Received November 25, 2019, accepted December 11, 2019, date of publication December 16, 2019,
date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959940

Spatial Transformer Generative Adversarial Network for Robust Image Super-Resolution

HOSSAM M. KASEM[1,2,3], **KWOK-WAI HUNG**[1,2], **AND JIANMIN JIANG**[1,2]

[1]Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China
[3]Faculty of Engineering, Tanta University, Tanta 31512, Egypt

Corresponding author: Jianmin Jiang (jianmin.jiang@szu.edu.cn)

This work was supported in part by the Natural Science Foundation China (NSFC) under Grant 61620106008 and Grant 61602312, and in part by the Shenzhen Commission for Scientific Research and Innovations under Grant JCYJ20160226191842793.

ABSTRACT Recently, there have been significant advances in image super-resolution based on generative adversarial networks (GANs) to achieve breakthroughs in generating more images with high subjective quality. However, there are remaining challenges needs to be met, such as simultaneously recovering the finer texture details for large upscaling factors and mitigating the geometric transformation effects. In this paper, we propose a novel robust super-resolution GAN (i.e. namely RSR-GAN) which can simultaneously perform both the geometric transformation and recovering the finer texture details. Specifically, since the performance of the generator depends on the discreminator, we propose a novel discriminator design by incorporating the spatial transformer module with residual learning to improve the discrimination of fake and true images through removing the geometric noise, in order to enhance the super-resolution of geometric corrected images. Finally, to further improve the perceptual quality, we introduce an additional DCT loss term into the existing loss function. Extensive experiments, measured by both PSNR and SSIM measurements, show that our proposed method achieves a high level of robustness against a number of geometric transformations, including rotation, translation, a combination of rotation and scaling effects, and a cobmination of rotaion, transalation and scaling effects. Benchmarked by the existing state-of-the-arts SR methods, our proposed delivers superior performances on a wide range of datasets which are publicly available and widely adopted across research communities.

INDEX TERMS Super-resolution, generative adversarial networks, spatial transformer network, robust image super-resolution, robust generative adversarial network.

I. INTRODUCTION

Single image super-resolution (SISR) has attracted increasing attention in the research community and numerous image SR methods have been reported to deal with this non-trivial problem [1]. Single image super-resolution (SISR) is a process which recover high-resolution (HR) images from its lower resolution (LR) version with better visual quality and refined details. SISR enjoys a wide range of real-world applications, such as medical imaging [2]–[4], surveillance and security [5]–[7], amongst others.

In recent years, a variety of classical non deep learning super-resolution methods have been proposed, including prediction-based methods [8]–[10], edge-based methods [11], [12], statistical methods [13], [14], patch-based methods [15], [16] and sparse representation methods [13], [17], etc.

With the rapid development of deep learning techniques in recent years, deep learning based super-resolution models have been extensively explored and often achieve the state-of-the-art performance on various benchmarks of super-resolution. A variety of deep learning methods using convolutional neural networks (CNN) have been applied to tackle SR tasks [18]–[35]. These deep learning techniques can easily learn the correlation between the LR and HR images and then achieve better performance compared with conventional methods. The CNN based SR networks are utilized common mean square error (MSE) loss function to minimize the error between the LR and HR images. Due to the fact that MSE is not able to measure the complex signal structure including the texture regions, the network tend to output over-smoothed results without sufficient high-frequency details.

The associate editor coordinating the review of this manuscript and approving it for publication was Qiangqiang Yuan.

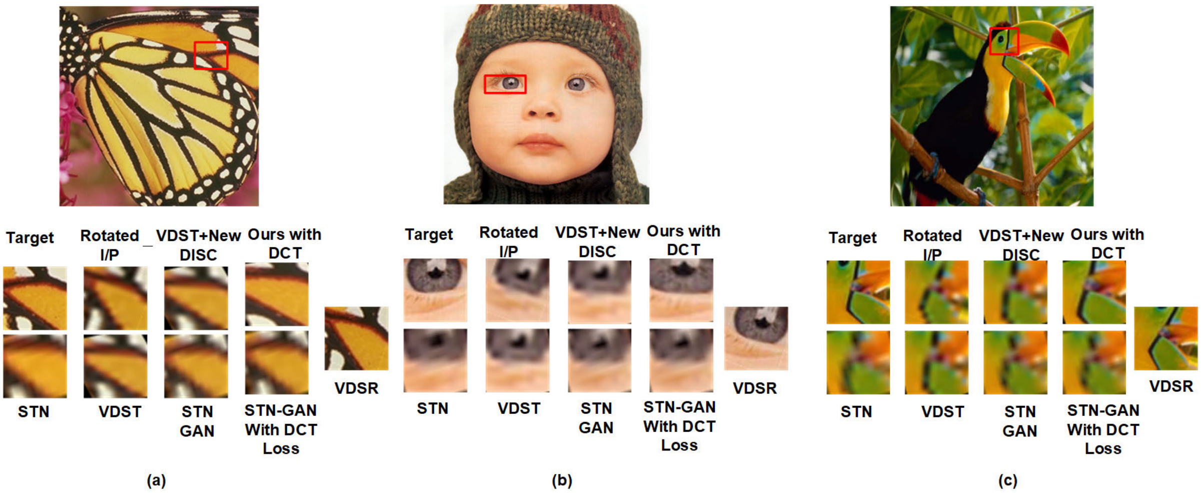


FIGURE 1. **Visual comparison samples of our proposed framework and various benchmarks: Top and third rows: original image. Under each row: The desired output, rotated input patch, our proposed framework output with generative loss function (PSNR/SSIM), our proposed framework output with our loss function output (PSNR/SSIM), existing STN, existing VDST, STN based GAN network with generative loss function, STN based GAN network with our proposed loss function output (PSNR/SSIM), and VDSR, respectively.**

Thus, the MSE loss is not always proportional to human perception quality, and it is not sufficient for recovering HR images with finer details.

