EasyChair Preprint

№ 7807

SRTGAN: Triplet Loss Based Generative Adversarial Network for Real-World Super-Resolution

Dhruv Patel, Abhinav Jain, Simran Bawkar, Manav Khorasiya, Kalpesh Prajapati, Kishor Upla, Kiran Raja, R. Raghavendra and Christoph Busch

April 18, 2022

SRTGAN: Triplet Loss based Generative Adversarial Network for Real-World Super-Resolution

Dhruv Patel[*1], Abhinav Jain[*1], Simran Bawkar[1], Manav Khorasiya[1], Kalpesh Prajapati[1]
Kishor Upla[1], Kiran Raja[2], Raghavendra Ramachandra[2], Christoph Busch[2]
[1]*Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India.*
[2]*Norwegian University of Science and Technology (NTNU), Gjøvik, Norway.*

***Abstract*—Many applications such as surveillance, forensics, satellite imaging, medical imaging, etc., demand High-Resolution (HR) images. However, obtaining an HR image is not always possible due to the limitations of optical sensors and their costs. An alternative solution called Single Image Super-Resolution (SISR) is a software-driven approach that aims to take a Low-Resolution (LR) image and obtain the HR image. Most supervised SISR solutions use the HR image as a target and do not include the information provided in the LR image, which could be valuable. In this work, we introduce Triplet Loss-based Generative Adversarial Network hereafter referred to as *SRTGAN* for image SR problem on real-world degradation. We introduce a new triplet-based adversarial loss function which exploits the information provided in the LR image by using it as a negative sample. Allowing the patch-based discriminator with access to both HR and LR images optimizes to better differentiate between HR and LR image; hence, improving the adversary. Further, we propose to fuse the adversarial loss, content loss, perceptual loss, and quality loss to obtain an SR image with high perceptual fidelity. We validate the superior performance of the proposed method over the other existing methods on the RealSR dataset in terms of quantitative and qualitative metrics.**

I. Introduction

Single Image Super-Resolution (SISR) refers to the reconstruction of High Resolution (HR) images from the Low Resolution (LR) input image. It is widely applicable in various fields such as medical, satellite imaging, forensics, security, robotics, and so on where LR images are abundant. It is an inherently ill-posed problem since obtaining the SR image from an LR image that might correspond to any patch of the HR image which is intractable. The most employed solutions are the supervised super-resolution methods due to the availability of the HR data and the development of many state-of-the-art models.

The SISR includes image deblurring, denoising, and super-resolution, which makes it a highly complex task to reconstruct HR image from the LR input. Due to recent technological advances, such as computational power and availability of data, there has been substantial development in various CNN architectures and loss functions to improve SISR methods [1]–[6]. These models have been primarily tested on the simulated datasets in which the downsampled LR images are obtained from the HR images by using known degradation such as bicubic sampling. For instance, Fig. 1 shows that the characteristics such as blur and details of true and bicubic downsampled LR images do not correspond exactly for both RealSR and DIV2KRK dataset. Such differences can be attributed to underlying sensor noise and unknown real-world degradation. Hence, the models perform well on those synthetically degraded images, they generalize poorly on the real-world dataset [7]. Further, most of the works have shown that increasing the number of the CNN layers, do increase the performance of the model up to some extent. However, they are unable to capture the high frequency information such as texture in the images as they rely on the pixel-wise losses and hence suffer from poor perceptual quality [8]–[11].

To address the issues mentioned above, the research community has also proposed using Generative Adversarial Networks (GANs) for SISR task. The first GAN-based framework called SRGAN [14] introduced the concept of perceptual loss for super-resolution, which has both content and adversarial losses. Subsequently, numerous GAN-based methods were introduced that have shown improvements in the super-resolution results [14]–[16]. GANs are also used for generating perceptually better images [14], [15], [17]. Motivated by such works, we propose SR using Triplet loss based GAN (SRTGAN) - a triplet loss based patch GAN comprising a generator trained in a multi-loss setting and a patch-based discriminator. The discriminator takes its input as an LR and corresponding HR and SR images. The inherent formulation of the triplet loss implicitly forces the discriminator to penalize LR foreground patch more than a LR background patch, which is missing when directly trained using vanilla GAN. In a vanilla GAN, in which we train the discriminator to classify image as HR or LR, each patch of the image would be scored in quality, the caveat being, even in HR images, the background is blurred and could be considered lower quality which would cause spurious loss to back propagate. We overcome this issue in triplet loss by introducing the LR and HR images as negative and positive samples, respectively and the SR image as an anchor.

Further, the proposed SR method is trained on a fusion of

* denotes equal contribution

RealSR dataset [12] DIV2KRK Dataset [13]

Fig. 1: True LR and corresponding bicubic downsampled LR image from ground truth HR of the RealSR dataset [12] and DIV2KRK dataset [13].

losses namely the content, multi-layer perceptual, triplet-based adversarial and quality assessment. Minimizing the content and perceptual losses train the network to keep the information consistent in the HR and LR images; however, training just on content loss can cause blurring. Perceptual loss overcomes this problem by not directly computing the loss on pixel values but on the latent space. The quality assessment network assesses an image based on human rankings; hence. minimizing this loss would lead to increasing the quality of image based on human perception. Finally, the adversarial loss is a loss function that is aimed to aid the generator in creating SR image through a min-max setting. Using such fusion of different loss functions, we obtain superior visual quality of SR results. Additionally, the proposed method also gains better PSNR and competing SSIM values on RealSR dataset, which are still not a valid metric as they fail to capture the perceptual features. Hence, we evaluate the performance of the proposed method on the perceptual measure, i.e., LPIPS score, and the proposed method outperforms the other state-of-the-art methods in the quantitative evaluation in addition to the visual performance.

