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

October 24, 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], and Christoph Busch[2]

[1] Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India
{dhruv.r.patel14, abhinav98jain, sim017bawkar, manavkhorasiya, kalpesh.jp89, kishorupla}@gmail.com
[2] Norwegian University of Science and Technology (NTNU), Gjøvik, Norway.
{kiran.raja, raghavendra.ramachandra, christoph.busch}@ntnu.no

Abstract. Many applications such as forensics, surveillance, 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 ground truth 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 as *SRTGAN* for Image Super-Resolution problem on real-world degradation. We introduce a new triplet-based adversarial loss function that 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 images; hence, improving the adversary. Further, we propose to fuse the adversarial loss, content loss, perceptual loss, and quality loss to obtain Super-Resolution (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.

1 Introduction

Single Image Super-Resolution (SISR) refers to reconstructing a High Resolution (HR) image from an input Low Resolution (LR) image. It has broad applications in various fields, including satellite imaging, medical imaging, forensics, security, robotics, where LR images are abundant. It is an inherently ill-posed problem since obtaining the SR image from an LR image might correspond to any patch of the ground truth HR image, which is intractable. The most employed solutions

⋆ denotes equal contribution

RealSR dataset [6] DIV2KRK Dataset [7]

Fig. 1: True LR and corresponding bicubic downsampled LR image from ground truth HR of the RealSR dataset [6] and DIV2KRK dataset [7]

are the supervised super-resolution methods due to the availability of ground truth information and the development of many novel methods.

Reconstructing the HR image from LR input includes image deblurring, denoising, and super-resolution operations which makes the SISR a highly complex task. 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 [1–5]. These models have been primarily tested on the synthetic datasets. Here, the LR images are downsampled from the ground truth HR images by using known degradation model such as bicubic downsampling. For instance, Fig. 1 shows that the characteristics like blur and that details of true and bicubic downsampled LR images do not correspond exactly for both RealSR [6] and DIV2KRK dataset [7]. 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 [8]. Further, most of the works have shown that adding more CNN layers does increase the performance of the model by 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 [9–12].

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 [13], introduced the concept of perceptual loss, calculated from high-level feature maps, and tried to solve the problem of poor perceptual fidelity as mentioned before. Subsequently, numerous GAN-based methods were introduced that have shown improvements in the super-resolution results [13–15]. GANs are also used for generating perceptually better images [13, 14, 16]. 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 with a patch-based discriminator.

Our proposed method - SRTGAN gains superior Peak Signal-to-Noise Ratio (PSNR) and competing Structural Similarity Index (SSIM) [17] values on the RealSR dataset (real-world degradation) [6], which still cannot be considered a valid metric as they fail to capture the perceptual features. Hence, we also evaluate our performance on the perceptual measure, i.e. Learned Perceptual Image Patch Similarity (LPIPS) [18] score. Our SRTGAN outperforms the

other state-of-the-art methods in the quantitative evaluation of LPIPS and visual performance on the RealSR dataset. It also provides superior LPIPS results on the DIV2KRK dataset [7] (synthetic degradation). All our experiments on both RealSR and DIV2KRK datasets are done for an upscaling factor of $\times 4$. Even though DIV2KRK happens to be a synthetic dataset, 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 propose a new triplet-based adversarial loss function that 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.
- A patchGAN-based discriminator network is utilized that assists the defined triplet loss function to train the generator network.
- The proposed SR method is trained on a linear combination of losses, namely the content, multi-layer perceptual, triplet-based adversarial, and quality assessment. Such fusion of different loss functions leads to superior quantitative and subjective quality of SR results as illustrated in the results.
- Additionally, different experiments have been conducted in the ablation study to judge the potential of our proposed approach. The superiority of the proposed method over other novel SR works has been demonstrated from the undertaken quantitative and qualitative studies.

The structure of the paper is designed in the following manner. Section 2 consists of the related work in the field. Section 3 includes the proposed framework, the network architecture, and loss formulation for training the Generator and Discriminator networks. The experimental validation is presented in Section 4, followed by the limitation and conclusion in Section 5 and 6 respectively.

