing Conference, EUSIPCO 2017, Kos, Greece, August 28 - September 2, 2017.* IEEE, 2017, pp. 803–807. [Online]. Available: <https://doi.org/10.23919/EUSIPCO.2017.8081318>
- [39] C. Ying, P. Zhao, and Y. Li, “Low-light-level image super-resolution reconstruction based on iterative projection photon localization algorithm,” *Journal of Electronic Imaging*, vol. 27, no. 1, pp. 1 – 11, 2018. [Online]. Available: <https://doi.org/10.1117/1.JEI.27.1.013026>
- [40] K. Guo, M. Hu, S. Ren, F. Li, J. Zhang, H. Guo, and X. Kui, “Deep illumination-enhanced face super-resolution network for low-light images,” vol. 18, no. 3, mar 2022. [Online]. Available: <https://doi.org/10.1145/3495258>
- [41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [42] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12346. Springer, 2020, pp. 213–229. [Online]. Available: https://doi.org/10.1007/978-3-030-58452-8_13
- [43] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: deformable transformers for end-to-end object detection,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: <https://openreview.net/forum?id=gZ9hCDWe6ke>
- [44] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, “Learning delicate local representations for multi-person pose estimation,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12348. Springer, 2020, pp. 455–472. [Online]. Available: https://doi.org/10.1007/978-3-030-58580-8_27
- [45] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, “Pose recognition with cascade transformers,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 1944–1953. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Li_Pose_Recognition_With_Cascade_Transformers_CVPR_2021_paper.html
- [46] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *CoRR*, vol. abs/2105.15203, 2021. [Online]. Available: <https://arxiv.org/abs/2105.15203>

References

- [47] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 6881–6890. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Zheng_Rethinking_Semantic_Segmentation_From_a_Sequence-to-Sequence_Perspective_With_Transformers_CVPR_2021_paper.html
- [48] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *IEEE International Conference on Computer Vision Workshops*, 2021.
- [49] Q. Qin, J. Yan, Q. Wang, X. Wang, M. Li, and Y. Wang, "Etdnet: An efficient transformer deraining model," *IEEE Access*, vol. 9, pp. 119 881–119 893, 2021.
- [50] T. H. Kim, M. S. M. Sajjadi, M. Hirsch, and B. Schölkopf, "Spatio-temporal transformer network for video restoration," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11207. Springer, 2018, pp. 111–127. [Online]. Available: https://doi.org/10.1007/978-3-030-01219-9_7
- [51] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28
- [52] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 1874–1883. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.207>
- [53] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. A. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 12 894–12 904. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Vaswani_Scaling_Local_Self-Attention_for_Parameter_Efficient_Visual_Backbones_CVPR_2021_paper.html
- [54] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," *CoRR*, vol. abs/2107.00641, 2021. [Online]. Available: <https://arxiv.org/abs/2107.00641>
- [55] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016. [Online]. Available: <http://arxiv.org/abs/1607.06450>

References

- [56] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," *CoRR*, vol. abs/2103.11816, 2021. [Online]. Available: <https://arxiv.org/abs/2103.11816>
- [57] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
- [58] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018, pp. 464–468. [Online]. Available: <https://doi.org/10.18653/v1/n18-2074>
- [59] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, "Conditional positional encodings for vision transformers," *Arxiv preprint 2102.10882*, 2021. [Online]. Available: <https://arxiv.org/pdf/2102.10882.pdf>
- [60] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1800–1807. [Online]. Available: <https://doi.org/10.1109/CVPR.2017.195>
- [61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004. [Online]. Available: <https://doi.org/10.1109/TIP.2003.819861>
- [62] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *CoRR*, vol. abs/2004.07728, 2020. [Online]. Available: <https://arxiv.org/abs/2004.07728>
- [63] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 334–355.
- [64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
- [65] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, M. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12370. Springer, 2020, pp. 492–511. [Online]. Available: https://doi.org/10.1007/978-3-030-58595-2_30
- [66] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2242–2251. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.244>

References

Paper G

Single-Loss Multi-Task Learning for Improving Semantic Segmentation Using Super-Resolution

Andreas Aakerberg, Anders S. Johansen, Kamal Nasrollahi, and
Thomas B Moeslund

The paper has been published in
Computer Analysis of Images and Patterns - 19th International Conference (CAIP),
Lecture Notes in Computer Science, vol. 13053, pp. 403-411, 2021.