The proposed novel method provides superior results on the synthetic data and outperforms to the current state-of-the-art methods on the Real-Data for upscaling factor $\times 4$. We demonstrate this using two datasets - RealSR (real-world degradation) [12] and DIV2KRK dataset (synthetic degradation) [13]. In addition, DIV2KRK which happens to be a synthetic dataset also has this variation as it has a highly complex and unknown degradation model. Hence, our proposed method has been trained and validated on these datasets proving the generalizability on the real-world data.

Our key contributions in this work can therefore be listed as:

- We introduce a new triplet-based adversarial loss function which exploits the information provided in the LR image by using it as a negative sample as well as the HR image which is used as a positive sample.
- Further, a patchGAN based discriminator network is utilized that assists the defined triplet loss function for training of the generator network.
- Additionally, the proposed SR method is trained on weighted combination of various losses. Such fusion of different loss functions leads to superior quantitative and subjective quality of SR results as illustrated in the results.
- The sensitivity of the proposed network is analyzed with experiments in the ablation study. Additionally, different experiments have been carried out in order to judge the potential of the proposed model. Also, quantitative and qualitative studies have been performed, which show the superiority of the proposed method-SRTGAN over the other state-of-the-art SR works.

The structure of the paper is designed in the following manner. Section II consists of the related work in the field. Further, the description of the proposed method is elaborated in Section III. It includes the proposed framework, the network architecture and loss formulation related to the training the Generator and Discriminator networks. The experimental validation of the proposed methods is presented in Section IV, followed by the conclusion of the work in Section VI.

II. Related Works

Dong et al. [18] proposed Convolutional Neural Network (CNN) based SR approach (referred to as SRCNN) where only three layers of convolution had been used to correct finer details in an upsampled LR image. Similarly, FSRCNN [5] and VDSR [19] were inspired by SRCNN with suitable modifications to further improve the performance. VDSR [19] is the first model that uses deep convolutional neural network and introduces the use of the residual design that helps in the faster convergence with improvement in SR performance. Such residual connection also helps to avoid vanishing gradient problem, which is the most common problem with deeper networks. Inspired from VDSR [19], a number of works [6], [14], [20]–[22] have been reported with the use of residual connection to train deeper models. Apart from residual network, an alternative approach using dense connections has been used to improve SR images in many recent networks [4], [23], [24]. The concept of attention was also used in several efforts [20], [25] to focus on important features and allow sparse learning for the SR task. Along similar lines, adversarial training [26] has been reported to obtain better perceptual SR results. Ledig et al. introduced adversarial learning for super-

resolution termed as SRGAN [14], which shows perceptual enhancement in the SR images even with low fidelity metrics such as PSNR and SSIM. Many works such as SRFeat [17] and ESRGAN [15] inspired by *SRGAN* have been recently reported to improve the perceptual quality in obtaining SR images. A variant of GAN, *TripletGAN* [27] demonstrated that a triplet loss setting will theoretically help the generator to converge to the given distribution. Inspired by TripletGAN, *PGAN* [28] has been proposed, which uses triplet loss to super-resolve medical images in a multistage manner.

Most of the works suggested above have the limitation of the training data prepared by artificial degradation such as bicubic downsampling. The CNN model trained on such dataset often generalizes poorly for real-world data where degradation is significantly different from that of bicubic downsampling (see Fig. 1). In order to make models generalizable to the real-world data, the supervised deep networks require real LR-HR paired images which is challenging. To this extent, Cai et al. [12] introduced the RealSR dataset and a baseline network named Laplacian Pyramid-based Kernel Prediction Network (LP-KPN) to recover real-world HR images. Many representative SR works on the RealSR dataset have been reported recently with considerations to real data [29]–[35].

Further, Cheng et al. [30] proposed an encoder–decoder based residual network for the real SR approach. They employ coarse to fine method, which gradually restores lost information and reduces the noise effects. Kwak et al. [35] introduced a fractal residual network to super-resolve the real-world LR image by using autoencoder-based loss function. They also proposed an inverse pixel shuffle at the beginning of the network architecture which helps them to reduce the number of training parameters. For high fidelity recovery of image details, Du et al. [33] proposed an Orientation-Aware Deep Neural Network (OA-DNN), which consists of several Orientation Attention Modules (OAMs). Here, in each OAM, three well-designed convolution layers are used to extract orientation-aware features in different directions. Further, Xu and Li [34] have introduced a spatial color attention-based network called SCAN for real SR. In SCAN, the spatial color attention module is designed to jointly exploit the spatial and spectral dependency within color images. In this direction, to improve the perceptual quality of SR images on realSR dataset, we propose a novel framework based on triplet loss in the manuscript inspired from [27].

Although, there have been previous attempts to incorporate the triplet loss optimization for super-resolution such as PGAN [28], which progressively super-resolve the images in a multistage manner, it has to be noted that they are specifically targeted to medical images, and in addition, the LR images used are obtained through a known degradation (such as bicubic sampling) and blurring (Gaussian filtering). Thus, it fails to address the real-world degradation. Using the triplet loss, the proposed patch-based discriminator is able to differentiate better between low and high resolution images, thereby improving the perceptual fidelity. To the best of our knowledge, the application of triplet loss to the real-world SISR problem has not been explored before. We therefore propose the new approach as explained in the upcoming section.

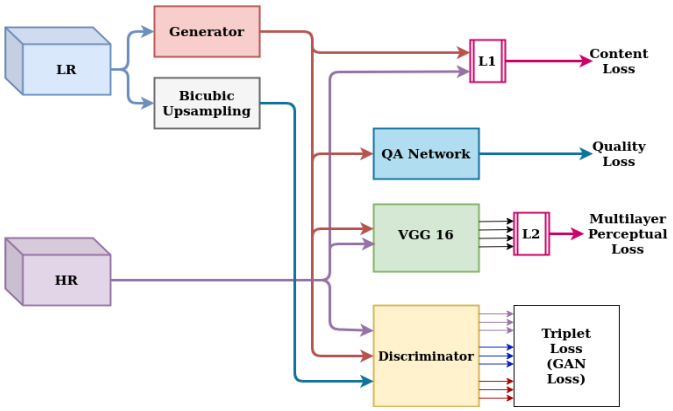

Fig. 2: The architecture framework of proposed method-SRTGAN for image SR.