2 Related Works

A Convolutional Neural Network (CNN) based SR approach (referred as SRCNN) was proposed by Dong et al. [2], where only three layers of convolution were used to correct finer details in an upsampled LR image. Similarly, FSRCNN [4] and VDSR [19] were inspired by SRCNN with suitable modifications to further improve the performance. VDSR [19] is the first model that uses a deep CNN and introduces the use of residual design that helps in the faster convergence with improvement in SR performance. Such residual connection also helps to avoid the vanishing gradient problem, which is the most common problem with deeper networks. Inspired by VDSR [19], several works [5, 13, 20–22] have been reported with the use of a residual connection to train deeper models. Apart from a residual network, an alternative approach using dense connections has been used to improve SR images in many recent networks [3, 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 problem. Similarly, adversarial training [26] has been shown to obtain better perceptual SR results. Ledig et al. introduced

adversarial learning for super-resolution termed as SRGAN [13], which shows perceptual enhancement in the SR images even with low fidelity metrics such as PSNR and SSIM. Recent works such as SRFeat [16] and ESRGAN [14], which were inspired by SRGAN, have also reported improvements in 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.

The limitation of the majority of the work mentioned above is the use of artificially degraded training data, such as bicubic downsampling. The CNNs typically fail to generalise well on the real-world data, because real-world degradation is considerably different than bicubic downsampling (see Fig. 1). The supervised approaches need real LR-HR pairs in order to generalise to real-world data, which is challenging. For recovering real-world HR images, Cai et al. [6] introduced the RealSR dataset and a baseline network called Laplacian Pyramid-based Kernel Prediction Network (LP-KPN). Thereafter, several research works for SR have been conducted on the RealSR dataset, considering factors from real data into account. [29–35].

Further, Cheng et al. suggested a residual network based on an encoder-decoder architecture for the real SR problem [30]. A coarse-to-fine approach was used by them, where lost information was gradually recovered and the effects of noise were reduced. By adopting an autoencoder-based loss function, a fractal residual network was proposed by Kwak et al. [35] to super-resolve real-world LR images. At the outset of network architecture, an inverse pixel shuffle was also proposed by them to minimise the training parameters. Du et al. [33] suggested an Orientation-Aware Deep Neural Network(OA-DNN) for recovering of images with high fidelity. It is made up of many Orientation Attention Modules(OAMs) which are designed for extracting orientation-aware features in different directions. Additionally, Xu and Li have presented SCAN, a spatial colour attention-based network for real SR [34]. Here, the attention module simultaneously exploits spectral and spatial dependencies present in colour images. In this direction, we provide a novel framework based on triplet loss in the manuscript inspired by [27] to enhance the perceptual quality of SR images on the realSR dataset.

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 real-world degradation. Using the triplet loss, the proposed patch-based discriminator can better distinguish between generated and high-resolution images, thereby improving the perceptual fidelity. To the best of our knowledge, the utilization 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.

3 Proposed Method

Fig. 2 shows the detailed training framework of our proposed method. The proposed supervised SR method expects the LR and its corresponding ground truth HR image as the input. It performs super-resolution on the LR image using the generator network, which is trained in a multi-loss setting using a fusion of losses namely content, perceptual, adversarial, and quality assessment. As depicted in Fig. 2, the content Loss is calculated as L_1 loss (pixel-based difference) between the generated(SR) and ground truth(HR) images. It assists the generator in preserving the content of ground truth HR. As the generator network is trained in an adversarial setting with the discriminator, we use a triplet-based GAN loss, which also boosts the stability of the learning. Apart from the GAN loss, we incorporate multi-layer perceptual loss, which is calculated as L_2 loss between the features of HR and SR, obtained from a pre-trained VGG network as suggested in SRGAN [13]. Moreover, we also use a quality assessment loss based on Mean Opinion Score (MOS) for improving the perceptual quality of generated images [22]. The validation of each setting in the framework is demonstrated in the ablation section later.

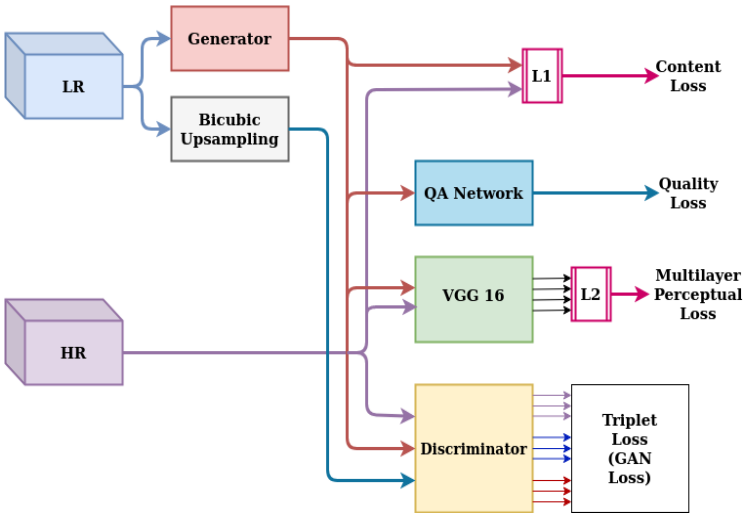


Fig. 2: The training framework of our proposed method - SRTGAN.