To address the weakness of the existing single image super-resolution (SISR) approaches for recovering the finer details of HR images and producing more realistic images, several perceptual-driven methods have been proposed to improve the visual quality of SR results. The discovery of Generative Adversarial Networks (GAN) leads to better perceptual quality. (GAN) [36], [37] employ a game-theoretic approach where two components of the model, namely a generator and discriminator, where the generator try to fool the discriminator. Specifically, the generator creates SR images that a discriminator cannot distinguish as a real HR image or an artificially super-resolved output. In this manner, HR images with better perceptual quality are generated. For example, the authors in [25], [38] have proposed the perceptual loss to minimize the error between the target image and the generated image by the deep network in the feature space instead of pixel space. The authors in [39] proposed utilizing the generative adversarial network (GAN) [36] in SR to encourage the network to favor solutions that look more like natural images.

However, the GAN based network does not take into account the practical scenario. Specifically, practical applications of single image super-resolution indicate that the real LR measurements usually suffer from various types of corruptions, such as geometric transformations, noise, and blurring. Additionally, although the GAN is able to generate high frequency (HF) details, some artifacts and noise are generated with HR images.

In this paper, we propose a novel robust framework which is able to solve both challenges mentioned above. Specifically, we introduce a robust generative adversarial network which is able to mitigate the geometric transformation and recovering the high frequency details of the images simultaneously. To be more precise, our proposed framework contains two parts including the generator and discriminator. Structurally, we propose to use our spatially transformed module which is presented in [26] to overcome the geometric transformation distortion. Moreover, we propose a new combined loss function to produce more realistic images. Additionally, a novel discriminator design is presented to improve the generator ability to produce HR with more high frequency details. We have added a spatial transformer module with the existing discriminator design. The spatial transformer module is able to remove the geometric transformation and background noises. Thus, the ability of the discriminator to distinguish between the fake and real images increased. Consequently, the generator ability to produce more realistic images is improved. Fig.1, shows the robustness of our proposed framework against the transformation effect and produce more realistic images than the existing state-of-the-art methods. As shown, VDSR network is not able to mitigate the transformation effect but it is able to generate HR with non correct direction. Moreover, the results in Fig. 1, show that the performance of our proposed framework still better that the existing spatial transformer network [40].

In summary, our contributions can be highlighted as follows:

- We propose a robust super-resolution generative adversarial network which is designed to simultaneously perform geometric corrections and generate more realistic images;
- We propose a novel discriminator design which is able to improve the capability of discriminating between fake and true images;
- We propose to add a DCT loss term to the generator loss function to bridge the gap between MSE loss and

adversarial loss to improve the perceptual quality of the recovering images.

- In comparison with existing SR image methods, our proposed framework achieves much lower training cost and learning complexity due to the fact that our proposed deep model requires significantly less number of model parameters.;
- In comparison with the existing state of the art SR methods reported in the literature, experimental results support and verify that our proposed achieves superior performances in terms of both PSNR and SSIM.

II. RELATED WORK

Super-resolution methods can be broadly divided into two main categories: traditional and deep learning methods. Traditional algorithms have been around for decades now, but are out-performed by their deep learning based counterparts. Therefore, most recent algorithms rely on data driven deep learning models to reconstruct the required details for accurate super-resolution.

Super-Resolution Convolutional Neural Network abbreviated as SRCNN [1] is the first successful attempt towards using CNN for super-resolution. SRCNN be considered as the pioneering work in deep learning based SR that inspired several later attempts in this direction. Unlike the shallow network architectures used SRCNN, Very Deep Super- Resolution (VDSR) is based on a deep CNN architecture originally proposed in [19]. VDSR shows that deeper networks can provide better contextualization and learn generalizable representations that can be used for multi-scale super-resolution. The authors in [20]–[25], propose deep neural network structure which is relay on the skip connections, residual blocks, and Laplacian pyramids. Although, these networks show a better performance but it still suffer from the geometric transformation effects. Thus, its performance is degraded with these effects.

The authors of [40] introduce the spatial transformer network (STN) which is able to mitigate the transformation effects and increase the robustness of the network against the various transformation effects. Spatial transformer unit [40] aims to perform a geometric transformation on an input map so that CNNs are provided with the ability to be spatially invariant to the input data in a computationally efficient manner. This differentiable module can be inserted into existing CNN architectures since the parameters of the transformation that are applied to feature maps are learnt by means of a backpropagation algorithm. Spatial transformer networks consist of 3 elements: the localisation network, the grid generator and the sampler. Although, STN shows a wide success in various computer visions problems. Inspired by STN, we proposed a novel spatial transformer module In [26] and then we have called it Very Deep Spatial Transformer (VDST) module. In addition, we propose a novel robust SR framework which is able to mitigate the geometric transformation effects and generate HR image with better PSNR and SSIM values than existing STN. Our simulation results confirmed the superiority of VDST over the existing STN under different geometric transformation effects. However, all the aforementioned networks are utilized the MSE as loss function to minimize the error between HR and target image, they produce a blury images, since it has an average effect.

The generative adversarial network (GAN) [36], shows an improvement in the subjective quality of the recovered images. Consequently, various networks have been proposed utilizing GAN in SR field [39], [41]–[45]. The authors in [39], propose to use the a combined loss function which contains two parts including the adversarial and content losses. They proposes to use the MSE loss and VGG loss to improve the visual quality with adversarial loss. Consequently, the proposed network successfully improver th e subjective quality over the existing networks. However, there are still artifacts produces with the generated images. In [41], the authors propose to add texture matching loss to the generator loss function to produce more realistic textures and further reduce the occurrence of visually implausible artifacts. Although the success of [41] to produce more ralistic images with more HF details, the PSNR and SSIM have been reduced. Also, the authors in [42] propose a novel generator loss function with the EUSR network [46]. In the proposed generator, the authors propose to uses content loss and differential content loss, which both use L1-norm. Additionally, they have utilized DCT loss to measure the similarity on the frequency domain between the recovered and the target images. Although it reduces unpleasing HF and succeeds in achieving high objective quality in terms of PSNR and SSIM, it has a higher perceptual index (PI) value than SR-GAN (a high PI value indicates a low subjective quality).

