Semantic Segmentation Guided Real-World Super-Resolution

Aakerberg, Andreas; Johansen, Anders Skaarup; Nasrollahi, Kamal; Moeslund, Thomas B.

Published in:
Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, 2022

DOI (link to publication from Publisher):
10.1109/WACVW54805.2022.00051

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

Semantic Segmentation Guided Real-World Super-Resolution

Andreas Aakerberg[1], Anders S. Johansen[1], Kamal Nasrollahi[1,2], Thomas B. Moeslund[1]
[1] Visual Analysis and Perception, Aalborg University, Denmark
[2] Research Department, Milestone Systems A/S, Denmark
anaa,asjo,kn,tbm@create.aau.dk

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

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 [20]. 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 cor-

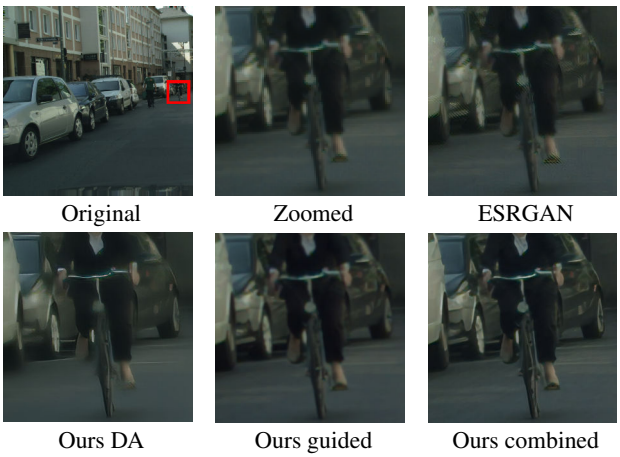


Figure 1: Super-resolution ($\times 4$) of a real image from the Cityscapes dataset [11]. By combining domain adaptation (DA) and guidance by semantic segmentation, our proposed method reconstructs visually pleasing images. In contrast, ESRGAN fails to handle the corruptions in the real image, resulting in many artifacts.

ruptions, 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 [17, 43, 23], where [17] was the winner of the NTIRE 2020 Challenge on RWSR [25]. 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 [3, 30] 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 [37].

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 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 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 [21], which is based on a deep residual CNN, the ResNet based SRResNet proposed by [20], and RCAN [42], 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 [41]. This problem is addressed in [20], which presents an SR model based on Generative Adversarial Networks (GANs), optimized with a combination of MSE, GAN, and VGG loss [18]. This approach leads to more photo-realistic images with better correlation to human perception of good image quality. In ESRGAN [36] 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 [24]. This is mainly due to the domain gap between the real and synthetic LR images. To overcome this issue, ZSSR [29] introduced a zero-shot approach which learns an image specific SR model at test time. In MSZR [30] this concept is extended to exploit information from an external dataset as well. In KernelGAN [3], ZSSR is used together with a GAN based network for estimation of image-specific blur kernels. DAN [26] 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 [37]. In contrast, the prediction times of our method are similar to [36]. In [17], 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 [39], 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*. [12], 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 [10] 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 [16, 45, 9] it has been shown that semantic information can be utilized to generate detailed textures and realistic looking images. In [27], 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 [35] and [22]. It is shown in [35] that the conditions can strongly influence the textures generated and result in much more realistic looking textures that are more semantically appropriate. While [27] shows that CNNs learn some categorical information, [28] 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 backbone 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 [40], 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 [38, 7], and as such SoTA SS methods have tended towards architectures that retain spatial resolution to some extent. PSPNet [38] 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 [7] 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 [8] with depth-wise-separable convolutions resulting in the network being able to learn more fine-grained control of the details in each layer. HRNet [33] 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 upsampling of the prediction is avoided, which leads to more accurate predictions [1], 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.