Generator Network (G): The design of generator network is shown in Fig. 3, which was published in [36]. The architecture can be divided into Feature Extraction (Low-level Information Extraction (LLIE), High-level Information Extraction (HLIE)) and Reconstruction (SR reconstruction (SRRec)) modules based on their functionality. The LLIE module is initially fed with LR input (I_{LR}) for extracting the low-level details (*i.e.*, I_l). It consists of a convolutional layer with kernel size 3 and 32 channels. This can be expressed mathematically

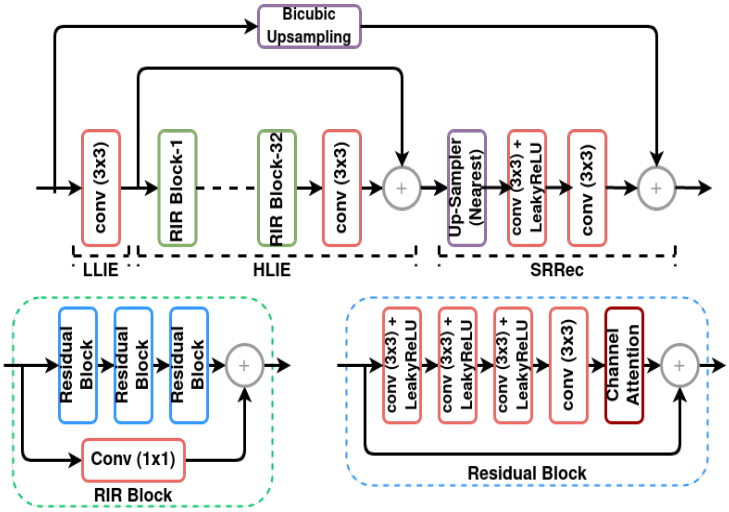


Fig. 3: Generator network [36].

as,

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

The edges and fine structural details present in the LR image are extracted by the HLIE module using the low-level information I_l. HLIE module comprises of 32 Residual-In-Residual (RIR) blocks, one 3×3 convolutional layer, and have one long skip connection. The long skip connection here stabilizes the network training [13,14,22,36]. Each RIR block is created using three residual blocks and a skip connection with a 1×1 convolutional layer. The Residual Block comprises of four 3×3 convolutional layers with a serially attached Channel Attention (CA) module. Using the statistical average of each channel, each channel is independently re-scaled via the CA module [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,

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

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

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

where the reconstruction function of the SRRec module is f_{REC}. The nearest neighbour is used to perform a $2\times$ upsampling with a 3×3 convolutional layer and 32 feature maps in each up-sampling block. 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 a patch with scale of 70×70 pixels. The proposed architecture is shown in Fig. 4. It is designed by adhering to the recommendations made in the work of PatchGAN [37]. It consists of five convolutional layers with strided convolutions. After each convolution, the number of channels doubles, excluding the last output layer which has a single channel. The network uses a fixed stride of two except for the second last and last layer where the 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,

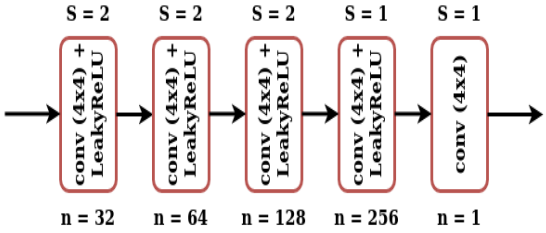


Fig. 4: Discriminator Network. Here, n stands for the number of channels, while S represents stride.

each convolutional layer except the output layer uses leaky ReLU activation and padding of size one. All intermediate convolutional layers except the first and last layer use Batch Normalisation.

