Real-world super-resolution of face-images from surveillance cameras

Aakerberg, Andreas; Nasrollahi, Kamal; Moeslund, Thomas B.

Published in:
IET Image Processing

DOI (link to publication from Publisher):
10.1049/ipr2.12359

Creative Commons License
CC BY 4.0

Publication date:
2022

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

Received: 18 January 2021 | Revised: 3 September 2021 | Accepted: 7 October 2021

DOI: 10.1049/ipr2.12359

IET Image Processing

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Real-world super-resolution of face-images from surveillance cameras

Andreas Aakerberg[1] | **Kamal Nasrollahi**[1,2] | **Thomas B. Moeslund**[1]

[1] Visual Analysis and Perception, Aalborg University, Rendsburggade 14, Aalborg, Denmark

[2] Research Department, Milestone Systems A/S, Milestone Systems, Brøndby, Denmark

Correspondence
Andreas Aakerberg, Visual Analysis and Perception, Aalborg University, Rendsburggade 14, Aalborg, Denmark.
Email: anaa@create.aau.dk

Funding information
Independent Research Fund Denmark, Grant/Award Number: 8022-00360B

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, a novel framework for the generation of realistic LR/HR training pairs is proposed. The 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 to train an SR model using high-quality face images as Ground-Truth (GT). For better perceptual quality, a Generative Adversarial Network (GAN) based SR model is used, where the commonly used VGG-loss [1] is exchanged 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 the existing State-of-the-Art (SoTA) methods. In addition, it is shown 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 SR is a special case of SR which aims to restore HR face images from their LR counterparts. This is useful in many different applications such as video surveillance and face enhancement. Current 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.

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 an 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, that is, the source domain images. Creating LR images from clean high-quality images, that is, the target domain, allows us to train an SR model that learns to super-resolve images with similar characteristics. This approach is inspired by the work of Ji et al. [23] 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 [23] to include JPEG compression artifacts. We use the ESRGAN [11] 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 [38] loss for the discriminator similar to [23]. Inspired by Jo et al.

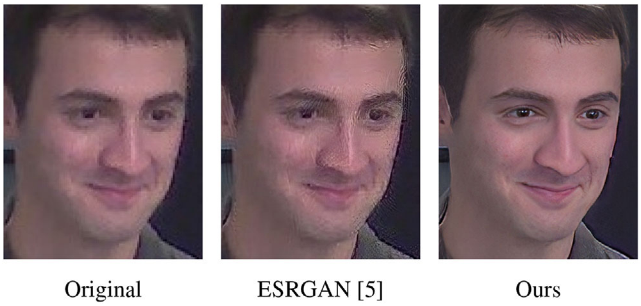


FIGURE 1 ×4 SR of a real low-quality face image (100 × 128 pixels) from the Chokepoint DB [34]. Our method enhances details and removes noise while the ESRGAN [11] amplifies the corruptions

[39], 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 [31–33], 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. A comparison of a reconstructed face image produced by our method and the baseline can be seen in Figure 1.

We evaluate our method on two different face image datasets and one dataset of general images. To enable comparison of the SR performance against GT reference images, we artificially corrupt high-quality images from Flickr-Faces-HQ Dataset (FFHQ) [40] and DIV2k [43] and report quantitative results using conventional IQA methods and the most recent methods for the assessment of the perceptual quality. For the evaluation on real LR face image from surveillance cameras we use the Chokepoint DB [34]. In this case, as no GT image is available, we report the results using 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 of the 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:

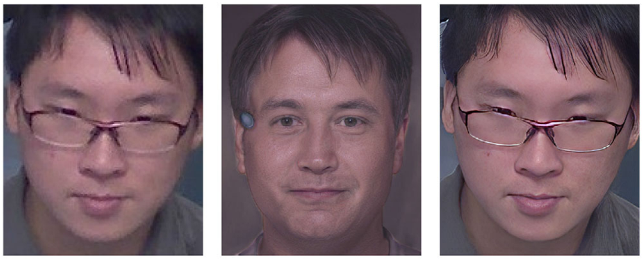
- A novel framework for the generation of LR/HR training pairs, where we introduce realistic image compression artifacts, and improve upon the noise collection method from [35], for noise injection, by adding additional constraints.
- Improving the ESRGAN [11] SR model with a novel combination of loss functions including local patch-wise adversarial loss [38], 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 [34] and artificially corrupted face images from the FFHQ DB [40]. Furthermore, we also evaluate on general images from the DIV2K dataset [43], 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. [4] 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 [5, 6], recursive learning [7], multi-path learning [8], and different loss functions [9] 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. [10] proposed to use GANs for SR with the SRGAN, to achieve realistic looking images according to human perception. The ESRGAN [11] further improves the SRGAN [10] 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 [12].