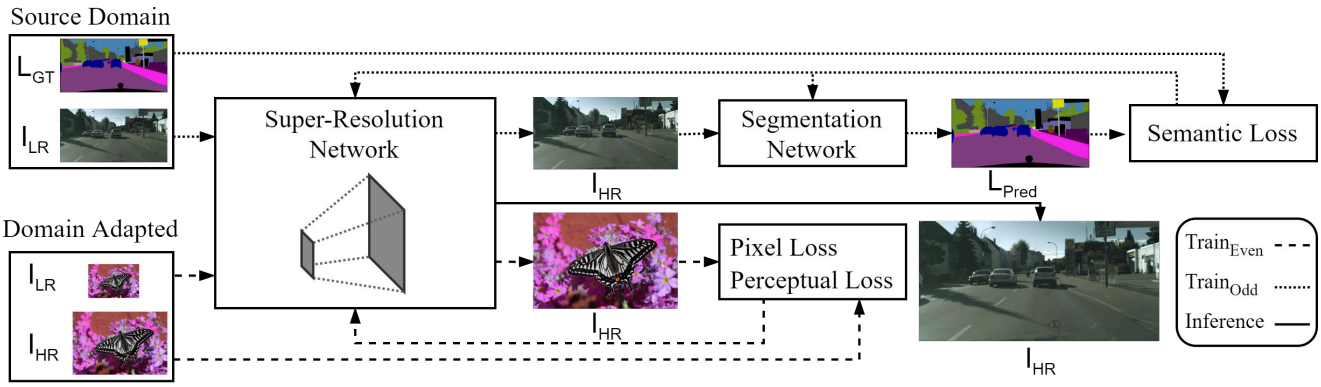


Figure 2: Schematic overview of our proposed SSG-RWSR. To learn to perform RWSR we leverage both guiding from an auxiliary semantic segmentation task and domain adaptation. At test time, the semantic segmentation network is de-coupled, and as such no semantic labels are required to super-resolve the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

3. The proposed method

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:

$$I_{LR} = (I_{HR} * k) \downarrow_s + n \tag{1}$$

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 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 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 [28], 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 [15]. 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 [17]. For estimation of realistic blur kernels, we use KernelGAN [3], on real LR images in X to build a pool of image-specific blur kernels that can be used to degrade the clean HR im-

ages in Y.

To generate artificial LR images which are more similar to the real LR images we employ the method from [6] 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:

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

and

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

where $Mean$ and Var denote the mean and variance, respectively, and μ and γ are scaling factors. Different from [6] we add an additional constraint to ensure that saturated patches are not extracted:

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

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 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:

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

where I_D is the downscaled image, Y_n is a HR image, k_i refers to a kernel from the degradation pool $\{k_1, k_2, \cdots k_m\}$ 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:

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

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

3.3. Backbone networks

Super-resolution Our SR network consist of 23 Residual-in-Residual Dense Blocks (RRDBs) [36]. 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 Learned Perceptual Image Patch Similarity (LPIPS) loss, $\mathcal{L}_{lpips}$, 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 [41]. The total loss for learning the SR model from the domain adapted images is defined as:

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

where λ_{pix}, and λ_{lpips} 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 [33] 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:

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

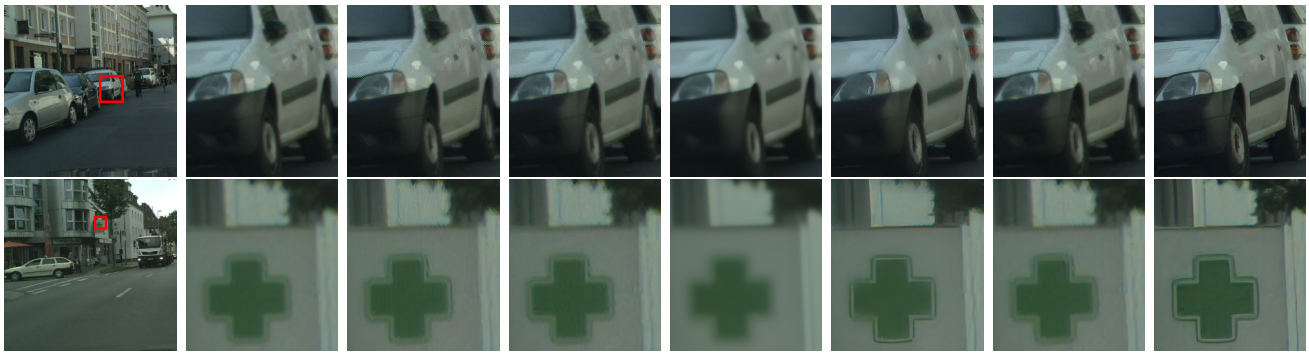
where λ_{ce} is a scaling parameter.