Quality Assessment (QA) Network: Inspired by [36], a novel quality-based score obtained from QA Network is employed which serves as a loss function in training. The design of QA network is shown in Fig. 5, which is inspired by the VGG. The addition of the QA loss in the overall optimization enhances the image quality based on human perception as the QA network is trained to mimic how humans rank images based on their quality. Instead of using a single path to feed input to the network, two paths have been employed in this case. To proceed forward, both of these features are subtracted. Each VGG block has two convolutional layers, the second of which uses a stride of 2 to reduce the spatial dimensions. The network uses Global Average Pooling (GAP) layer instead of flattening layer to minimize the trainable parameters. At fully connected layers, a drop-out technique is used to overcome the issue of over-fitting. The KADID-10K [38] dataset, consisting of $10,050$ images, was used to train the QA network. The dataset has been divided in 70%-10%-20% ratio for train-validate-test purposes respectively during the training process.

3.1 Loss Functions

As depicted in Fig. 2, the generator 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 the following formula:

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

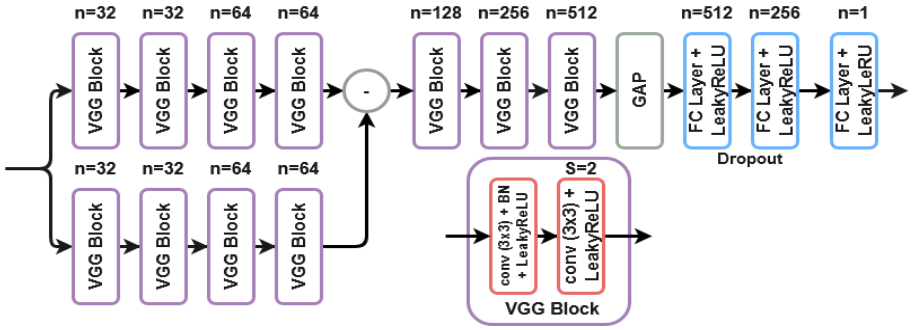


Fig. 5: The architecture of QA network [36].

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.

The Discriminator network is trained using triplet-based GAN loss. This can be expressed as,

$$L^{disc} = \lambda_3 L^D_{GAN} \tag{5}$$

where λ_3 is emperically set to 1×10^{-1}.

Both L^G_{GAN} and L^D_{GAN} are defined in Eqn. 9 and 10 respectively.

The content loss in Eqn. 4 has been used to preserve the content of the ground truth, which is an L_1 loss between ground truth HR (i.e., I_{HR}) and generated image SR (i.e., I_{SR}), and same can be expressed as,

$$L_{content} = \sum^{N} \|G(I_{LR}) - I_{HR})\|_1, \tag{6}$$

where N denotes the batch size in training, and G represents the function of generator. $\sum^N[\cdot]$ denotes an average operation across all images in the mini-batch. The perceptual loss $L_{perceptual}$ is used here for improving the perceptual similarity of the generated image with respect to its ground truth, which can be expressed as,

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

Here, $MSE(a, b)$ 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 [10]. 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. To further improve the quality of SR images based on human perception, a Quality Assessment (QA) loss is also introduced. It rates the SR image on a scale of 1-5, with a higher value indicating

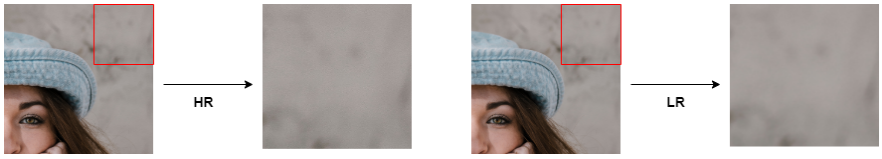


Fig. 6: Comparison of background patch in LR and HR images.

better quality. This predicted value is used to calculate the QA loss i.e., L_{QA}, which is expressed as [36],

$$L_{QA} = \sum^{N} \big(5 - Q(I_{SR})\big), \tag{8}$$

where $Q(I_{SR})$ represents the quality score of SR image from the QA network.