2.1 | 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 [13–15]. 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. [16] 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. [17] extend the work of [16] by using meta-transfer learning phase to exploit information from an external dataset. Gu et al. [18] 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. [19] propose a deep plug-and-play framework which takes advantage of existing blind deblurring methods for blur kernel estimation. Bell-



Original PULSE [38] Ours

FIGURE 2 An example of SR of a real low-quality face image from the Chokepoint DB [34], where it can be seen that the PULSE [29] method changes the identity of the person, while our method preserves the identity and enhances details

Kligler et al. [20] trains a GAN to estimate blur kernels from LR images and combines it with the ZSSR SR model [16]. Fritsche et al. [21] 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. [22] 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. [23] 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.

2.2 | Face super-resolution

Face SR is an SR technique specialized for reconstruction of face images. One of the first methods for face SR was proposed by Baker and Kanade [24]. This method reconstructed face details by searching for the most optimal mapping between LR and HR patches. Wang et al. [25] used an eigen transformation to map between LR and HR faces. Yang et al. [26] 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. [27] use pixel recursive learning with two CNNs to synthesize realistic hair and skin details. Chen et al. [28] 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, Menon et al. [29] 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 2. Additionally, none of the above mentioned methods are robust against noise or other corruptions in the input images [30].

There are very few publications available in the literature which address the problem of RWSR of face-images [30]. 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 [31–33]. 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 ×4 RWSR of face images of arbitrary sizes, which we evaluate on real LR face images from surveillance cameras without any prior re-scaling.

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 Y, such that these have similar image characteristics as the ones in the source domain X. The second step involves training an SR model on the constructed paired data, and optimizing for perceptual quality.

3.1 | Novel image degradation

Traditional approaches for SR assumes that an 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:

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

However, real LR images from cameras are influenced by multiple other factors that degrade the image as well. The RealSR [23] 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 [23] 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}.

$$I_{LR} = c((I_{HR} * k) \downarrow_s + n), \tag{2}$$

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. [20]. 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.

3.2.1 | 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:

$$I_D = (Y_n * k_i) \downarrow_s, i \in \{1, 2 \cdots m\}, \tag{3}$$

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.

3.3 | Noise estimation

For degradation with realistic image noise, we adopt the method from [35] 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 following constraints are met:

$$|Mean(q_j^i) - Mean(p_i)| \leq \mu \cdot Mean(p_i), \tag{4}$$

and

$$|Var(q_j^i) - Var(p_i)| \leq \gamma \cdot Var(p_i), \tag{5}$$

where $Mean$ and Var denote the mean and variance, respectively, and μ and γ are scaling factors. Different from [35] we add an additional constraint to ensure that saturated patches are not extracted:

$$Var(p_i) \geq \phi, \tag{6}$$

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.

3.3.1 | 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:

$$I_N = I_D + n_i, i \in \{1, 2 \cdots l\}, \tag{7}$$

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

3.4 | 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 [11], which is one of the SoTA networks for perceptual SR with ×4 upscaling, and train it on the paired LR and HR images generated with our degradation framework. Different from the SRGAN [10], 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.

3.5.1 | 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 [11] model uses several different loss functions during training. More specifically, the generator uses adversarial loss $\mathcal{L}_{adv}$ [36] in combination with VGG perceptual loss $\mathcal{L}_{vgg}$ [1] and pixel loss $\mathcal{L}_{pix}$, while the discriminator use VGG-128 [37] loss $\mathcal{L}_{vgg}$. However, we find that this combination of loss functions is not ideal for high perceptual quality. Following the work of [23], we first exchange the VGG-128 [37] discriminator loss with a PatchGAN discriminator from [38] 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 the 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 [39], we find that using the LPIPS, perceptual loss $\mathcal{L}_{lpips}$ [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:

$$\mathcal{L}_{lpips} = \sum_{k} \tau^{k}(\phi^{k}(I_{gen}) - \phi^{k}(I_{gt})), \tag{8}$$