The authors in [43], proposes a pre-trained feature domain discriminator using a generator network with short-and-long range skip connections. They justify their use of multi-discriminators by obtaining high PSNR and SSIM values along with clarity close to the ground-truth. However, the use of VGG based discriminator has a limitation in terms of signal accuracy. Since VGG has been proposed for image classication, pooling layers are included and the VGG features are low-resolution. Thus, VGG feature maps tend to be global rather than local, leading to the occurrence of new artifacts around the edge region. The authors of [44], proposed an novel network with relativistic discriminator. The authors utilized the relativistic discriminator to judge whether one image is more realistic than another, which guides the generator to synthesize more detailed textures. In addition, the authors use L_1 content loss instead of the MSE. Moreover, the authors propose to improve the perceptual loss for the generator loss function. Consequently, they achieve consistently better performance than conventional SR methods in terms of subjective and objective qualities. The author of [47], propose multi perspective discriminator fo image super resolution. The authors propose to use multiple discriminator to improve the subjective qality through reducing the effect of the artifacts. Specifically, they propose three different discriminators including DCT perspective discriminator, gradient perspective discriminator, and spatial perspective discriminator. These proposed multi-perspective discriminators can easily identify artifacts, and they can help the generator reproduce artifact-less SR images.

To this end, all aforementioned proposed GAN networks lead to improve the subjective quality, but there are still challenges need to be met. First, the aforementioned networks still suffer form the noise generated with the recovered HR images. Additionally, all the above study dose not take into account the robustness of the network against the various geometric transformation effects including rotation(R), scaling (S), translation(T), a combination of rotation and scaling (RS), and combination of rotation, translation and scaling. In this paper, we introduce a robust GAN for image super-resolution to be able to mitigate the geometric transformation effects and reduce the noise generated with recovered images. Specifically, we propose our VDST [26] module as a generator to overcome the transformation effects. Additionally, we propose a novel discriminator design through integration of our spatial transformer module with the existing discriminator in order to improve its ability to distinguish the real and fake image. Moreover, we introduce a new combined loss function to add more high frequency details to the recovered images. The proposed loss function contains from three parts including adversarial loss, content loss and the DCT loss.

III. OUR PROPOSED ROBUST SUPER RESOLUTION GAN NETWORK

Our main objective is to create a robust SR network to alleviate the geometric transformation effects and generating more natural HR images that similar to the ground truth images simultaneously. The key aspect of achieving this goal is to alleviate the effect of spatial transforms for corrupted LR images. In other words, our proposed framework is able to mitigate the geometric transformation effects and recovering high frequency details of HR images. The flow chart of our proposed framework is shown in Fig.2. As shown in Fig.2, our proposed framework contains from two parts including the generator and discriminator. The flow of our proposed framework is started by feeding the transformed input to the generator. Our proposed generator tries to alleviate the transformation effects by producing a correct HR image. Then, this output is connected to the discriminator with the ground truth. The discriminator tries to decide that the generator output is fake or real image. Additionally, the DCT loss and MSE loss are calculated using both generator output and ground truth images. Then, the total loss is obtained by combining the adversarial loss with DCT and MSE losses. After that, the generator parameter is updated utilizing the value of the total loss function. In this section, we illustrate the details of our proposed framework, including the proposed discriminator, the deep residual learning based spatial transform module, and the improved loss functions.

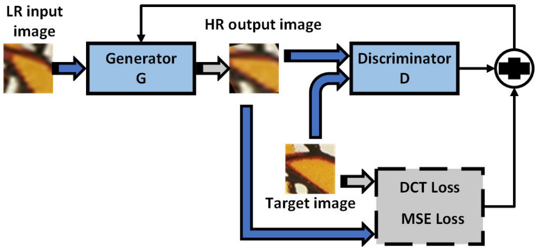
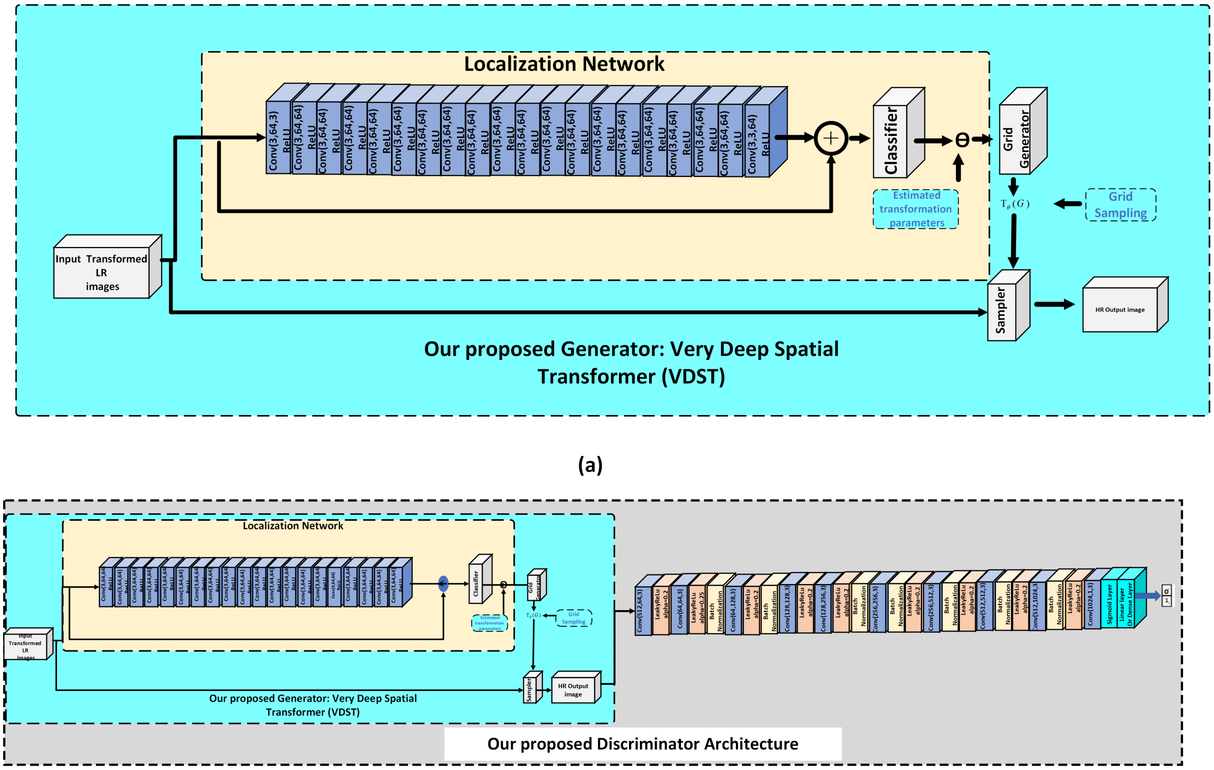