© 2023 by the authors.

The layout has been revised.

Abstract

We propose a novel means to improve the accuracy of semantic segmentation based on multi-task learning. More specifically, in our Multi-Task Semantic Segmentation and Super-Resolution (MT-SSSR) framework, we jointly train a super-resolution and semantic segmentation model in an end-to-end manner using the same task loss for both models. This allows us to optimize the super-resolution model towards producing images that are optimal for the segmentation task, rather than ones that are of high-fidelity. Simultaneously we adapt the segmentation model to better utilize the improved images and thereby improve the segmentation accuracy. We evaluate our approach on multiple public benchmark datasets, and our extensive experimental results show that our novel MT-SSSR framework outperforms other state-of-the-art approaches.

1 Introduction

Semantic Segmentation (SS) is a widely studied computer vision problem that helps scene understanding by assigning dense labels to all pixels in an image. SS has several applications in fields such as autonomous driving, robot sensing, and similar tasks that require a semantic understanding with pixel-level localization. The accuracy of SS is highly correlated with the spatial resolution of the input images [1]. This is particularly prominent for segmentation of small objects, where High-Resolution (HR) is essential to obtain a high accuracy [2]. However, obtaining HR image data is not always possible. One possible solution is therefore to upsample Low-Resolution (LR) images as a pre-processing step. This can be done with classical interpolation-based methods, such as bicubic interpolation, or with the more recent deep-learning based Super-Resolution (SR) methods. The latter has shown to be the most effective in terms of restoring HR details from LR images [3, 4]. Deep-learning based SR models are trained by minimizing the loss, typically Mean Squared Error (MSE) loss, between the reconstructed HR image and the Ground-Truth (GT). Hence, these methods require paired LR/HR images for training. However, in the case of improving another computer vision task, such as SS, the objective and subjective quality of the super-resolved image is not necessarily the best metrics. Therefore, we hypothesize that by only using the segmentation loss, it is possible to optimize the SR model jointly, to produce super-resolved images that result in improved segmentation accuracy.

In this paper, we therefore propose a novel framework named Multi-Task Semantic Segmentation and Super-Resolution (MT-SSSR), for joint learning of SS and super-resolution as seen in Fig. G.1. We use ESRGAN [4] and HRNet [5] respectively as SR and SS backbones, in our joint framework, and rely on a single loss for learning both models, namely the loss of the SS task. We evaluate our method on two different publicly available datasets, and present

new State-of-The-Art (SoTA) results on both. In summary, the contributions of this paper are:

- A novel multi-task learning framework, which uses a single loss to improve the segmentation performance together with SR.
- Our method does not require LR/HR training image pairs for the SR model when jointly learning in the multi-task learning framework.
- We outperform SoTA SS methods on the challenging CityScapes and IDD-Lite datasets by respectively 4.2% and 2.2%, compared to the best existing published results.

2 Related Work

Super-resolution: Dong *et al.* [3] proposed the first deep-learning-based method for SR, which successfully learned to perform non-linear mapping from LR to HR images. Since then, most successful SR methods have been based on convolutional neural networks. One of the SoTA SR methods is ESRGAN [4], which uses a relativistic Generative Adversarial Network (GAN) with Residual-in-Residual Dense Blocks (RRDBs). Besides improving Signal-to-Noise Ratio (SNR), or the perceptual quality of images, SR can also be used to assist other computer vision tasks to achieve better accuracy [6, 7]. Recently, it has been shown that SR can improve optical character recognition accuracy by up to 15% [8] and object detection in satellite imagery by up to 30% [9].