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:

$$\mathcal{L}_{generator} = \lambda_{pix} \cdot \mathcal{L}_{pix} + \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{lpips} \cdot \mathcal{L}_{lpips}, \tag{9}$$

where λ_{pix}, λ_{adv} and λ_{lpips} 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 [34] 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 split the dataset, to obtain 72,282 images for training and 3,805 images for testing. The average resolution of the cropped faces is ≈ 92 × 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 [41], 8,040 face images from the Radboud Faces Database [42] and 5,000 randomly selected face images from FFHQ database [40] 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 [43] consiting 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

3.7.1 | 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, for example, 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 NIQUE 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.

3.7.2 | 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) [52]. However, these metrics focus more on

TABLE 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 the evaluation of perceptual quality

Method	NIQE ↓	BRISQUE ↓	PIQE ↓	NRQM ↑	PI ↓	NIMA ↑	MOR ↓
Bicubic [56]	5.77	56.77	86.28	3.09	6.34	3.92	–
MZSR [17]	7.36	50.09	77.63	3.75	6.81	3.97	–
EDSR [6]	5.43	50.63	81.97	3.82	5.81	4.08	–
ESRGAN [11]	3.75	19.35	19.20	7.08	**3.34**	4.34	4.72
USRNet [22]	6.10	59.13	87.70	3.19	6.46	4.75	3.11
RealSR [23]	**3.50**	**17.20**	**9.11**	5.45	4.00	4.93	3.39
DPSR [57]	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**

signal fidelity rather than perceptual quality [53]. 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) [54], LPIPS, [2], and Deep Image Structure and Texture Similarity (DISTS) [55].

4 | EXPERIMENTS AND RESULTS

4.1 | 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 [11]. We use LR patches of size 32 × 32, and empirically set λ_{pix}, λ_{adv} and λ_{lpips} to 0.01, 0.005 and 0.001, respectively. For noise estimation, we set p_i to match the LR patch size and q^i_j to 8. Similar to [35], 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 randomly chosen strength of 15 to 30.

4.2 | Comparison with state-of-the-art

We did not find any other ×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 [11], EDSR [6]), two SR methods for arbitrary blur kernels (DPSR [57], USRNet [22]), two real-world SR models (MZSR [17], and RealSR [23]). We fine-tune or adjust the competing models for optimal performance for a fair comparison. For the unsupervised MZSR [17], we enable back-projection with 10 iterations and set a noise level of 0.5. We re-train the RealSR [23] using the framework provided by the authors. The remaining methods all require 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 [22] and DPSR [57], we input blur kernels estimated with KernelGAN [20], and set noise levels for real images as recommended by the authors.

4.2.1 | Real-world face images

In this experiment, we evaluate the SR performance on real LR face images from the Chokepoint testset. Quantitative and qualitative results can be seen in Table 1 and Figure 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]) fail 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.

4.2.2 | Artificially corrupted face images

This experiment evaluates the SR performance on artificially corrupted images from the FFHQ testset. We show quantitative and qualitative results in Table 2 and Figure 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.