FIGURE 2. Illustration of our proposed framework flow chart.

A. OUR PROPOSED NETWORK ARCHITECTURE

As seen in Fig.3, our proposed network contains two parts: a) the generator and b) the discriminator. Structurally, Our proposed generator consists of VDST module [26]. We have shown in [26] that VDST is able to mitigate the geometric transformation effects. But, in our previous work [26], we have used VDSR module to refine the output of VDST in order to obtain HR image with better PSN and SSIM values. However, in this work, we propose to use only VDST module to overcome the transformation effects. Consequently, the number of the parameter needed to be estimated are reduced. Our VDST module of three parts namely localization network, grid generator, and sample. The first part is a localization network, which takes the input image through convolutional neural network (CNN) and estimate the warping parameters from the input image. We proposed a new localization design network namely deeper residual learning localization network (DRLN). DRLN is able to exploit wider contextual information inside the input images. In the DRLN, we propose to stack 20 convolutional layers to extract the features of the input LR image. the proposed DRLN is a deeper network with only 64 feature maps per layer, which is more powerful yet requires less parameters. In the second part of VDST, the predicted transformation or warping parameters are utilized to form a sampling grid, which is implemented by a grid generator. In the third part, the input image and the sampling grid are taken as inputs to the sampler, in order to interpolate the output image. Then the final correct HR image can be obtained. This HR image is passed to our proposed discriminator to decide if it is fake or real image compared with the ground truth.

Fig.3 shows our proposed discriminator. As seen, our proposed discriminator contains from two parts. First, we propose to use VDST again as a first stage. The principle behind adding of VDST as first stage is that spatial transformer networks learn to remove the geometric noise and background so that only the interesting zones of the input are forwarded to the next layers of the network [48]. Thus, the discriminator ability to distinguish tehe fake and real image increased. Consequently, the discriminator si drive the generator to produce more real mages similar to the ground truth images. The second part contains from eight convolution layer with kernel size 3 × 3. In addition, these convolution layer are followed by two dense layers and sigmoid layer.



(a)

(b)

FIGURE 3. Illustration of our proposed framework architecture: a) our proposed generator b) our proposed discriminator.

B. OUR PROPOSED LOSS FUNCTION

Recently, most of the SR networks utilize only the pixel-wise loss functions such as MSE as an objective function. However, minimizing MSE encourages finding pixel-wise averages of plausible solutions which are typically overly-smooth and thus have poor perceptual quality. As proposed in [36], the authors tackled this problem by employing (GANs) [36] used for the application of image generation To obtain a more realistic images. To achieve this task, they proposed to utilize the adversarial loss function to train the GAN network. Specifically, the generator is feed by LR image and then producing HR images. These images are passed to the discriminator network to distinguishes between the original image (i.e the target image) and the generated HR images. As the generator and discriminator try to compete each other, the generator improve its ability to produce more realistic images similar to the target images. However, this technique achieves better results, but the are some artifacts are generated inside HR images that produced from the generator. So, in this paper. we propose a combined loss function in order to produce more realistic images with lower artifcats. Specifically, we propose to combine the loss functions including the conventional adversarial loss function for GAN (l_{gen}) and content loss function (i.e MSE loss function (l_{MSE})). In addition, we propose to add one more content related loss term, (l_{dct}), based on the discrete cosine transform (DCT) to strengthen the reconstruction quality in high frequency details for all output images. To this end, we can describe the total loss function for the generator as:

$$l^{SR} = \alpha_1 \times l_{gen} + \alpha_2 \times l_{MSE} + \alpha_3 \times l_{dct} \tag{1}$$

where α_1,α_1, α_1 are the loss functions wights. The details of the loss functions are described below:

1) ADVERSARIAL LOSS

The general idea behind the adversarial loss is to train the generator model with the goal of fooling a differentiable discriminator that is trained to distinguish super-resolved images from the real images. This approach encourages the generative component to favour the solution which is perceptually more similar to the natural images by trying to fool the discriminator network.

The generator network is trained as a feed-forward CNN G_{θ_G} parametrized by θ_G. Hence, θ_G denotes the weights and biases of the generator layers. These weights is obtained by optimizing l^{SR} loss function. Then, θ_G can be described as:

$$\hat{\theta}_G = \mathbf{argmin}_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{sr} \left(G_{\theta_G} \left(I_n^{LR} \right), I_n^{HR} \right) \tag{2}$$

where I^{LR}, I^{HR} are LR and HR images, respectively. N is the number of the training images. In this work, we propose to use the loss function as in Eq.1, which is as a weighted combination of several loss components.