4. Implementation details

Similar to recent RWSR literature [24, 5, 25] 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 [2] and Flickr2K [31], 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_{lpips}, 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 [36].



Original Bicubic ESRGAN[36] MZSR[30] DPSR[39] RealSR[17] DAN[26] Ours

Figure 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.

Original Bicubic ESRGAN[36] MZSR[30] DPSR[39] RealSR[17] DAN[26] Ours GT

Figure 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.

Original Bicubic ESRGAN[36] MZSR[30] DPSR[39] RealSR[17] DAN[26] Ours

Figure 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. Experiments and results

We compare our proposed method to four recent SoTA methods for SR of real images, namely MZSR [30], DPSR [39], RealSR [17], and DAN [26]. We adjust the competing models for optimal performance for a fair comparison. We use KernelGAN [3] to estimate blur kernels for use with MZSR [30]. For DPSR [39] and DAN [26], we set noise levels as recommended by the authors. With RealSR [17] we use the degradation framework provided by the authors,

and re-train the model to the respective datasets. We also include the ESRGAN [36] 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 [11] and IDD [32] 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 [24], 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 of human perceived perceptual quality [25]. 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 [14] and MetaIQA [44] 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 [41], and DISTS [13]. 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 [41]. Note that low distortion and high perceptual quality are at odds with each other, making it impossible to two obtain both [4]. 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.

Cityscapes (Real LR images)

Method	NIMA ↑	Meta-IQA ↑	MOR ↓
Bicubic [19]	4.62	0.245	-
ESRGAN [36]	4.95	0.247	-
MZSR [30]	4.88	0.231	3.33
DPSR [39]	4.83	0.240	4.41
RealSR [17]	4.87	0.236	2.75
DAN [26]	4.65	0.246	3.47
Ours	**5.04**	**0.254**	**1.21**

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

IDD (Real LR images)

Method	NIMA ↑	Meta-IQA ↑	MOR ↓
Bicubic [19]	4.73	0.330	-
ESRGAN [36]	4.94	0.325	-
MZSR [30]	5.00	0.330	2.96
DPSR [39]	4.92	0.330	3.16
RealSR [17]	4.83	0.296	4.88
DAN [26]	4.77	**0.330**	2.48
Ours	**5.03**	0.323	**1.45**

Table 2: Quantitative results on the IDD validation sets. ↑ and ↓ 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.

5.3. Quantitative results

Real images As show in Table 1 and 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 [26] is slightly better according to the Meta-IQA scores on the IDD dataset.

Cityscapes (Synthesized LR images)				
Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
Bicubic [19]	27.51	0.62	0.64	0.19
ESRGAN [36]	18.17	0.12	1.29	0.20
MZSR [30]	26.68	0.55	0.73	0.16
DPSR [39]	**33.11**	**0.90**	0.42	0.13
RealSR [17]	25.88	0.77	0.26	0.10
DAN [26]	27.16	0.58	0.60	0.20
Ours	29.08	0.83	**0.19**	**0.07**

Table 3: Quantitative results on the artificially degraded Cityscapes validation set. ↑ and ↓ 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.

However, this is in contrast to the visual appearance of the images, as the digits on the licence plates shown in Figure 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 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 [4].

5.4. Qualitative results

Real images In Figure 3 and 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 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.