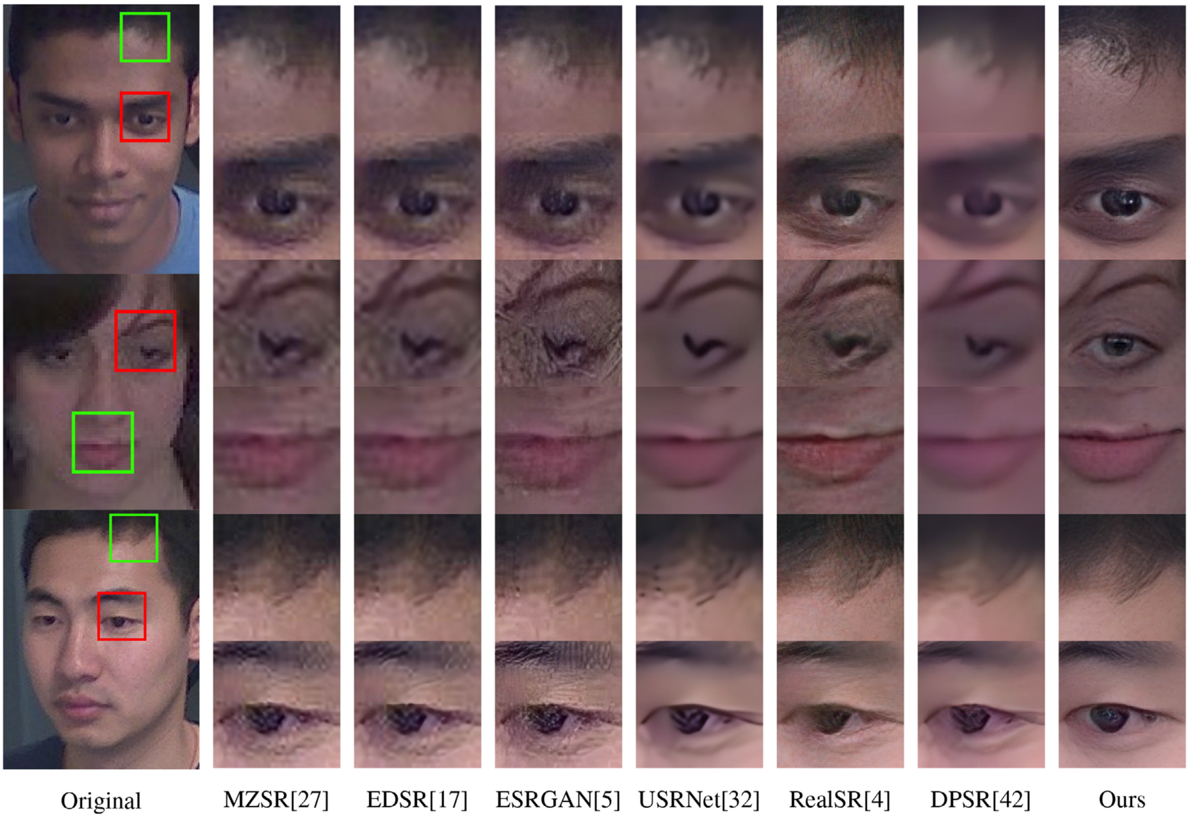


FIGURE 3 Comparison with SoTA methods for ×4 SR of real low-quality face images from the Chokepoint DB [34]. As visible, our method generates superior reconstructions over the existing methods for different faces

TABLE 2 Quantitative results on the FFHQ testset. ↑ and ↓ indicate whether higher or lower values are desired, respectively

Method	PSNR ↑	SSIM ↑	MS-SSIM ↑	NLPD ↓	LPIPS ↓	DISTS ↓
Bicubic [56]	28.39	0.79	0.88	0.32	0.52	0.20
MZSR [17]	29.56	0.78	0.89	0.29	0.43	0.18
EDSR [6]	28.27	0.78	0.88	0.33	0.50	0.19
ESRGAN [11]	28.09	0.77	0.88	0.34	0.40	0.19
USRNet [22]	28.53	**0.80**	0.89	0.32	0.53	0.21
RealSR [23]	29.14	0.79	0.90	0.29	0.29	0.18
DPSR [57]	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**

4.2.3 | Artificially corrupted general images

Finally, we also evaluate on artificially corrupted generic images from the DIV2K [43] validation set. Quantitative and qualitative results can be seen in Table 3 and Figure 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 the method is closer to the ground truth.

TABLE 3 Quantitative results on the DIV2K validation set. ↑ and ↓ indicate whether higher or lower values are desired, respectively

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Bicubic	25.16	0.65	0.67
ESRGAN [11]	16.40	0.14	0.99
RealSR [23]	18.37	0.50	0.34
Ours	20.95	0.58	0.31