III. Proposed Method

Fig. 2 shows the detailed framework of the proposed architecture. The proposed supervised SR method expects the LR and its corresponding HR images as the input. It performs SR on the LR image using the generator network and the training of the generator network rely on a fusion of losses consisting of content, perceptual, adversarial, and quality assessment. As depicted in Fig. 2, the content Loss is calculated as the pixel based difference as L_1 loss between the SR and HR images. It helps to guide the generator to preserve the content of HR image in the SR image. As the generator network is trained in an adversarial setting with the discriminator, we use a triplet setting for calculating GAN loss, which also boosts the stability of the learning. Apart from GAN loss, we incorporate perceptual loss, which is calculated as L_1 loss between features obtained from pre-trained VGG network as suggested in SRGAN [14]. Moreover, to improve the perceptual quality of SR image, we employ quality assessment loss based on Mean Opinion Score (MOS), which is introduced by Prajapati et al. [22]. The validation of each setting in the framework is demonstrated in the ablation section later. The different networks utilised in the proposed method is described in the following texts.

Generator Network (G): Fig. 3 shows the design architecture of the generator network published in [36]. Based on its functionality, the architecture can be divided into three different modules: Low-Level Information Extraction (LLIE), High-Level Information Extraction (HLIE), and SR reconstruction (SRRec) modules. In order to extract the low-level details $(i.e., I_l)$, LR input (I_{LR}) is first fed to the LLIE module. This module consists of a convolutional layer having kernel size 5 and 32 channels. Larger kernel is used, which leads to larger reception area for predicting the accurate low-level information. This can be expressed mathematically as,

$$I_l = f_{LLIE}(I_{LR}), \tag{1}$$

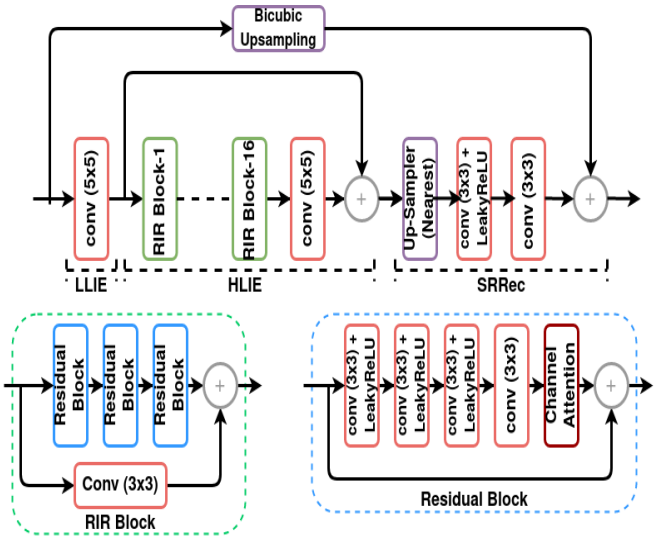

Fig. 3: The design architecture of the generator network [36].

where f_{LLIE} indicates the operation of the LLIE module.

The HLIE module uses the low-level information obtained from the LLIE module to extract edges and fine structural details present in HR image. The HLIE module consists of 16 Residual-In-Residual (RIR) blocks with one long skip connection. It is noted that the long skip connection is used to stabilize the network [14], [15], [22], [36]. Each RIR block is designed using 3 residual blocks with 1×1 convolutional layer in skip connection. The Residual block consists of 4 convolutional layers with a kernel size of 3 and a Channel Attention (CA) module. The CA module re-scales each channel individually on the basis of the statistical average of each channel [20]. As depicted in Fig. 3, skip connections are also used in residual blocks, which aids in stabilizing the training of deeper networks and resolving the vanishing gradient problem. The output from HLIE module can be expressed as,

$$I_h = f_{HLIE}(I_l). \tag{2}$$

Here, the f_{HLIE} denotes the function of the HLIE module. Now, the feature maps with high-level information are passed to the SR Reconstruction (SRRec) module, which comprises of 1 up-sampling block and 2 convolutional layers. This helps in mapping the high-level information feature maps to the required number of channels needed for SR output image (I_{SR}). This can be expressed as,

$$I_{SR} = f_{REC}(I_h), \tag{3}$$

where f_{REC} is the reconstruction function of the SRRec module. In each up-sampling block, the nearest neighbor is employed to perform $2\times$ up-sampling with convolutional layer having kernel size of 3 and feature maps of 32. Finally, a convolutional layer is used to map 32 channels into 3 channels of SR image in the generator network.

Discriminator (D) Network: We further use a PatchGAN [37] based discriminator network to distinguish foreground and background on patch with scale of 70×70 pixels. The proposed architecture is displayed in Fig. 4. It has been designed by following the guidelines suggested in the work of PatchGAN [37]. It consists of five convolutional layers with strided convolutions. The number of channels is increased by a factor of two after each convolution, excluding the last output layer where channel is kept one. It uses a fixed stride of two except for the second last and output layer where stride is set to 1. It is noted that a fixed kernel size of 4 is used for all layers throughout the discriminator network. Further, each

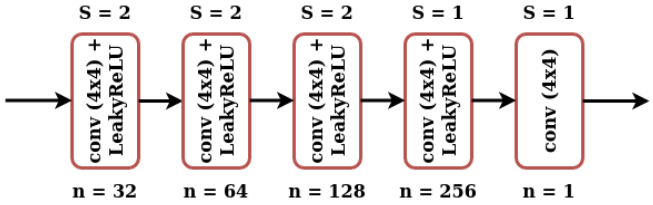

Fig. 4: The architecture design of discriminator. Here, n represents number of channels and S represents stride, respectively.

convolutional layer except the output layer is used with leaky ReLU activation having leaky constant of 0.2, and padding of size one. All intermediate convolutional layers except the first and last layer use Batch Normalisation.