The discriminator D_{θ_D} and parameters θ_D are similar to the G_{θ_G} and θ_G for the generator, respectively.

$$\hat{\theta}_D = \mathbf{argmin}_{\theta_D} \frac{1}{N} \sum_{n=1}^{N} l^{sr} \left(G_{\theta_G} \left(I_n^{LR} \right), D_{\theta_D}(I_n^{HR}) \right) \quad (3)$$

Then, the adversarial min-max problem of GAN can be formulated as:

$$\min_{\theta_G} \max_{\theta_D} \sum_{n=1}^{N} logD_{\theta_D} \left(I^{HR} \right) + log(1 - D_{\theta_D} \left(G_{\theta_G} \left(I^{LR} \right) \right)) \quad (4)$$

where D_{θ_D} is the probability generated by the discriminator to distinguish the likelihood of reconstructed image $G_{\theta_G}\left(I^{LR}\right)$ generated by the generator to be a natural HR image.

As mentioned above, by adding the generative loss function to our proposed loss function as in Eq.1 encourages the G to to favour solutions that reside on the manifold of natural images, by trying to fool the discriminator network. Then, the generative loss l_{gen} is defined based on the the probabilities of the discriminator $D_{\theta_D}\left(G_{\theta_G}\left(I^{LR}\right)\right)$ and cab be represented as:

$$l_{gen} = \sum_{n=1}^{N} -logD_{\theta_D} \left(G_{\theta_G} \left(I^{LR} \right) \right) \quad (5)$$

In our proposed framework, we improve the ability of the discriminator to distinguish between the real and generated HR images by incorporating the spatial transformer into the discriminator (as shown in Fig.3 (b)). As a results, the discriminator is able to remove the geomatic transformation and background noise. Thus, the discriminator pushes the generator to reduce the error between the generated image and the real image, and hence producing images more similar to the target image.

2) CONTENT LOSS

In this section, we propose to use the pixel-wise MSE loss as a content loss function. As mentioned in Section. II, utilizing the adversarial loss leads to reduce the PSNR and SSIM values of the generated HR images. So, we propose to use the MSE loss function to increase these values. We propose to control in the contribution amount of the content loss by introducing α_2 weight. This content loss function tries to minimizes the error between the super-resolved images and the real images. For our proposed generator, the MSE Loss can be described as:

$$l_{MSE} = \left\| \hat{I}^{HR} - I^{HR} \right\|_2 \quad (6)$$

where $\hat{I}^{HR}$ is the generated image by our proposed generator.

As we mentioned, our proposed generator not only tries to alleviate the geometric transformation, but also tries to produce HR images more similar to the target images. Then, we can rewrite the content loss function as:

$$l_{MSE} = argmin_{\theta_A} \left\| \mathbf{I}^{LR}(\theta_A) - \hat{\mathbf{I}}_T \right\|_2^2 + argmin_{\theta_G} \left\| N(\hat{\mathbf{I}}_T, \theta_G) - \mathbf{I}_{HR} \right\|_2^2 \quad (7)$$

where $\hat{\mathbf{I}}_T$ is the output image after performing the spatial transformation, and $\hat{\mathbf{I}}_{HR} = N(\hat{\mathbf{X}}_T, \theta_G)$ is the estimated HR image, θ_G is the model parameters of the generator network. θ_A represents the estimated geometric/affine transformation parameters. As seen from Eq. 7, we can conclude that the content loss function can be divide into two parts. The first part of this equation is represented the first task of our proposed generator which is mitigation the geometric transformation effect by estimation θ_A. Then, we interpolate the transformed images by utilizing these parameters. The second part of the equation is used to minimize the errors between the estimated output after performing the spatial transformation and the target images.

3) DCT LOSS

As mentioned in Section. II that the various GAN networks suffer from some artifacts with the recovered image which leads to reduce the PSNR and SSIM values [47]. So, we propose to add a new DCT loss term to the total loss function to train the overall network as in Eq.1. For further performance improvement, the added DCT loss term enables our proposed generator to explicitly compare the two images in the frequency domain. More specifically, while different SR images can have the similar value of l_{MSE} and l_{gen}, the DCT loss term encourages the model to generate the final image having the frequency distribution as similar to the HR image as possible. In this way, the frequency distribution constraint compensates the gap between l_{MSE} and l_{gen} to generate more statistically consistent results in frequency domain. The forward DCT transform can be represented as:

$$DCT_{u,v}(I) = \frac{1}{2N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I(x, y) \quad (8)$$

$$cos[\frac{\pi}{N} u(x + \frac{1}{2})]cos[\frac{\pi}{N} v(y + \frac{1}{2})] \quad (9)$$

where each frequency band u, v corresponds to different DCT frequency components of texture units to be optimized to match the grouth truth. Hence, the $L2$-norm is applied for the DCT loss function:

$$l_{dct} = \left\| DCT(I^{HR}) - DCT(I^{SR}) \right\|_2 \quad (10)$$

IV. EVALUATIONS AND EXPERIMENTAL RESULT ANALYSIS

To test the robustness of our proposed framework, we carry out various experiments and illustrate our experimental results as well as their analyses. To make it convenient for benchmarking and comparative studies, we follow the experimental procedures as described in [26], [39]. First, to test