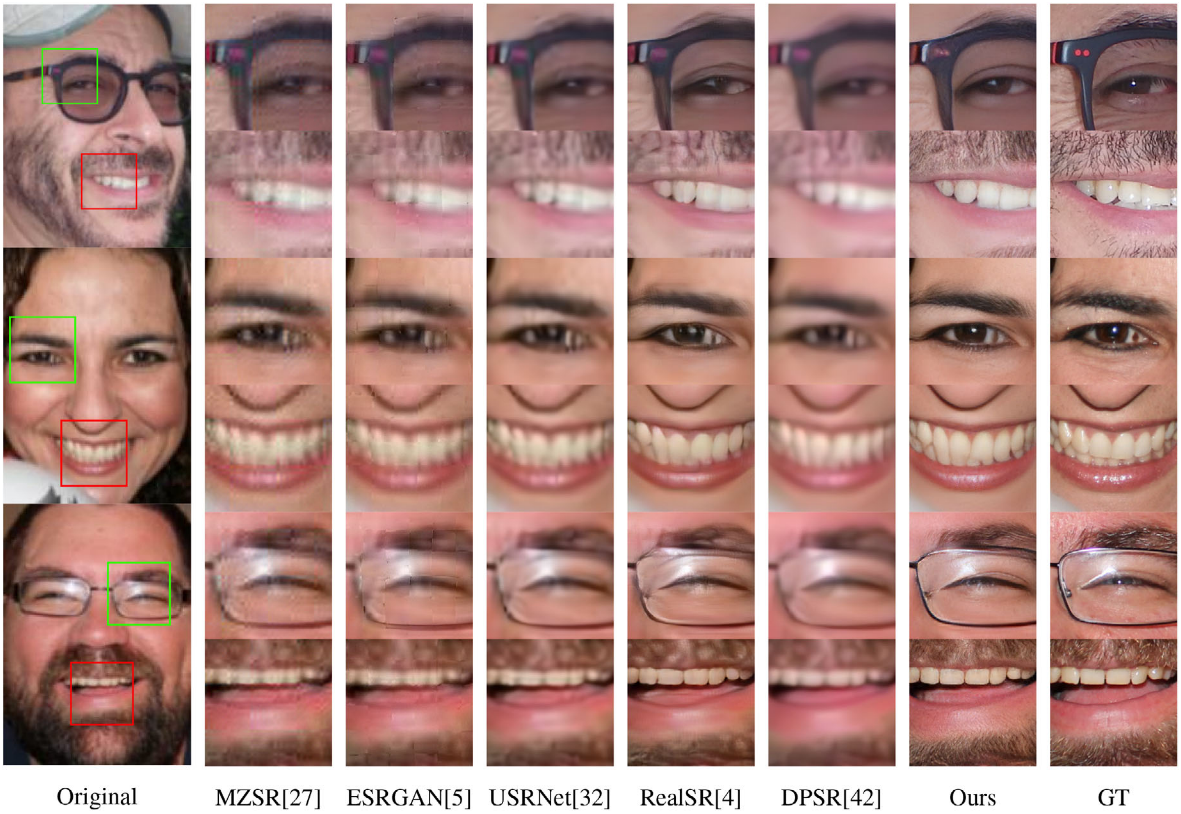


FIGURE 4 Comparison with SoTA methods for ×4 SR of artificially corrupted face images from the FFHQ [40] testset. As seen, our method hallucinates faces with richer detail and less artifacts compared to the existing methods

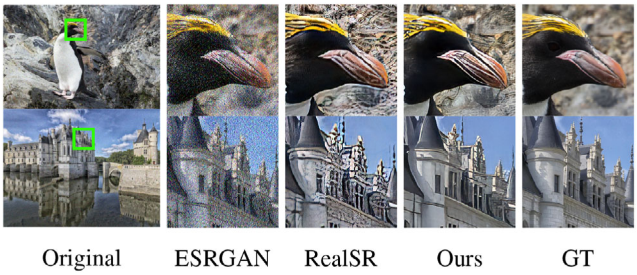


FIGURE 5 Super-resolution results of artificially corrupted LR images from the DIV2K dataset

4.3 | 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 4.2. A qualitative comparison can be seen in Figure 6.

4.3.1 | Baseline

Here, we use kernel estimation and noise injection to generate training data for the ESRGAN with patch discriminator, similar to [23]. 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.

4.3.2 | 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.

4.3.3 | 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 become sharper with better texture and details.