Quality Assessment (QA) Network: To improve the perceptual quality of SR images, a novel quality-based score which serves as a loss function in training as inspired by Prajapati et al. [36] is also employed. The design of such deep network is inspired by the VGG, as shown in Fig. 5 [14]. The QA network is trained to mimic how humans rank images based on its quality; hence, adding the QA loss in the overall optimization improves the image quality based on human perception. Here, two paths have been used to provide input to the network instead of a single one, and both of these features are subtracted to proceed further. Each VGG block comprises of two convolutional layers, the second of which utilizes a stride value of 2 to reduce the spatial dimension of the features. To limit the number of trainable parameters, the network uses Global Average Pooling (GAP) layer instead of flattening layer. To overcome the issue of over-fitting, a drop-out technique is employed at fully connected layers. The KADID-10K [38] dataset, consisting of $10,050$ images, was used to train the QA network. Further, the dataset has been separated in 70%-10%-20% ratio for train-validate-test purposes respectively during the training process.

A. Loss Functions

As depicted in Fig. 2, the proposed framework is trained using a fusion of content loss (pixel-wise L_1 loss), GAN loss (triplet based), QA loss, and perceptual loss. Mathematically, we can describe the loss of generator by following formula:

$$L^{gen} = \lambda_1 L_{content} + \lambda_2 L_{QA} + \lambda_3 L^{G}_{GAN} + \lambda_4 L_{perceptual}. \tag{4}$$

The values of λ_1, λ_2, λ_3 and λ_4 are set empirically to 5, 2×10^{-7}, 1×10^{-1} and 5×10^{-1}, respectively. As mentioned

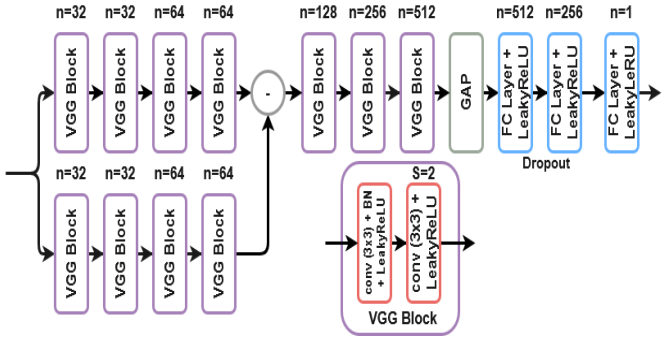

Fig. 5: The architecture of QA network [36].

earlier, the content loss has been used to preserve the content, which is an L_1 loss between original HR image (i.e., I_{HR}) and generated SR image (i.e., I_{SR}), and same can be expressed as,

$$L_{content} = \sum^{N} \|G(I_{LR}) - I_{HR})\|_1, \tag{5}$$

where N is the number of batch in training process, and G represents the function of generator.

As discussed earlier, to further improve the perceptual quality of SR image, a Quality Assessment (QA) loss is also introduced in the proposed framework, which is computed using trained QA network. It rates the quality of SR image on a scale of 1-5, with a higher value indicating better quality. This predicted value is used to calculate the QA loss i.e., L_{QA}, which is expressed as [36],

$$L_{QA} = \sum^{N} \big(5 - Q(I_{SR})\big), \tag{6}$$

where $Q(I_{SR})$ represents the quality score of SR image obtained from the proposed QA network. The perceptual loss $L_{perceptual}$ is used here to improve the perceptual similarity of the generated image with its ground truth, which can be expressed as,

$$L_{perceptual} = \sum^{N} \big[\sum_{i=1}^{4} MSE(F^i_{HR}, F^i_{SR})\big]. \tag{7}$$

Here, $\sum^{N}[\cdot]$ denotes an average operation of all super resolved (fake data) in the mini-batch, $MSE(a, b)$ represents Mean Square Error (MSE) between a and b, F^i: Normalised features taken from $layers[i]$ and $layers$ = [$relu_{12}$, $relu_{22}$, $relu_{33}$, $relu_{43}$]. Here, $layers$ is the list of four layers of VGG-16 used for the calculation of perceptual loss [9]. Such loss is calculated as the MSE between the normalized feature representations of generated image (F_{SR}) and ground truth HR (F_{HR}) obtained from a pre-trained VGG-16 network. It is not dependent on low-level per-pixel information that leads to blurry results. Instead, it depends on the difference in high-level feature representations which helps to generate images of high perceptual quality. In addition, the idea of using multi-layer feature representations adds to its robustness.

The GAN loss used here is a triplet-based loss function to a patch-based discriminator. An image consists of 2 parts, Background and Foreground; according to human perceptions, we rate images to be higher quality based on the foreground, which is the focus of the image. On the other hand, background between LR and HR images is hard to differentiate as shown in Fig 7. When using a Patch GAN with L_2 loss, network try to give SR and HR images with opposite labels, and they do not consider the fact that a patch could correspond to the background image, and it will be very hard for the discriminator to give an LR background and an HR background opposite labels. A background patch could lead to a high loss and cause instability and noise in training. However, in the case of foreground patches, this idea will work well. To solve this problem, we introduce the use of triplet loss: Instead of forcing the discriminator output for HR and SR to be opposite labels, we calculate the loss using the relative output produced by the discriminator for HR, LR, and SR images. Here let us say for a background image, the quality of the image for HR, LR, and SR will be similar; hence the loss would not be large; however, in the case of a foreground image, the discriminator output of HR and LR is going to vary as the patch quality will be different, and the discriminator and generator would be trained likewise. Thus, the triplet loss has three variables in it - positive, negative, and anchor. The cost function is such that it minimizes the distance between the anchor and positive, while maximizing the distance between the anchor and the negative. For the generator, the anchor is defined as the SR image (I_{SR}), the positive is defined as the HR (I_{HR}), and the negative is low-resolution image ($n(I_{LR})$) where n is the bicubic upsampling function. For discriminator training, the positive and negative are interchanged. Thus, the generator and discriminator losses are then defined as,