5.5. 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 1 and Table 4 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 ↑	Meta-IQA ↑
Ours (DA)	4.33	0.206
Ours (Guided only)	5.00	0.251
Ours	**5.04**	**0.254**

Table 4: The effect of the different components in our proposed method on the Cityscapes validation set. ↑ and ↓ 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] A. Aakerberg, A. Johansen, K. Nasrollahi, and T. Moeslund. Single-loss multi-task learning for improving semantic segmentation using super-resolution. *In Press, CAIP*, 2021.

[2] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR*, 2017.

[3] S. Bell-Kligler, A. Shocher, and M. Irani. Blind super-resolution kernel estimation using an internal-gan. In *NIPS*. 2019.

[4] Y. Blau and T Michaeli. The perception-distortion tradeoff. In *CVPR*. IEEE, 2018.

[5] Jianrui Cai, Shuhang Gu, Radu Timofte, and Lei 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.

[6] J. Chen, J. Chen, H. Chao, and M. Yang. Image blind denoising with generative adversarial network based noise modeling. In *CVPR*, 2018.

[7] L. 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.

[8] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[9] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.

[10] X. Cheng, Z. Fu, and J. Yang. Zero-shot image super-resolution with depth guided internal degradation learning. In *ECCV*, 2020.

[11] 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 *CVPR*, 2016.

[12] R. De Lutio, S. D'Aronco, J. D. Wegner, and K Schindler. Guided super-resolution as pixel-to-pixel transformation. In *ICCV*, 2019.

[13] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, 2020.

[14] H. T. Esfandarani and P. Milanfar. NIMA: neural image assessment. *TIP*, 2018.

[15] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.

[16] P. Isola, J. Zhu, T. Zhou, and A. A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[17] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPR Workshops*, 2020.

[18] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[19] R. G. Keys. Cubic Convolution Interpolation for Digital Image Processing. *IEEE Trans Acoust. Speech Signal Process*, 1981.

[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 *CVPR*, 2017.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017.

[22] L. Liu, S. Wang, and L. Wan. Component semantic prior guided generative adversarial network for face super-resolution. *IEEE Access*, 2019.

[23] A. Lugmayr, M. Danelljan, and R. Timofte. Unsupervised learning for real-world super-resolution. In *CVPR Workshops*.

[24] 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 *ICCV Workshops*, 2019.

[25] R. Lugmayr, M. Danelljan, and R. Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. *CVPR Workshops*, 2020.

[26] Zhengxiong Luo, Y. Huang, S. Li, L. Wang, and T. Tan. Unfolding the alternating optimization for blind super resolution. In *NeurIPS*, 2020.

[27] 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.

[28] 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.

[29] A. Shocher, N. Cohen, and M. Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, June 2018.

[30] J. W. Soh, S. Cho, and N. I. Cho. Meta-transfer learning for zero-shot super-resolution. In *CVPR*, 2020.

[31] R. Timofte, E. Agustsson, L. Van Gool, M. Yang, and L. Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR Workshops*, 2017.

[32] 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 *WACV*, 2019.

[33] 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.

[34] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Dual super-resolution learning for semantic segmentation. In *CVPR*, pages 3774–3783, 2020.

[35] 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.

[36] 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 *ECCV*, 2019.

[37] Z. Wang, J. Chen, and S. Hoi. Deep learning for image super-resolution: A survey. *TPAMI*, 2020.

[38] Hengshuang Z., Jianping S., Xiaojuan Q., Xiaogang W., and Jiaya J. Pyramid scene parsing network. In *CVPR*, 2017.

[39] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Deep plug-and-play super-resolution for arbitrary blur kernels. In *CVPR*, 2019.

[40] Qian Zhang, Guang Yang, and Guixu Zhang. Collaborative network for super-resolution and semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–12, 2021.

[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[42] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *ECCV*, 2018.

[43] Ruofan Zhou and Sabine Süsstrunk. Kernel modeling super-resolution on real low-resolution images. In *ICCV*, 2019.

[44] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *CVPR*, 2020.

[45] S. Zhu, R. Urtasun, S. Fidler, D. Lin, and C. Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017.