4.4 | Failure cases

While our method produces reconstructed faces of better visual quality than the compared SoTA methods, it does not solve the problem RWSR of face images. Figure 7 shows several failure cases of our method. These occur when the input image is severely corrupted, for example, by motion blur or harsh

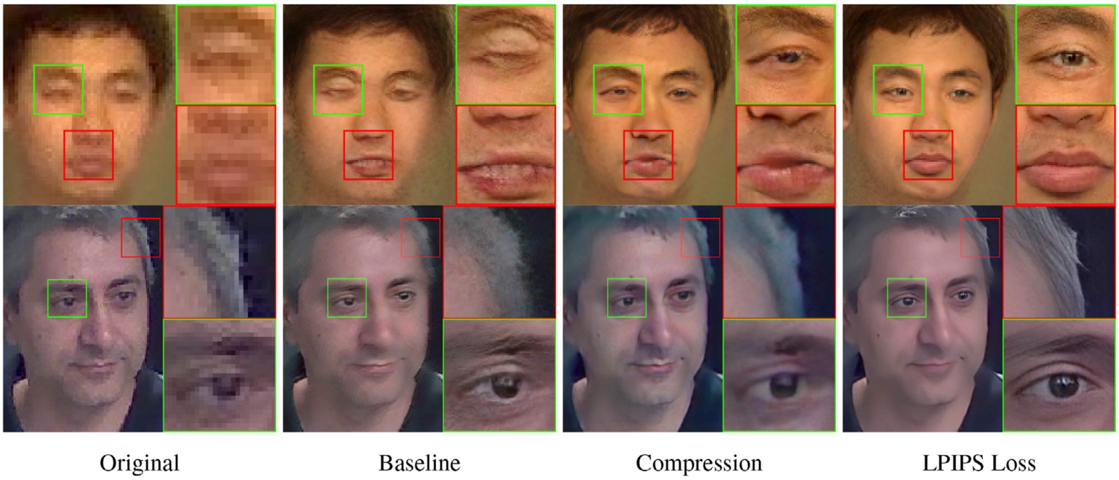


FIGURE 6 Ablation study of the effect of including compression artifacts in the degradation framework and exchanging the VGG-loss with LPIPS-loss for the generator in the SR model, compared to the baseline and the original LR images

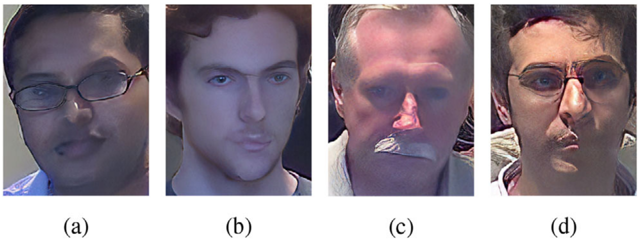


FIGURE 7 Examples of failure cases; (a) and (b) illustrate cases where only parts of the image is super-resolved; (c) shows a case where almost no high-frequency details are restored; (d) shows a case where unrealistic facial features are introduced

lighting, or when out-of-focus. In these cases, our method might only super-resolve some parts of the face, for example, a single eye, or even hallucinate unrealistic facial features.

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 in the framework.

ACKNOWLEDGEMENTS

This work was supported by Independent Research Fund Denmark under grant number 8022-00360B.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at http://arma.sourceforge.net/chokepoint/ and https://github.com/NVlabs/ffhq-dataset, and https://data.vision.ee.ethz.ch/cvl/DIV2K/ for the Chokepoint DB, FFHQ, and DIV2K datasets, respectively.

ORCID

Andreas Aakerberg https://orcid.org/0000-0002-3911-2638
Kamal Nasrollahi https://orcid.org/0000-0002-1953-0429
Thomas B. Moeslund https://orcid.org/0000-0001-7584-5209

REFERENCES

1. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of Computer Vision - ECCV 2016-14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016. Lecture Notes in Computer Science, vol. 9906, Part II, pp. 694–711. Springer, Berlin (2016). Available from: https://doi.org/10.1007/978-3-319-46475-6_43
2. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595. IEEE, Piscataway (2018)
3. Esfandarani, H.T., Milanfar, P.: NIMA: neural image assessment. IEEE Trans. Image Process. 27(8), 3998–4011 (2018). https://doi.org/10.1109/TIP.2018.2831899
4. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38(2), 295–307 (2016)
5. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral). IEEE, Piscataway (2016)
6. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1132–1140. IEEE, Piscataway (2017)
7. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral). IEEE, Piscataway (2016)