Semantic Segmentation: A popular method to achieve SS is to use an encoder-decoder architecture [10–12] which encodes the input image to dense representational feature-maps and then decodes to regain spatial information [13, 14]. Eff-UNet [15] utilizes Efficientnet [16] as an encoder and UNet [11] as a decoder, to achieve SoTA performance the IDD-Lite dataset [17]. DeepLabV3 [18] uses atrous convolutions and skip-connections for decoding. ERFNet [19] uses deconvolutional layers, combined with a non-bottleneck-1d layer to reduce computational cost. PSPNet [20] proposes a spatial pyramid pooling layer that gathers information by pooling over an increasingly smaller region of the image, then fusing those feature-maps with the original feature-map. Unlike the previously mentioned methods, HRNet [5], aims to retain as much of the resolution of the input image, by combining a HR branch with parallel LR branches to achieve representational information, and subsequently fusing the information from all branches before the final layer. Segmentation models are often optimized using cross-entropy loss, which is a per-pixel evaluation. In [21], Region Mutual Information (RMI) loss is proposed, which utilizes neighboring pixels in a statistical approach, allowing the model to adjust the loss based on how difficult the prediction is, resulting in an overall improvement in accuracy [21].

3. The Proposed Framework

Multi-Task Learning: Multi-task learning has proven to be effective for different computer vision problems, when multiple tasks need to be solved at once. By jointly learning multiple related tasks, the performance of the individual tasks can be further improved, compared to learning them separately. In [22], multi-task learning is used to jointly learn image segmentation and depth estimation. In [23] it is proposed to use two GANs for joint de-noising and SR. In [24] a network that can perform a selection of tasks with the same weights is proposed. This is done with task-specific feature modulation, and residual adapters to adjust the forward pass. The work most closely related to ours is DSRL proposed in [1], where multi-task learning is used to jointly learn SR and SS. As the main purpose of multi-task learning in [1] is to improve the encoder of the segmentation model, the SR is considered an auxiliary task that is removed at test time. A key difference in our approach is that we use our SR model to upsample the input images during both training and testing. Additionally, we use the segmentation loss for optimizing our SR model, while [1] uses MSE, which requires a HR ground truth version of the input images for supervised learning.

3 The Proposed Framework

While the use of SR has shown to improve the performance of other vision tasks, experiments show that traditional SR metrics cannot be used as full proxies to recover all the lost details [6]. We postulate that using traditional SR metrics as auxiliary loss for multi-task learning, serves to optimize the model on some implicit assumptions rather than a global optimum for the entire system. We therefore propose using the segmentation task’s loss for improving the performance of the SR task as well.

The block diagram of our proposed method, MT-SSSR, is shown in Fig. G.1. By jointly training both models using the segmentation loss, we remove the need for LR/HR image pairs during training. This makes our method applicable to real-world applications where such data are not available. The SR backbone in our framework is built upon the RRDB generator from ESRGAN [4]. Hence, we do not perform traditional GAN training with ESRGAN, and instead replace all pixel and feature-based loss functions with our task loss. For SS it is vital to have a high spatial resolution to accurately segment the contents of an image. Hence, we chose HRNet [5] as the backbone architecture. Other than replacing the Online Hard Example Mining (OHEM) cross-entropy loss [25] with RMI loss, we do not modify the HRNet architecture further.

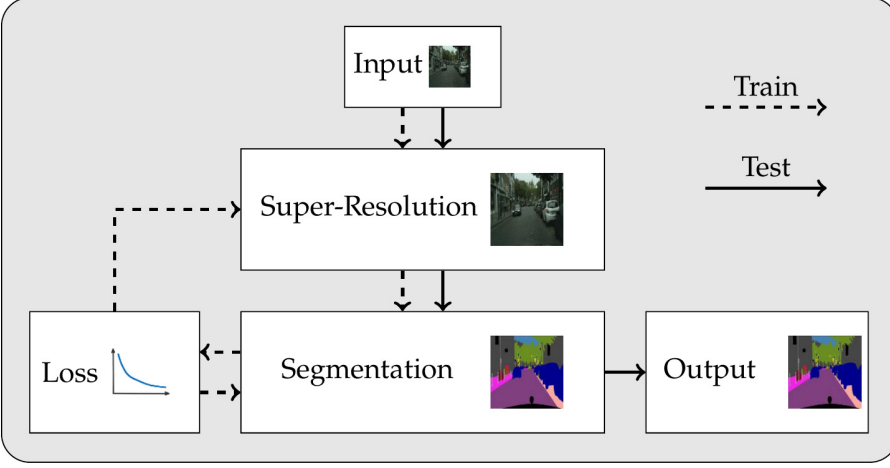


Fig. G.1: Our proposed framework, MT-SSSR. Dashed and full lines represent training and testing phases, respectively. The SR model learns to upsample and enhance the input image based on the segmentation task loss. The segmentation model uses the same loss to improve the accuracy of its prediction.