$$L^G_{GAN} = \sum^{N} \big[MSE(D(I_{SR}), D(I_{HR})) - MSE(D(I_{SR}), D(n(I_{LR}))) + 1\big],$$

$$L^D_{GAN} = \sum^{N} \big[MSE(D(I_{SR}), D(n(I_{LR}))) - MSE(D(I_{SR}), D(I_{HR})) + 1\big].$$

Here, $MSE(a, b)$ represents MSE between a and b; n denotes upsampling factor. Here, L^D_{GAN} and L^G_{GAN} loss functions are used to train the discriminator and generator networks, respectively. This triplet based GAN loss teaches the Generator to generate sharp and high-resolution images by trying to converge SR embeddings $D(I_{SR})$ and HR embeddings $D(I_{HR})$ and diverge SR embeddings with LR embeddings $D(n(I_{LR}))$ from the Discriminator. Simultaneously, it also trains the patch-based Discriminator to distinguish the generated SR image from ground-truth HR.

IV. Experimental Results

In order to see the effectiveness of the proposed method, we have conducted numerous experiments on two different datasets. All experiments have been conducted on a computer with Intel Xeon(R) CPU with 128GB RAM and NVIDIA Quadro P5000 GPU with 16GB memory. Hyper-parameter tuning, visual and quantitative evaluations of the proposed

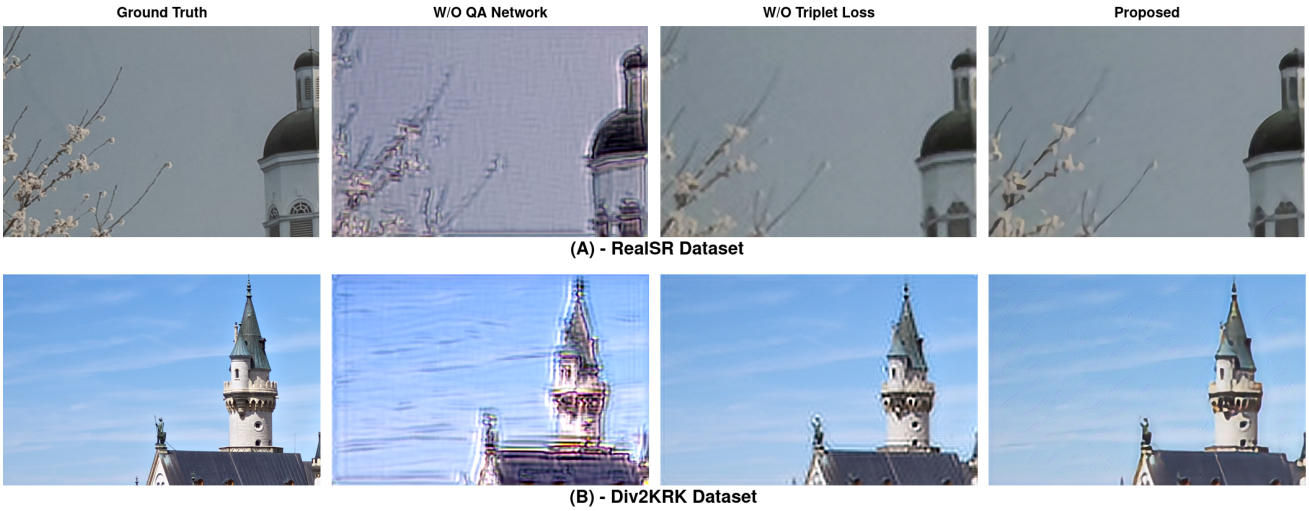

Fig. 6: The comparison of the SR results obtained using the proposed method-*SRTGAN* and without QA Loss and without Triplet Loss (Vanilla GAN Loss) method on (A)-RealSR validation dataset [12] and (B)-DIV2KRK dataset [13].

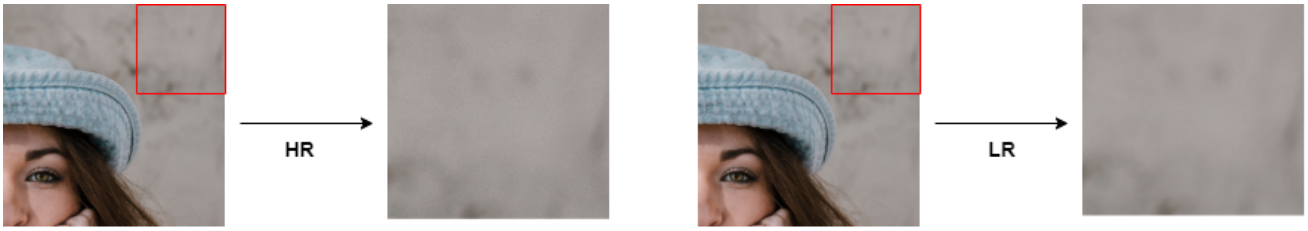

Fig. 7: Comparison of background patch in LR and HR images.

approach with other state-of-the-art methods have been elaborated in the following subsections.