8. Han, W., Chang, S., Liu, D., Yu, M., Witbrock, M., Huang, T.S.: Image super-resolution via dual-state recurrent networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1654–1663. IEEE, Piscataway (2018)
9. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2017)
10. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114. IEEE, Piscataway (2017)
11. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., et al.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Computer Vision – ECCV 2018 Workshops, pp. 63–79. Springer, Berlin (2019)
12. Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., et al.: Aim 2019 challenge on real-world image super-resolution: Methods and results. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3575–3583. IEEE, Piscataway (2019)
13. Cai, J., Zeng, H., Yong, H., Cao, Z. and Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3086–3095. IEEE, Piscataway (2019)
14. Vaezi Joze, H., Zharkov, I., Powell, K., Ringler, C., Liang, L., Roulston, A., et al.: Imagepairs: Realistic super resolution dataset via beam splitter camera rig. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE, Piscataway (2020). https://www.microsoft.com/en-us/research/publication/imagepairs-realistic-super-resolution-dataset-via-beam-splitter-camera-rig/
15. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., et al.: Component divide-and-conquer for real-world image super-resolution. Proceedings of the European Conference on Computer Vision, (2020)
16. Assaf Shocher, M.I. Nadav Cohen: ”zero-shot” super-resolution using deep internal learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Piscataway (2018)
17. Soh, J.W., Cho, S., Cho, N.I.: Meta-transfer learning for zero-shot super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2020)
18. Gu, J., Lu, H., Zuo, W., Dong, C.: Blind super-resolution with iterative kernel correction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Piscataway (2019)
19. Zhang, K., Zuo, W., Zhang, L.: Deep plug-and-play super-resolution for arbitrary blur kernels. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 1671–1681. IEEE, Piscataway (2019). http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_Deep_Plug-And-Play_Super-Resolution_for_Arbitrary_Blur_Kernels_CVPR_2019_paper.html
20. Bell Kligler, S., Shocher, A., Irani, M.: Blind super-resolution kernel estimation using an internal-gan. In: Advances in Neural Information Processing Systems, vol. 32, pp. 284–293. Curran Associates, Inc., Red Hook (2019)
21. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. IEEE, Piscataway (2019)
22. Zhang, K., Van Gool, L., Timofte, R.: Deep unfolding network for image super-resolution. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3217–3226. IEEE, Piscataway (2020)
23. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE, Piscataway (2020)
24. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. In: 2000 Conference on Computer Vision and Pattern Recognition (CVPR 2000), pp. 2372–2379. IEEE, Piscataway (2000). https://doi.org/10.1109/CVPR.2000.854852
25. Wang, X., Tang, X.: Hallucinating face by eigentransformation. IEEE Trans. Syst. Man Cybern. Part C 35(3), 425–434 (2005). https://doi.org/10.1109/TSMCC.2005.848171
26. Yang, C., Liu, S., Yang, M.: Structured face hallucination. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1099–1106. IEEE, Piscataway (2013). https://doi.org/10.1109/CVPR.2013.146
27. Dahl, R., Norouzi, M., Shlens, J.: Pixel recursive super resolution. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 5449–5458. IEEE, Piscataway (2017). https://doi.org/10.1109/ICCV.2017.581
28. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Piscataway (2018)
29. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: 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, pp. 2434–2442. IEEE, Piscataway (2020). https://doi.org/10.1109/CVPR42600.2020.00251
30. Grm, K., Pernus, M., Cluzel, L., Scheirer, W.J., Dobrisek, S., Struc, V.: Face hallucination revisited: An exploratory study on dataset bias. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, pp. 2405–2413. IEEE, Piscataway (2019). http://openaccess.thecvf.com/content_CVPRW_2019/html/Biometrics/Grm_Face_Hallucination_Revisited_An_Exploratory_Study_on_Dataset_Bias_CVPRW_2019_paper.html
31. Bulat, A., Tzimiropoulos, G.: 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, pp. 109–117. IEEE, Piscataway (2018). http://openaccess.thecvf.com/content_cvpr_2018/html/Bulat_Super-FAN_Integrated_Facial_CVPR_2018_paper.html
32. Cheng, Z., Zhu, X., Gong, S.: Characteristic regularisation for super-resolving face images. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2020, pp. 2424–2433. IEEE, Piscataway (2020). https://doi.org/10.1109/WACV45572.2020.9093480
33. Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a GAN to learn how to do image degradation first. In: Proceedings of 15th European Conference on Computer Vision - ECCV 2018, Part VI. Lecture Notes in Computer Science, vol. 11210, pp. 187–202. Springer, Berlin (2018). https://doi.org/10.1007/978-3-030-01231-1_12
34. Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C.: 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, pp. 81–88. IEEE, Piscataway (2011)
35. Chen, J., Chen, J., Chao, H., Yang, M.: Image blind denoising with generative adversarial network based noise modeling. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pp. 3155–3164. IEEE, Piscataway (2018)
36. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680. Curran Associates, Inc. Red Hook (2014). http://papers.nips.cc/paper/5423-generative-adversarial-nets
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
38. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 2242–2251. IEEE, Piscataway (2017). https://doi.org/10.1109/ICCV.2017.244
39. Jo, Y., Yang, S., Kim, S.J.: Investigating loss functions for extreme super-resolution. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1705–1712. IEEE, Piscataway (2020). https://doi.org/10.1109/CVPRW50498.2020.00220
40. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 4401–4410. IEEE, Piscataway (2019)
41. Vieira, T.F., Bottino, A., Laurentini, A., De Simone, M.: Detecting siblings in image pairs. The Visual Comput. 30(12), 1333–1345 (2014). https://doi.org/10.1007/s00371-013-0884-3

42. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition Emotion 24(8), 1377–1388 (2010)
43. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE, Piscataway (2017)
44. Lugmayr, A., Danelljan, M., Timofte, R.: Unsupervised learning for real-world super-resolution. In: 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, pp. 3408–3416. IEEE, Piscataway (2019). https://doi.org/10.1109/ICCVW.2019.00423
45. Lugmayr, A., et al.: Ntire 2020 challenge on real-world image super-resolution: Methods and results. In: CVPR Workshops. IEEE, Piscataway (2020)
46. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a ”Completely Blind” image quality analyzer. IEEE Signal Process. Lett. 20(3), 209–212 (2013)
47. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. 21(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050
48. N., V., D., P., Bh., M.C., Channappayya, S.S., Medasani, S.S.: Blind image quality evaluation using perception based features. Twenty First National Conference on Communications, NCC 2015, pp. 1–6. IEEE, Piscataway (2015). https://doi.org/10.1109/NCC.2015.7084843
49. Ma, C., Yang, C., Yang, X., Yang, M.: Learning a no-reference quality metric for single-image super-resolution. Comput. Vis. Image Underst. 158, 1–16 (2017). Available from: https://doi.org/10.1016/j.cviu.2016.12.009
50. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: Computer Vision – ECCV 2018 Workshops, pp. 334–355. Springer, Berlin (2019)
51. Lennan, C., Nguyen, H., Tran, D.: Image quality assessment (2018) https://github.com/idealo/image-quality-assessment
52. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multi-scale structural similarity for image quality assessment. In: Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, pp. 1398–1402. IEEE, Piscataway (2003)
53. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18-22, 2018, pp. 6228–6237. IEEE, Piscataway (2018). http://openaccess.thecvf.com/content_cvpr_2018/html/Blau_The_Perception-Distortion_Tradeoff_CVPR_2018_paper.html
54. Laparra, V., Ballé, J., Berardino, A., Simoncelli, E.P.: Perceptual image quality assessment using a normalized laplacian pyramid. Human Vision and Electronic Imaging, HVEI 2016, pp. 1–6. SPIE, Washington DC (2016). http://ist.publisher.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000016/art00008
55. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. CoRR, abs/2004.07728 (2020). https://arxiv.org/abs/2004.07728
56. Keys, R.G.: Cubic Convolution Interpolation for Digital Image Processing. IEEE Trans. Acoust. Speech Signal Process. 29, 1153–1160 (1981)
57. Zhang, K., Zuo, W., Zhang, L.: Deep plug-and-play super-resolution for arbitrary blur kernels. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 1671–1681. IEEE, Piscataway (2019)

How to cite this article: Aakerberg, A., Nasrollahi, K., Moeslund, T.B.: Real-world super-resolution of face-images from surveillance cameras. IET Image Process. 16, 442–452 (2022). https://doi.org/10.1049/ipr2.12359