4 Experiments and Results

4.1 Datasets

The IDD dataset [26] contains driving scenes in unstructured environments, including both urban and rural scenes. In our experiments we use the IDD-Lite dataset [17] which is a sub-sampled version of the IDD dataset. The IDD-Lite dataset contains pixel annotations for 1404 training, 204 validation, and 408 test images, respectively. The dataset has a resolution of 320×227 pixels and contains 7 classes. Ground truth labels are only publicly available for the training and validation images.

The CityScapes dataset [27] contains driving scenes from 50 different cities recorded across several months. The dataset contains finely annotated semantic maps for 2975 training, 500 validation, and 1525 test images, respectively, which have a resolution of 2048×1024 pixels. Following [1], we sub-sample the CityScapes dataset to 1024×512 pixels. There are 19 classes to be segmented. We report our results on the test set, based on submission to the CityScapes Online Server.

4.2 Implementation Details

For both our experiments on CityScapes and IDD-Lite, we initialize the segmentation backbone with weights pre-trained on CityScapes training data. For

4. Experiments and Results

the SR backbone, we use transfer-learning by pre-training the model on generic LR/HR image pairs before the model is used in the multi-task framework. For this, we use the DF2K dataset, which is a merge of DIV2K [28] and Flickr2K [4], and use bicubic interpolation to downsample the HR images. We denote the pre-trained SR model as SR_{ST} (Super-Resolution_{Single-Task}).

For our experiments on CityScapes, we use the sub-sampled images, but test against the full-resolution labels by upsampling our predictions with bi-linear interpolation. For our experiments on IDD-Lite we train at the native resolution training images and labels, and test against 256×128 pixel labels according to [17]. We experiment with both $\times 2$ and $\times 4$ upsampling in our MT-SSSR framework.

Training Setup: Due to memory constraints, we use a cyclic approach for training our MT-SSSR framework, where we alternate between training on patches and the full image. For patch training, we randomly crop 128×128 pixel LR patches from the training images and update both the weights of the SR and the SS model. When training on the full image, we only update the weights of the SS model.

We train all our models using gradient-descent with a mini-batch size of 12 on four V100 GPUs using a learning rate of 0.001 with an exponential decay ($lr \times \frac{iter_{cur}}{iter_{max}}^{0.9}$) trained until convergence. For the segmentation models we additionally use momentum (0.9) and weight decay (0.0005).

4.3 Results

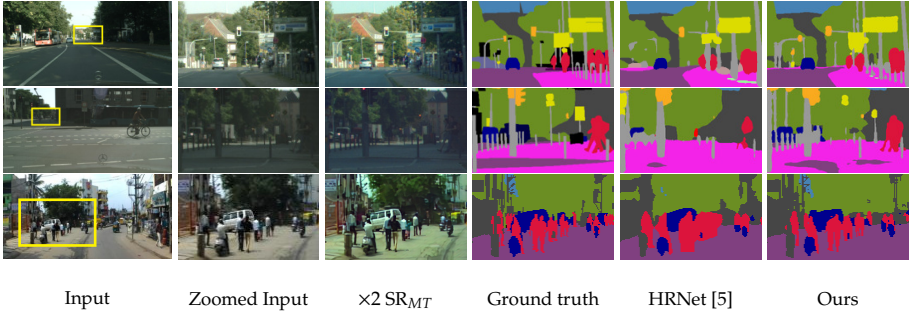


Fig. G.2: Comparison of segmentation results on CityScapes (rows 1,2) and IDD-Lite (row 3). The three first columns show the input image together with a zoomed-in crop of the input and super-resolved image respectively. The last three columns show the differences between the ground truth, HRNet [5] (baseline), and our best performing model respectively. Noticeable differences include; distant streetlights, poles and signs in row 1, traffic poles, and people in row 2, and distant poles and people in row 3.

Results on CityScapes: Table G.1 shows the segmentation accuracy on CityScapes. We include results for experiments with 1024×512 resolution input images and $\times 2$ upsampling of these. Most noticeably, our MT-SSSR framework provides 4.2% improvement over the current SoTA [1] and 3.6% improvement over the HRNet baseline [5] on the test set. As seen in the qualitative comparison in Fig. G.2, our jointly trained SR model enhances sharpness and details of the input images, which in turn helps the segmentation model to better segment smaller distant objects, compared to the baseline.