The GAN loss used here is a triplet-based loss function to a patch-based discriminator. An image can be simplified consisting 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, the background between LR and HR images is hard to differentiate as shown in Fig 6. A background patch with a vanilla GAN would be similar to the discriminator perceptually, hence forcing the output for the same to be real/fake could lead to a high erroneous loss and cause instability and noise in training. However, in the case of foreground patches, the idea of vanilla GAN 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. We formulate this as triplet loss optimization comprising of 3 variables - positive, negative and anchor. The distance between the anchor and the positive is minimised by the cost function, while the distance between the anchor and the negative is maximised. For the generator, the anchor is defined as the generated SR image (I_{SR}), the positive as the ground-truth HR image (I_{HR}), and the negative as the up-sampled LR input ($n(I_{LR})$), where n is the bicubic upsampling factor. The positive and negative are interchanged for training the discriminator. Thus, the triplet-based GAN losses for generator and discriminator can be defined as,

$$L^{G}_{GAN} = \sum^{N} \big[MSE(D(I_{SR}), D(I_{HR})) - MSE(D(I_{SR}), D(n(I_{LR}))) + 1\big] \tag{9}$$

$$L^{D}_{GAN} = \sum^{N} \big[MSE(D(I_{SR}), D(n(I_{LR}))) - MSE(D(I_{SR}), D(I_{HR})) + 1\big] \tag{10}$$

Here, $MSE(a, b)$ represents mean square error between a and b; n denotes up-sampling factor. 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 $D(n(I_{LR}))$, which are obtained from the Discriminator. Simultaneously, it also trains the patch-based Discriminator to distinguish the generated SR image from the ground-truth HR. The background patch as discussed before is similar for LR and HR images. Applying this triplet-based GAN loss patch-wise, improves

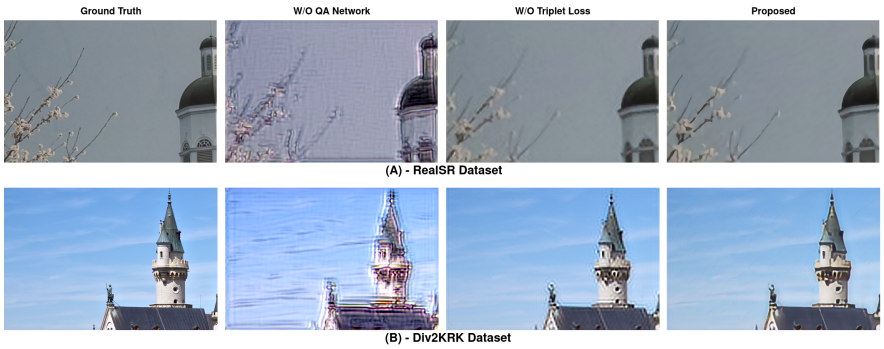


Fig. 7: Comparison of the results obtained through our proposed method-*SRTGAN* (with QA network and Triplet loss) Vs without incorporating QA Network or Triplet Loss on (A)-RealSR dataset [6] and (B)-DIV2KRK dataset [7]

the adversary as it allows the discriminator to better distinguish the main subject(foreground) of SR and HR images, which helps in generating images with better perceptual fidelity.

4 Experimental Results

4.1 Training Details

Using our proposed framework, we conduct supervised training on the RealSR dataset [6]. In this dataset, the focal length of a digital camera has been adjusted to collect LR-HR pairs of the same scene. To incrementally align the image pairs at various resolutions, an image registration method is developed. Our proposed network has been trained on 400 such images from the RealSR dataset and additionally it has been validated on 100 LR-HR image pairs provided in the same dataset. Finally, DIV2KRK [7] and test set of RealSR dataset [6] are employed for testing purposes. The LR images are subjected to several augmentations during the training phase, including horizontal flipping, rotation of $0°$ or $90°$, and cropping operations.The total trainable parameters of generator and discriminator networks are $3.7M$ and $2.7M$, respectively.

Additionally, we also employ QA network-based loss to enhance the quality of generated images. This method has been referenced from the work of [36]. Our proposed triplet loss optimization improves the visual appearance of the SR images to make them more realistic.

4.2 Ablation Study

We demonstrate the experimental support for incorporating the triplet loss and QA network in this section. Quantitative and Qualitative assessment conducted on the RealSR dataset [6], are shown in Table 1 and Fig. 7, respectively. Our method yields superior SR outcomes on both synthetic and real-world data (RealSR dataset). The proposed method with QA network and Triplet Loss performs better (see Table 1) when compared to the performance obtained using

Table 1: Quantitative evaluation of SRTGAN (with QA Network and Triplet Loss) Vs without incorporating these modules on the RealSR dataset [6].