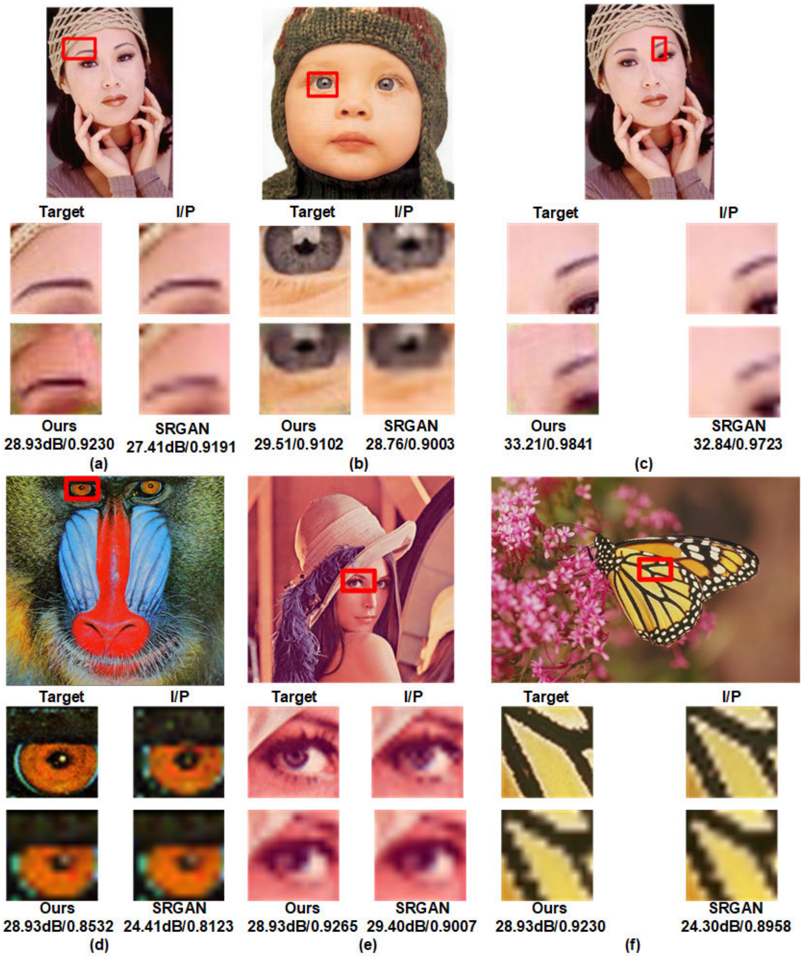


FIGURE 4. **Visual comparison samples of our proposed discriminator and existing SRGAN [39]. Top and fourth rows: original images. The target, input, and the output patches of our proposed and the existing SRGAN are given underneath each original image, and the bottom row presents the corresponding outputs measured in PSNR/SSIM values.**

the performance of our proposed discriminator, we repeat the experiment in [39] for scaling ×4. Specifically, we have used the same generator as in [39], but we replace the existing discriminator by our proposed as in Fig. 3.(b). To ensure a fair comparison, we have used the same loss function used in [39] to complete this experiment and then, we compare the obtained results with our proposed framework results.

Secondly, we contacted various experiments to show the robustness our proposed against the geometric transformation effects. We compare our the performance of our proposed network with various benchmarks including existing STN [40], VDST [26] and VDSR [19]. Finally, we compare our proposed with the different benchmarks over the number of modelling parameters to illustrate that the proposed network overwhelms the existing networks in terms of computing cost and learning complexity.

A. EXPERIMENT DESIGN AND SETUP

1) DATASET FOR TRANING AND TESTING

For the training purpose, we utilize 91 images from the Yang *et al.* [49] and 200 images from the training set of Berkeley Segmentation dataset [50] as our training data. To increase the size of the training dataset used, we augment the training data in three ways including the rotation, scaling and mirror effects. To generate the LR training images, we use the bicubic down-sampling and forming the images with size 48 × 48. In our experiments, we downscale all the training images using scale ×4.

For the testing purpose, we carry out experiments using 5 publicly available datasets, including: BSDS100 [50], SET5 [51], SET14 [52], URBAN100 [53] and MANGA109 [54].

2) IMPLEMENTATION DETAILS

To test our proposed framework with different geometric transformation effects, we simulate the effect of the geometric transformations, we generate the transformed LR training image by four different transformations, including: (i) the rotation effect represented by R, in which the original image is rotated clockwise by 20 degrees; (ii) the effect of both rotation and scaling represented by RS; (iii) translation represented by T, in which the LR images are translated by 5 pixels in both X and Y directions; and finally (iv) combinational

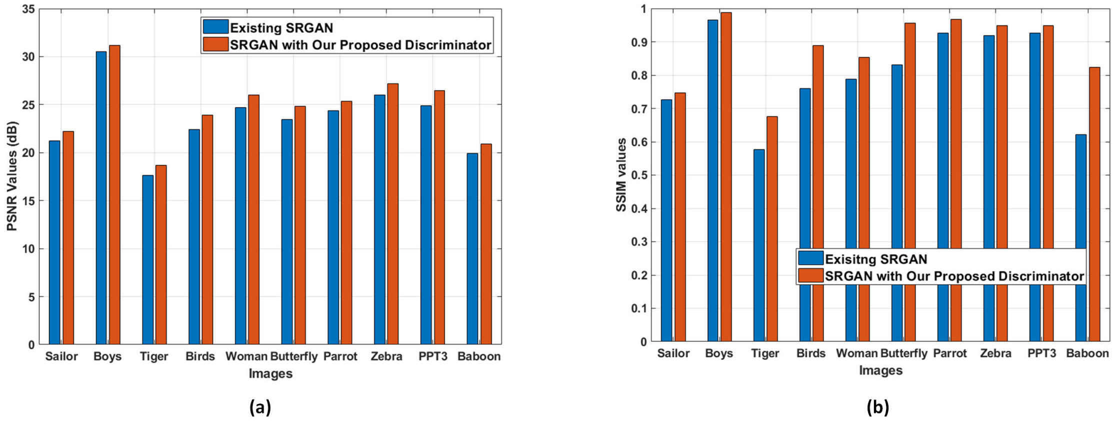


FIGURE 5. PSNR and SSIM values of our proposed and the existing SRGAN: The PSNR values are shown in (a), and SSIM values are shown in (b).