Method	Scale Factor	Val. (%)	Test (%)
DeepLabV3+ [18]	Native	70.0	67.1
PSPNet [20]	Native	71.5	69.1
HRNet [5]	Native	77.3	75.4
DSRL [1]	$\times 2$ SR_{MT}	75.7	74.8
MT-SSSR (ours)	$\times 2$ SR_{MT}	80.3	79.0

Table G.1: Quantitative segmentation results on CityScapes.

Results on IDD-Lite: The segmentation accuracy on IDD-Lite reported in Table G.2 shows that the performance increases with the upsampling factor in our MT-SSSR. In particular, our method with $\times 4$ upsampling provides 2.5% improvement compared to the current SoTA [15] and 6.9% improvement over the baseline HRNet [5]. In the qualitative segmentation results in

Method	Scale Factor	Val. (%)
DeepLabV3+ [18]	Native	64.3
ERFNet [19]	Native	66.1
HRNet [5]	Native	69.4
Eff-UNet [15]	Native	73.8
MT-SSSR (ours)	$\times 2$ SR_{MT}	74.1
MT-SSSR (ours)	$\times 4$ SR_{MT}	76.3

Table G.2: Quantitative segmentation results on IDD-Lite.

Fig. G.2, it can be seen that our method more accurately segments fine details in the image, compared to the baseline. This is also reflected in the per-class performance in Table G.3. An interesting example can be seen for the triangular part of the pole in the upper left corner (row three), where our method can label the sky correctly, even though this is mislabeled in the GT.

4.4 Ablation Study

Effect of Upsampling and RMI-loss: We investigate the effect of SR and RMI-loss on the CityScapes and IDD-Lite datasets. As seen in Table G.4, RMI-

4. Experiments and Results

Method	Drivable	Non Drivable	Living Things	Vehicles	Roadside Objects	Far Objects	Sky
HRNet + RMI	94.78	43.16	51.10	77.80	51.93	75.97	94.72
Eff-UNet [15]	94.86	50.12	61.96	81.31	54.99	77.54	95.55
MT-SSSR (ours best)	95.07	47.69	68.50	85.97	59.01	80.91	96.66

Table G.3: Per-class accuracy on IDD-Lite. Compared to the baseline and Eff-UNet our method improves significantly on small objects such as living things and roadside objects.

loss improves slightly over using OHEM-loss in the baseline HRNet on both datasets. Furthermore, naively upsampling the input images with $\times 2$ bicubic interpolation also improves the performance slightly. However, when using $\times 4$ upsampling with bicubic interpolation on the IDD-Lite dataset, the accuracy drops 2.3% below baseline. Using images upsampled in a pre-processing step with the pre-trained single-task SR model, SR_{ST} , together with HRNet + RMI, provides 0.7 and 1.3% improvement over the baseline on CityScapes and IDD-Lite, respectively. When combining RMI-loss and SR in our multi-task framework we improve the performance by 3.6% and 6.9% for the CityScapes and IDD-Lite datasets, respectively.

Method	Scale Factor	Val. (%)
HRNet [5]	Native	69.4
HRNet + RMI	Native	69.9
HRNet + RMI	$\times 2$ Bicubic	70.9
HRNet + RMI	$\times 4$ Bicubic	67.1
HRNet + RMI	$\times 2$ SR_{ST}	71.2
MT-SSSR (ours)	$\times 2$ SR_{MT}	74.1
MT-SSSR (ours)	$\times 4$ SR_{MT}	76.3

Table G.4: The effect of RMI-loss and SR on segmentation accuracy on the IDD-Lite dataset. MT and ST denote multi-task and single-task, respectively.

Inference Time: We compare our method in terms of inference time against the baseline model on the CityScapes dataset, on a V100 GPU. The inference time is 101ms and 1888ms per image for the baseline HRNet and MT-SSSR, respectively. The inference time of HRNet at the increased resolution alone is 592ms per image. This means that the increased performance comes at a significant computational cost. However, no particular efforts has been made in order to optimize the inference time.