Method	**PSNR** ↑	**SSIM** [17]↑	**LPIPS** [18]↓
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**

Table 2: Quantitative evaluation of SRTGAN with other state-of-the-art SR methods on RealSR and DIV2KRK dataset

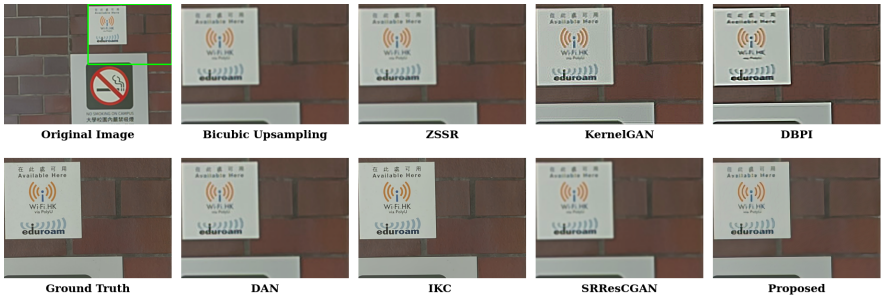
Method	**PSNR** ↑	**SSIM** [17] ↑	**LPIPS** [18] ↓	**PSNR** ↑	**SSIM** [17] ↑	**LPIPS** [18] ↓
	DIV2KRK [7] Dataset			*RealSR [6] Dataset*		
Bicubic	23.89	0.6478	0.5645	25.74	0.7413	0.4666
ZSSR [39]	24.05	0.6550	0.5257	25.83	0.7434	0.3503
KernelGAN [7]	24.76	0.6799	0.4980	24.09	0.7243	0.2981
DBPI [40]	24.92	0.7035	0.4039	22.36	0.6562	0.3106
DAN [41]	**26.07**	**0.7305**	0.4045	26.20	**0.7598**	0.4095
IKC [42]	25.41	0.7255	0.3977	25.60	0.7488	0.3188
SRResCGAN [15]	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**

the framework without those modules. This is quantitatively evaluated on various distortion metrics like PSNR and SSIM and perceptual measures, such as LPIPS. The SR images produced using our proposed approach with QA network and Triplet Loss are also perceptually better when compared to without adding these modules, which is shown in Fig. 7. It has been observed that our method without QA Network generates blurry output and variation in the natural color of the image. Our framework when optimized using vanilla GAN loss(instead of triplet loss), closely resembles the colour as anticipated in the real world, but fails to sharpen the edges, causing blurring. The proposed method's advantage may be observed in its ability to produce SR images with an adequate level of sharpening around the edges and preserving the color-coding of the original image. Here, by observing Fig. 7, one may quickly determine the perceptual improvement from our proposed strategy.

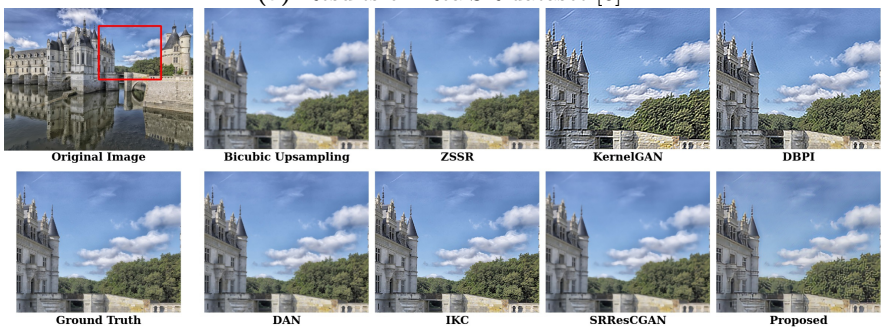
4.3 Quantitative Analysis

The PSNR and SSIM values, which are the accepted measurements for the SR problem, are often estimated for comparison of the results between different approaches. These metrics, however, do not entirely justify the quality based on human perception. Therefore, we also estimate a full-reference perceptual quality assessment score known as LPIPS [18]. A low LPIPS score indicates a better visual quality.

The comparison of all three metrics on the DIV2KRK [7] and RealSR datasets [6] is presented in Table 2. On both datasets, SRTGAN outperforms other novel approaches on LPIPS metric, demonstrating the proposed method's superiority in terms of perceptual quality. Our proposed approach also performs superior



(a) Results on RealSR dataset [6].



(b) Results on DIV2KRK dataset [7].