A. Training Details and Hyper-parameter Tuning

To perform supervised training using the proposed approach, we use RealSR dataset [12]. In this dataset, paired LR-HR images on the same scene are captured by adjusting the focal length of a digital camera. An image registration algorithm is developed to progressively align the image pairs at different resolutions. A total of 400 images have been used in the dataset for training of the proposed model. For validation, an additional 100 pairs of LR-HR images are used, which are provided in the same dataset. Finally, for testing purpose, images of validation sets of DIV2KRK [13] and testing images of RealSR [12] datasets are used. During training phase, the LR images are passed through different augmentations such as random horizontal flipping, random rotation of 0 or 90, and random cropping operations. We train our model using Adam optimizer upto $1,500$ iterations with batch size of 32. We keep β_1 value as 0.9, and set learning rate at 1×10^{-5}. We decrease this learning rate by half after every 500 iterations. Further, the total number of trainable parameters of generator and discriminator networks are $3.7M$ and $2.7M$, respectively.

We have also used QA network-based loss in order to improve the perceptual quality of the SR images. The reference of this method has been taken from the work of [36]. In addition, a Triplet Loss is used in loss optimization to enhance the visual appearance and perception of the SR images to make them more realistic. It uses the output from the previous stage as a baseline, thereby being able to improve the quality of the SR images in a stepwise manner.

TABLE I: The quantitative assessment of the proposed method without QA and without discriminator networks carried out on RealSR validation dataset [12].

Method	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓
w/o Triplet Loss (Vanilla GAN Loss)	25.879	0.72199	0.37095
w/o QA Network	16.126	0.39542	0.51217
Proposed	**26.47283**	**0.754585**	**0.283878**

B. Ablation Study

We show the experimental justification for employing Discriminator network in addition to Triplet Loss and QA network to improve the quality of SR images in the proposed method in this section. The quantitative and qualitative results of this experiment carried out on RealSR dataset [12] are depicted in Table I and Fig. 6, respectively. Our approach shows better SR results not only on synthetic data but also on the real world data for RealSR dataset. The quantitative assessments in terms of different distortion metrics such as PSNR & SSIM and in terms of perceptual measure such as LPIPS show that the proposed method with QA network and Triplet

TABLE II: The quantitative comparison of the proposed and other existing SR methods on RealSR validation and DIV2KRK datasets.

Method	**PSNR** ↑	**SSIM** [40] ↑	**LPIPS** [39] ↓	**PSNR** ↑	**SSIM** [40] ↑	**LPIPS** [39] ↓
	DIV2KRK [41] Dataset			*RealSR [12] Dataset*		
Bicubic	23.89	0.6478	0.5645	25.74	0.7413	0.4666
ZSSR [42]	24.05	0.6550	0.5257	25.83	0.7434	0.3503
KernelGAN [43]	24.76	0.6799	0.4980	24.09	0.7243	0.2981
DBPI [44]	24.92	0.7035	0.4039	22.36	0.6562	0.3106
DAN [45]	26.07	0.7305	0.4045	26.20	0.7598	0.4095
IKC [46]	25.41	0.7255	0.3977	25.60	0.7488	0.3188
SRResCGAN [16]	24.00	0.6497	0.5054	25.84	0.7459	0.3746
Proposed	**24.17**	**0.6956**	**0.3341**	**26.47**	**0.7546**	**0.2838**

Loss perform better (see Table I) when compared to the performance obtained using the proposed method without those modules. Further, as displayed in Fig. 6, the SR results obtained using the proposed network with QA network and Triplet Loss are perceptually better than the proposed approach without incorporating these modules. The efficacy of the proposed method can be verified by closely analyzing the visual appearance of the outputs of different methods. It is observed that the model without QA Network generates blurry output and variation in the image's natural color. The model without Triplet Loss (Vanilla GAN Loss) quite resembles the color as expected in the ground truth; however, it fails to sharpen the image around the edges resulting in blurring. The superiority of the proposed method can be seen as it is able to generate SR images, displaying an adequate level of sharpening around the edges and preserving the color-coding of the original image. Here, one can easily deduce the perceptual improvement from our proposed approach by observing at Fig. 6.

C. Quantitative Analysis

Generally, for comparison of SR results obtained using the proposed method with other state-of-the-art methods, the PSNR and SSIM values are estimated, which are the standard measurements for the SR problem. However, these metrics do not entirely justify the quality based on human perception. Therefore, we estimate an additional metric, called LPIPS [39] which is a deep network based full-reference perceptual quality assessment score. A low LPIPS value indicates better visual quality.

Table II shows the comparison of all three metrics obtained on the RealSR validation [12] and DIV2KRK [13] datasets. The proposed method-SRTGAN obtains better SSIM and LPIPS values than the other existing state-of-the-art methods indicating the superiority of the proposed method in terms of quantitative evaluation. In terms of PSNR, our proposed method outperforms to other state-of-the-art methods on the RealSR dataset [12], whereas it performs competitively to other methods on the DIV2KRK dataset [13]. The perceptual metric, LPIPS obtained using the proposed method is significantly better for both datasets using the proposed method (see Table II).

D. Qualitative Analysis

In addition to quantitative comparison, in this section, we demonstrate the effectiveness of the proposed method through visual inspection. We compare the SR results on an image of RealSR validation dataset [12] in Fig. 8 in which original HR image (Ground Truth) is available, and on two sample images of DIV2KRK dataset [13] (see Fig. 9 and Fig. 10) where original images are not provided. Additionally, we use different state-of-the-art pre-trained networks such as ZSSR [42], KernelGAN [43], DBPI [44], DAN [45], IKC [46], and SRResCGAN [16] to perform SR comparison. It can be observed from these SR results that the amount of noise present in the SR image of the proposed method is reduced considerably, and the clarity of the image is increased as compared to the recently proposed competing methods. In addition, our method is able to produce similar colors as the ground truth while competing approaches such as IKC and KernelGAN over-boosts the colors in the generated images.