5 Conclusion

In this paper, we propose a novel framework for SS based on multi-task learning with super-resolution. The super-resolution model learns to enhance the input images such that they become more suitable for the SS model, while the segmentation model jointly learns to predict more accurate segmentation maps. Our experimental results show that our proposed system outperforms existing SoTA SS methods significantly on the challenging CityScapes and IDD-Lite datasets.

6 Acknowledgements

This work was partially supported by the Milestone Research Programme at Aalborg University (MRPA) and Danmarks Frie Forskningsfond (DFF 8022-00360B)

References

- [1] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 3773–3782. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00383>
- [2] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *CVPR-W*, 2016.
- [3] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 38, no. 2, pp. 295–307, Feb 2016.
- [4] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé and S. Roth, Eds. Cham: Springer International Publishing, 2019, pp. 63–79.
- [5] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2019.
- [6] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?" in *WACV*, 2016.
- [7] B. Na and G. C. Fox, "Object classifications by image super-resolution preprocessing for convolutional neural networks," *ASTESJ*, vol. 5, no. 2, pp. 476–483, 2020.

References

- [8] V. Robert and H. Talbot, "Does super-resolution improve OCR performance in the real world? A case study on images of receipts," in *ICIP*, 2020.
- [9] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," in *CVPR-W*, 2019.
- [10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *TPAMI*, 2017.
- [11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28
- [12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, 2017.
- [13] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint*, 2020.
- [14] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *IJAC*, vol. 14, no. 2, pp. 119–135, 2017.
- [15] B. Baheti, S. Innani, S. Gajre, and S. N. Talbar, "Eff-unet: A novel architecture for semantic segmentation in unstructured environment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. IEEE, 2020, pp. 1473–1481. [Online]. Available: <https://doi.org/10.1109/CVPRW50498.2020.00187>
- [16] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.
- [17] A. Mishra, S. Kumar, T. Kalluri, G. Varma, A. Subramaian, M. Chandraker, and C. V. Jawahar, "Semantic segmentation datasets for resource constrained training," in *Computer Vision, Pattern Recognition, Image Processing, and Graphics*, R. V. Babu, M. Prasanna, and V. P. Namboodiri, Eds. Springer Singapore, 2020.
- [18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.
- [19] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, 2018. [Online]. Available: <https://doi.org/10.1109/TITS.2017.2750080>
- [20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.
- [21] S. Zhao, Y. Wang, Z. Yang, and D. Cai, "Region mutual information loss for semantic segmentation," in *NIPS*, 2019.

References

- [22] A. Jha, A. Kumar, S. Pande, B. Banerjee, and S. Chaudhuri, "MT-UNET: A novel u-net based multi-task architecture for visual scene understanding," in *ICIP*, 2020.
- [23] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *CVPR*, 2018.
- [24] K.-K. Maninis, I. Radosavovic, and I. Kokkinos, "Attentive single-tasking of multiple tasks," in *CVPR*, 2019.
- [25] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016.
- [26] G. Varma, A. Subramanian, A. M. Namboodiri, M. Chandraker, and C. V. Jawahar, "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*. IEEE, 2019, pp. 1743–1751. [Online]. Available: <https://doi.org/10.1109/WACV.2019.00190>
- [27] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [28] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Part III

Patent applications

Patent Application I

Patent Application I.



Concept House
Cardiff Road, Newport
South Wales
NP10 8QQ
United Kingdom
Telephone +44 (0)1633 814000
Website <https://www.gov.uk/ipo>

Electronic Filing Receipt

Canon Europe Limited
European Intellectual Property Group
4 Roundwood Avenue, Stockley Park
Uxbridge
Middlesex
United Kingdom
UB11 1AF

Your Ref: MILE-00044-GB-NP

07 April 2022

PATENT APPLICATION NUMBER 2205153.6

We have received your request for grant of a patent and recorded its details as follows:

Filing date(*)	07 April 2022	
Earliest priority date (if any)		
Applicant(s) / contact point	MILESTONE SYSTEMS A/S	
Application fee paid	Yes	
Description (number of pages or reference)	17	
Certified copy of referenced application	Not applicable	
If description not filed	Not applicable	
Claims (number of pages)	5	
Drawings (number of pages)	10	
Abstract (number of pages)	1	
Statement of inventorship (Form 7)	Yes	
Request for search (Form 9A)	Yes	
Request for examination (Form 10)	Yes	
Priority Documents	None	
Other Attachments Received	PDAS Registration Form	PDASRegistration.pdf
	Fee Sheet	FeeSheet.pdf