(c) Results on DIV2KRK dataset [7].

Fig. 8: Qualitative evaluation of SRTGAN with other state-of-the-art methods on RealSR and DIV2KRK dataset

to other methods on PSNR metric, whereas performs competitively in terms of SSIM, on the RealSR dataset [6]. SRTGAN also performs quite competitively in terms of PSNR and SSIM on the synthetic dataset - DIV2KRK [7]. The perceptual metric, LPIPS obtained using our proposed approach is significantly better for both datasets (see Table 2).

4.4 Qualitative Analysis

In this section, we show the efficacy of SRTGAN through visual inspection. We qualitatively evaluate the SR performance on one image of RealSR dataset (Fig. 8a) [6] and two sample images of DIV2KRK dataset (Fig. 8b and 8c) [7]. In addition, we also make comparison with other novel works such as KernelGAN

[7], ZSSR [39], DBPI [40], DAN [41], IKC [42], and SRResCGAN [15]. These SR results demonstrate that SRTGAN significantly reduces the amount of noise in the SR image and improves image clarity in comparison to other novel methods. In additon, SRTGAN can produce colours similar to the ground truth, while competing methods like IKC and KernelGAN over-boosts the colours in the generated images.

Our proposed method - SRTGAN produces SR images of better quality and with fewer noise artifacts than existing state-of-the-art methods. The quantitative assessment of several quality metrics (see Table 2) and the perceptual quality acquired on various datasets (see Fig. 7- 8c) support this conclusion.

5 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. 7, 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.

6 Conclusion

We have proposed an approach to the SISR problem 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 both high-resolution and low-resolution images and hence, it captures the essential information required in the SR image. Applying patch-wise triplet loss improves the adversary as it allows the discriminator to better distinguish the main subject(foreground) of SR and HR images, which helps in generating images with better perceptual fidelity. Through experiments, we have demonstrated that SRTGAN can super-resolve images by a factor of $\times 4$ with improved perceptual quality than other competing methods.

References

1. Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *IEEE CVPR*, vol. 1, no. 4, 2017.
2. C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE TPAMI*, vol. 38, no. 2, pp. 295–307, 2016.
3. 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.

4. C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *ECCV*, Oct 2016, pp. 391–407.
5. B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolutaion," *IEEE CVPR Workshops*, pp. 1132–1140, 2017.
6. 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.
7. S. Bell-Kligler, A. Shocher, and M. Irani, "Blind Super-Resolution Kernel Estimation using an Internal-Gan," in *NeurIPS*, 2019, pp. 284–293.
8. N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin, "Accurate blur models vs. image priors in single image super-resolution," in *ICCV*, 2013, pp. 2832–2839.
9. M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," Jan. 2016, 4th International Conference - ICLR 2016.
10. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, 2016, pp. 694–711.
11. A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *NeurIPS*, ser. NIPS'16, 2016, p. 658–666.
12. J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," *CoRR*, vol. abs/1511.05666, 2016.
13. C. Ledig, L. Theis, F. Huszár *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE CVPR*, 2017, pp. 4681–4690.
14. 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.
15. R. Muhammad Umer, G. Luca Foresti, and C. Micheloni, "Deep Generative Adversarial Residual Convolutional Networks for Real-World Super-Resolution," in *IEEE CVPR Workshops*, 2020, pp. 438–439.
16. S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "Srfeat: Single image super-resolution with feature discrimination," in *ECCV*, 2018, pp. 439–455.
17. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.
18. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.
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 *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 *CVPR Workshops*, June 2020.
23. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *IEEE CVPR*, 2018, pp. 2472–2481.
24. M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *IEEE 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*, 2014, pp. 2672–2680.
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 Process. Lett.*, 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 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 CVPR Workshops*, June 2019.
34. X. Xu and X. Li, "Scan: Spatial color attention networks for real single image super-resolution," in *The IEEE CVPR Workshops*, June 2019.
35. J. Kwak and D. Son, "Fractal residual network and solutions for real super-resolution," in *The IEEE 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 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 QoMEX*, 2019, pp. 1–3.
39. 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.
40. J. Kim, C. Jung, and C. Kim, "Dual Back-Projection-Based Internal Learning for Blind Super-Resolution," in *IEEE Signal Process Lett*, vol. 27, 2020, pp. 1190–1194.
41. 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.
42. J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *IEEE CVPR*, June 2019, pp. 1604–1613.