It can be concluded that the proposed method for SISR generates better quality SR images having fewer noise artifacts than those obtained with other state-of-the-art methods. Furthermore, this result is also demonstrated by the quantitative evaluation of various quality metrics (see Table II) and obtained perceptual quality on various datasets as illustrated in different figures (see Fig. 6 - Fig. 10).

V. Limitations

The proposed work obtains better results on real-world data; however, we note certain limitations as well. The network is stable only when fine-tuned for all the losses. As we can observe in Fig. 6, the removal of the QA loss leads to undesirable outputs. Thus, fine-tuning of each loss is an expensive process. Another limitation for using the current model is that the generator and discriminator are trained in a supervised manner and hence it requires true HR-LR image pairs which can be difficult to obtain as this will need the same image to be clicked by cameras of two different resolutions. However, our work can be easily extended to unsupervised approach, as the core idea of generative modeling is to treat such unsupervised problems in a supervised manner.

VI. Conclusion

We have proposed a solution to the problem of SISR based on TripletGAN that fuses the novel triplet loss and no-reference quality loss along with the other conventional losses. We further modify the design of discriminator to be a patch-based discriminator for improving image quality at the scale of local image patches. The triplet loss uses not only the high-resolution image but also the low-resolution image and hence, it captures the essential information required in the SR image. Through experiments, we have demonstrated that the proposed method-SRTGAN can super-resolve images by a factor of $\times 4$ of original LR image with improved perceptual fidelity. As demonstrated through the SR results, the proposed method is able to obtain superior performance to competing methods in the SR task in terms of perceptual quality.

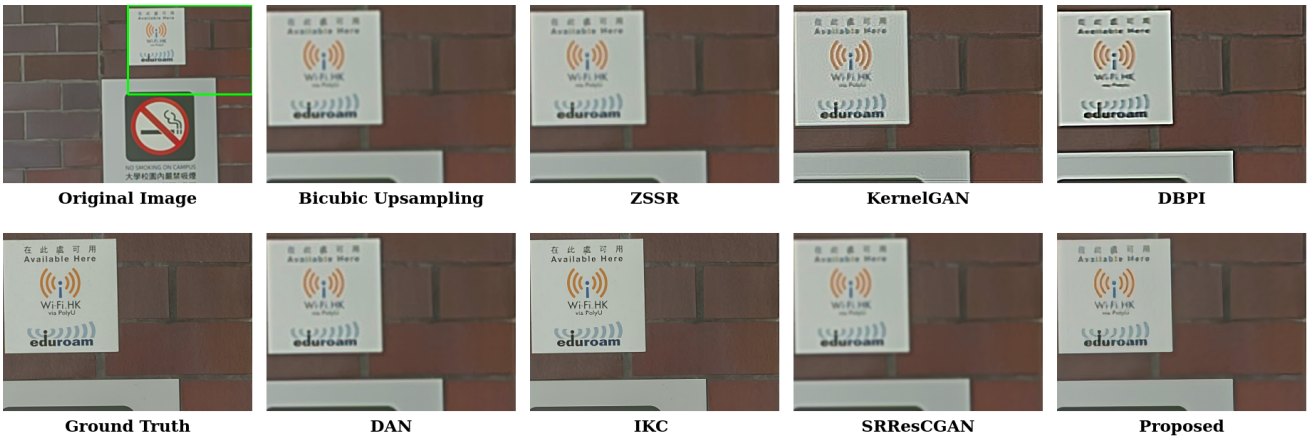

Fig. 8: The comparison of the SR results obtained using the proposed and other state-of-the-art methods on RealSR validation dataset [12].

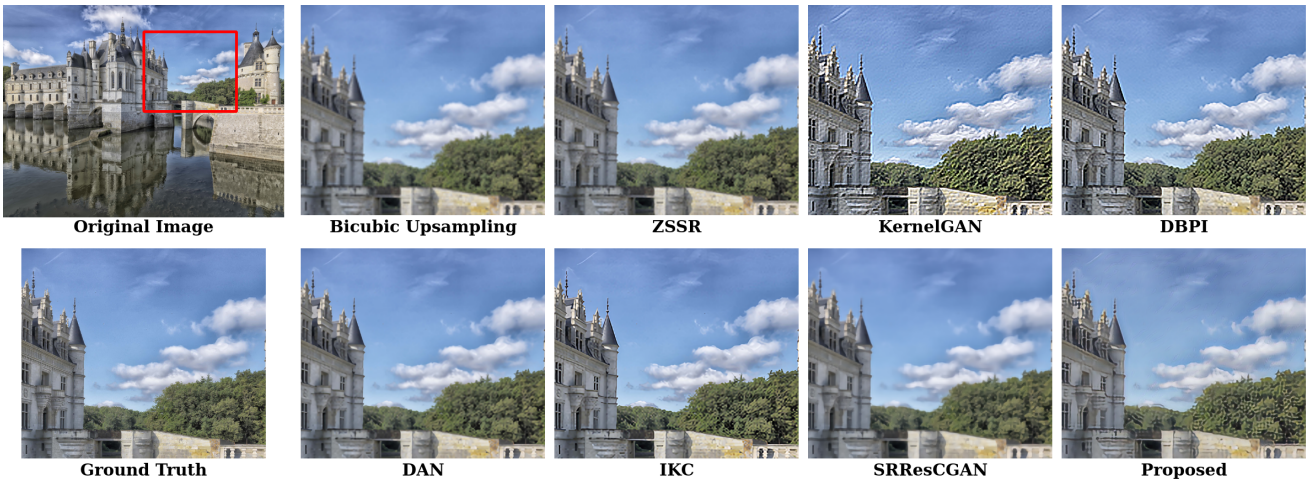

Fig. 9: The comparison of the SR results obtained using the proposed and other state-of-the-art methods on DIV2KRK dataset [41].

Fig. 10: The comparison of the SR results obtained using the proposed and other state-of-the-art methods on DIV2KRK dataset [41].