	Validation Log	ValidLog.pdf
Signed by	CN=Nicolas Gutlé 66700	
Submitted by	CN=Jeong-Mi Kim 66108	
Timestamp of Receipt	07 April 2022, 19:28:06 (BST)	
Digest of Submission	22:ED:8D:02:34:96:08:BA:B0:96:F5:F8:55:64:96:63:14:70:19:07	
Received	/Intellectual Property Office, Newport/	

Please quote the application number in the heading whenever you contact us about this application.

As requested your application as filed will be lodged in the Priority Document Access Service (PDAS) at WIPO. For further information relating to PDAS please see our website <https://www.gov.uk/government/publications/how-to-file-documents-with-the-intellectual-property-office/how-to-file-documents-with-the-intellectual-property-office#file-on-line> or contact our e-filing section on 01633 814870.

If you have any queries about the accuracy of this receipt, please phone the Document Reception Manager on +44 (0) 1633 814570. For all other queries, please phone our Information Centre on 0300 300 2000 if you are calling from the UK, or +44 (0) 1633 814000 if you are calling from outside the UK. Or e-mail information@ipo.gov.uk

* This date is provisional. We may have to change it if we find during preliminary examination that the application does not satisfy section 15(1) of the Patents Act 1977 or if we re-date the application to the date when we get any later filed documents.

Patent Application I.

Patent Application II

Patent Application II.



Intellectual
Property
Office

Concept House
Cardiff Road, Newport
South Wales
NP10 8QQ
United Kingdom

Telephone +44 (0)1633 814000

Website <https://www.gov.uk/ipo>

Electronic Filing Receipt

Canon Europe Limited
European Intellectual Property Group
4 Roundwood Avenue, Stockley Park
Uxbridge
Middlesex
United Kingdom
UB11 1AF

Your Ref: MILE-00057-GB

06 March 2023

PATENT APPLICATION NUMBER 2303244.4

We have received your request for grant of a patent and recorded its details as follows:

Filing date(*)	06 March 2023	
Earliest priority date (if any)		
Applicant(s) / contact point	MILESTONE SYSTEMS A/S	
Application fee paid	Yes	
Description (number of pages or reference)	10	
Certified copy of referenced application	Not applicable	
If description not filed	Not applicable	
Claims (number of pages)	4	
Drawings (number of pages)	None	
Abstract (number of pages)	No, file by 06 March 2024	
Statement of inventorship (Form 7)	No, file by 08 July 2024	
Request for search (Form 9A)	Yes	
Request for examination (Form 10)	None	
Priority Documents	None	
Other Attachments Received	PDAS Registration Form	PDASRegistration.pdf
	Fee Sheet	FeeSheet.pdf

	Validation Log	ValidLog.pdf
Signed by	CN=Dawn Perkins 51650	
Submitted by	CN=Puneet Bilon 66802	
Timestamp of Receipt	06 March 2023, 14:06:35 (GMT)	
Digest of Submission	33:F0:79:76:E5:F4:7A:C4:70:27:DB:38:6B:8D:E 0:14:76:56:D3:14	
Received	/Intellectual Property Office, Newport/	

Please quote the application number in the heading whenever you contact us about this application.

As requested your application as filed will be lodged in the Priority Document Access Service (PDAS) at WIPO. For further information relating to PDAS please see our website <https://www.gov.uk/government/publications/how-to-file-documents-with-the-intellectual-property-office/how-to-file-documents-with-the-intellectual-property-office#file-on-line> or contact our e-filing section on 01633 814870.

If you have any queries about the accuracy of this receipt, please phone the Document Reception Manager on +44 (0) 1633 814570. For all other queries, please phone our Information Centre on 0300 300 2000 if you are calling from the UK, or +44 (0) 1633 814000 if you are calling from outside the UK. Or e-mail information@ipo.gov.uk

* This date is provisional. We may have to change it if we find during preliminary examination that the application does not satisfy section 15(1) of the Patents Act 1977 or if we re-date the application to the date when we get any later filed documents.

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-690-4

AALBORG UNIVERSITY PRESS