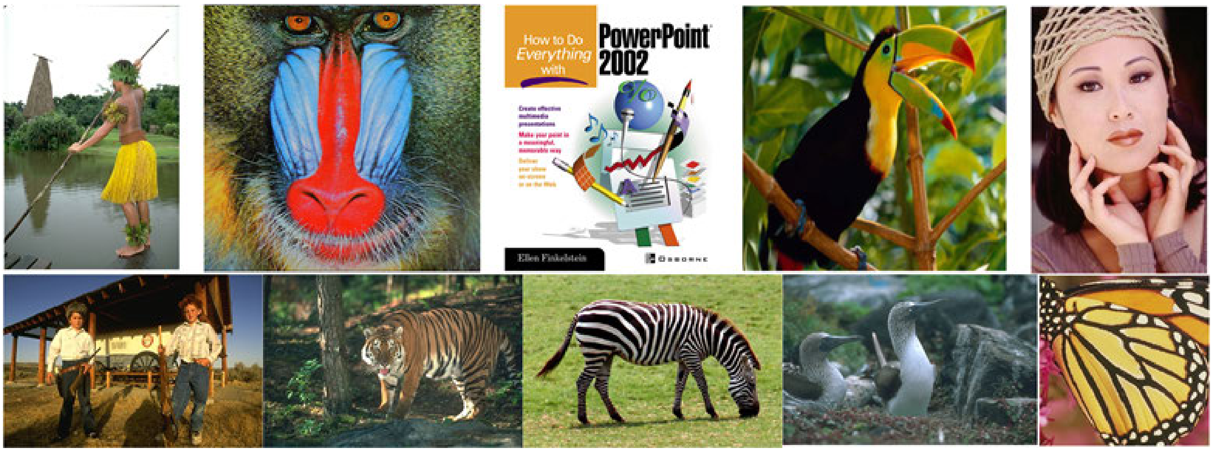


FIGURE 6. Illustration of ten natural images used to validate our proposed framework.

effect of rotation, scaling and translation represented by RTS. All the previous effects are performed on the downsampled images under scaling ×4.

For training our network, we utilize SGD optimization algorithm with learning rate 0.001 and no learning rate decay. The weight decay is set to 0.0001 and momentum is 0. Further, our proposed framework is trained in an end-to-end manner. We train all experiments over 10 epochs with batch size 25 and all the training is performed on NVIDIA Tesla P100. The evaluation results of all experiments are presented in term of two metrics widely used in the SR research community, which are Peak-signal-to-noise-ratio (PSNR) and Structural Similarity Index (SSIM).

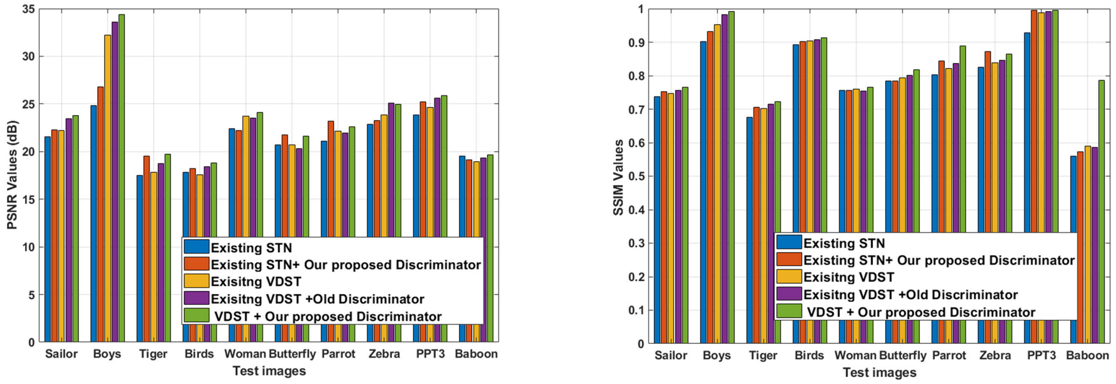
3) BENCHMARKS COMPARISONS

To validate the performance of the proposed framework, we compare the performance of our proposed network with different benchmarks. Firstly, we compare our proposed discriminator design with SRGAN [39], by replacing the discriminator used with SRGAN by our proposed. In this comparison, we use the same loss function presented in [39]. To fair comparison, we repeat the experiment by using our generating training and testing datasets as in section IV-A.2.

To test the performance of our proposed network against the geometric transformation effect, we compare our proposed with three different benchmarks, including the existing STN [40], VDST [26] and VDSR [19]. Moreover, we have formed one more benchmark by integrating the existing VDST and existing discriminator. Structurally, we construct a new GAN network by utilizing the VDST as a generator and the discriminator which is presented in [39], and called it as ‘‘VDST with old discriminator’’. One more Benchmarks, we construct another GAN network using the existing STN (i.e as a generator) and our proposed discriminator. These benchmarks can be divided into two groups: a)Pixel-wise benchmarks; b) adversarial benchmarks. First, the pixel-wise benchmarks are utilized only MSE loss to train the network. Secondly, the adversarial benchmarks, are utilized the conventional GAN loss function and our proposed loss function as loss function.

B. COMPARISON WITH THE STATE-OF-THE-ART

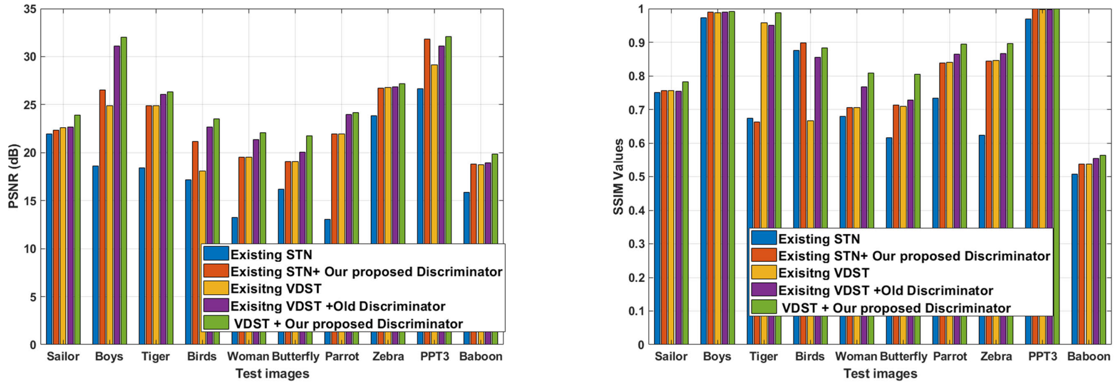
To validate the performance of our proposed discriminator design, we compare our proposed discriminator with one of the state-of-the-art network (i.e SRGAN [39]. Practically, we keep the generator design as described in [39].