REFERENCES

[1] Z. Fan, D. Bi, L. Xiong, S. Ma, L. He, and W. Ding, "Dim infrared image enhancement based on convolutional neural network," *Neurocomputing*, vol. 272, pp. 396 – 404, 2018.

[2] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *The IEEE Conference on CVPR*, vol. 1, no. 4, 2017.

[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[4] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE ICCV*, 2017, pp. 4799–4807.

[5] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV*, Oct 2016, pp. 391–407.

[6] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolutaion," *2017 IEEE Conference on CVPR Workshops*, pp. 1132–1140, 2017.

[7] N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin, "Accurate blur models vs. image priors in single image super-resolution," in *2013 IEEE ICCV*, 2013, pp. 2832–2839.

[8] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," Jan. 2016, 4th International Conference - ICLR 2016.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, pp. 694–711.

[10] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proceedings of the 30th International Conference on NeurIPS*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 658–666.

[11] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," *CoRR*, vol. abs/1511.05666, 2016.

[12] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *ICCV*, October 2019, pp. 3086–3095.

[13] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *NeurIPS*, 2019.

[14] C. Ledig, L. Theis, F. Huszár *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on CVPR*, 2017, pp. 4681–4690.

[15] X. Wang, K. Yu, S. Wu *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops*. Cham: Springer International Publishing, 2019, pp. 63–79.

[16] R. Muhammad Umer, G. Luca Foresti, and C. Micheloni, "Deep Generative Adversarial Residual Convolutional Networks for Real-World Super-Resolution," in *Proceedings of the IEEE Conference on CVPR Workshops*, 2020, pp. 438–439.

[17] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "Srfeat: Single image super-resolution with feature discrimination," in *Proceedings of the ECCV*, 2018, pp. 439–455.

[18] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[19] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE CVPR*, June 2016, pp. 1646–1654.

[20] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the ECCV*, 2018, pp. 286–301.

[21] Y. Li, E. Agustsson, S. Gu, R. Timofte, and L. Van Gool, "Carn: Convolutional anchored regression network for fast and accurate single image super-resolution," vol. 11133 LNCS, Leal-Taixé, Laura. Springer, 2019, pp. 166–181.

[22] K. Prajapati, V. Chudasama, H. Patel, K. Upla, R. Ramachandra, K. Raja, and C. Busch, "Unsupervised single image super-resolution network (usisresnet) for real-world data using generative adversarial network," in *Proceedings of the IEEE/CVF Conference on CVPR Workshops*, June 2020.

[23] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on CVPR*, 2018, pp. 2472–2481.

[24] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on CVPR*, 2018, pp. 1664–1673.

[25] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *ECCV*. Springer, 2020, pp. 56–72.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," in *Advances in NeurIPS 27*.

[27] G. Cao, Y. Yang, J. Lei, C. Jin, Y. Liu, and M. Song, "Tripletgan: Training generative model with triplet loss," *CoRR*, vol. abs/1711.05084, 2017. [Online]. Available: http://arxiv.org/abs/1711.05084

[28] D. Mahapatra and B. Bozorgtabar, "Progressive generative adversarial networks for medical image super resolution," *CoRR*, vol. abs/1902.02144, 2019. [Online]. Available: http://arxiv.org/abs/1902.02144

[29] Y. Shi, H. Zhong, Z. Yang, X. Yang, and L. Lin, "Ddet: Dual-path dynamic enhancement network for real-world image super-resolution," *IEEE Signal Processing Letters*, vol. 27, pp. 481–485, 2020.

[30] G. Cheng, A. Matsune, Q. Li, L. Zhu, H. Zang, and S. Zhan, "Encoder-decoder residual network for real super-resolution," in *CVPR Workshops*, June 2019.

[31] R. Feng, J. Gu, Y. Qiao, and C. Dong, "Suppressing model overfitting for image super-resolution networks," in *CVPR Workshops*, June 2019.

[32] S. Gao and X. Zhuang, "Multi-scale deep neural networks for real image super-resolution," in *The IEEE Conference on CVPR Workshops*, June 2019.

[33] C. Du, H. Zewei, S. Anshun *et al.*, "Orientation-aware deep neural network for real image super-resolution," in *The IEEE Conference on CVPR Workshops*, June 2019.

[34] X. Xu and X. Li, "Scan: Spatial color attention networks for real single image super-resolution," in *The IEEE Conference on CVPR Workshops*, June 2019.

[35] J. Kwak and D. Son, "Fractal residual network and solutions for real super-resolution," in *The IEEE Conference on CVPR Workshops*, June 2019.

[36] K. Prajapati, V. Chudasama, H. Patel, K. Upla, K. Raja, R. Raghavendra, and C. Busch, *Unsupervised Real-World Super-resolution Using Variational Auto-encoder and Generative Adversarial Network*, 02 2021, pp. 703–718.

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on CVPR*, 2017, pp. 5967–5976.

[38] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–3.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on CVPR*, 2018, pp. 586–595.

[40] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[41] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *2017 IEEE Conference on CVPR Workshops*, 2017, pp. 1122–1131.

[42] A. Shocher, N. Cohen, and M. Irani, "Zero-shot super-resolution using deep internal learning," in *2018 IEEE/CVF Conference on CVPR*, 2018, pp. 3118–3126.

[43] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind Super-Resolution Kernel Estimation using an Internal-Gan," in *33rd Conference on NeurIPS*, 2019, pp. 284–293.

[44] J. Kim, C. Jung, and C. Kim, "Dual Back-Projection-Based Internal Learning for Blind Super-Resolution," in *IEEE Signal Processing Letters*, vol. 27, 2020, pp. 1190–1194.

[45] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," *Advances in NeurIPS*, vol. 33, 2020.

[46] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *The IEEE Conference on CVPR*, June 2019, pp. 1604–1613.