(a)

(b)

FIGURE 7. **PSNR and SSIM values of our proposed framework and various benchmarks trained with generative loss function: The original images are rotated by 20 degrees in clockwise direction. The PSNR values are shown in (a), and SSIM values are shown in (b).**



(a)

(b)

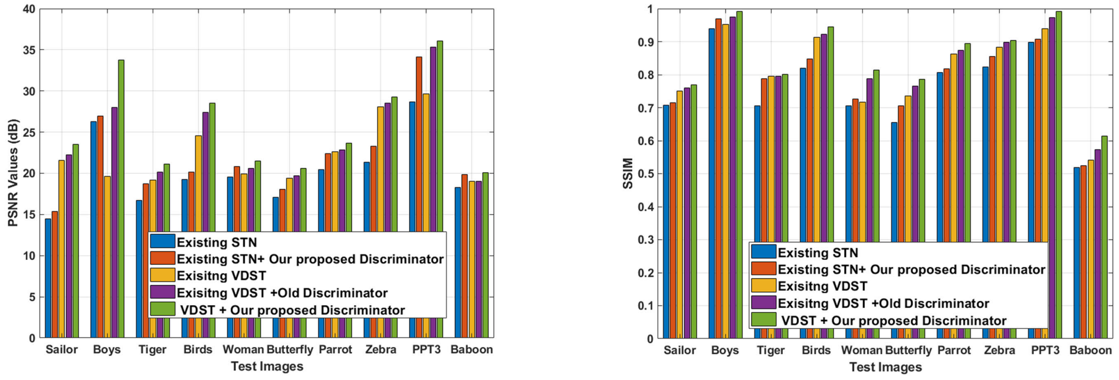
FIGURE 8. **PSNR and SSIM values of our proposed framework and various benchmarks trained with generative loss function: The original images are transformed by a combination of rotation and scaling effects. The PSNR values are shown in (a), and SSIM values are shown in (b).**

However, we replace the discriminator used with SRGAN by our proposed design. Then, we compare the the performance of our proposed design with the existing discriminator design. To fair comparison, we generate the LR traning dataset as described sectionIV-A.2. Specifically, we generate the LR images using bicubic downsample with scale ×4 and the image image size is 48 × 48. Then, we train the existing SRGAN with our generated dataset. The reason behind changing the size of the LR images than the original SRGAN is that the size limitation of our GPU memory. So, we reduce the LR images size to be 48 × 48.

To evaluate the performance of our proposed, we select various images from the test datasets (i.e. Set5 [51], Set14 [52]). For visual inspections and comparisons between our proposed and the existing SRGAN [39], Fig.4 shows six samples of the test images under scaling factor ×4. It can be seen that the GAN based SR network with our proposed discriminator design can generate images with more sharp details than SRGAN. For example, as seen in Fig.4, GAN based SR network with our proposed discriminator design is able to generate realistic images as in woman, boys and lena images. In addition, it is able to recover more details as shown in Baboon image. Further, the PSNR and SSIM values of our proposed discriminator is also much better than the exist SRGAN by 1dB. The reason behind this improvement is that our proposed discriminator able to remove the geometric noise and background so that only the interesting zones of the input are forwarded to the next layers of the network. Thus, the discriminator forces the generator to produce images which are more similar to the target images. Consequently, the error between the generated HR and its target images can be reduced. In other words, a discriminator network is essentially trained to discriminate those real HR images from the generated SR samples.

Fig.5 shows the objective evaluation results in terms of PSNR/SSIM under downsampled factor ×4. As seen again, our proposed achieves higher PSNR/SSIM values over the existing SRGAN.

To further evaluate our proposed, we conduct another experiment to test the performance of the proposed



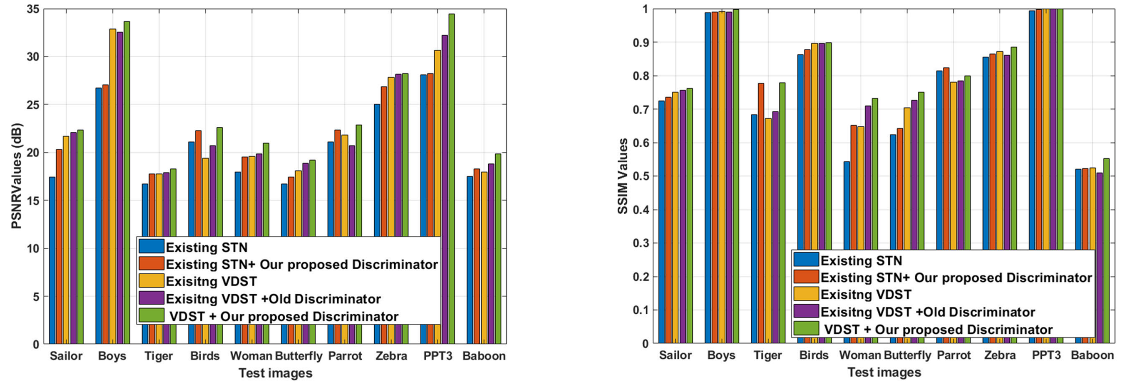
(a)

(b)

FIGURE 9. **PSNR and SSIM values of our proposed framework and various benchmarks trained with generative loss function: The original images are translated in X, and Y direction by 5 pixels. The PSNR values are shown in (a), and SSIM values are shown in (b).**

TABLE 1. **Experimental results (PSNR/SSIM) achieved by the existing SRGAN compared with our proposed framework with the scaling factor ×4.**

Test dataset / Network	SRGAN	SRGAN (Our Impelentation)	SRGAN with Our proposed Discriminator
BSD	27.58/0.7620	26.80/0.7598	28.01/0.7783
Manga	-	23.82/0.8028	24.02/0.8127
Set5	29.40/0.8472	29.96/0.8907	30.97/0.9211
Set14	28.49/0.8184	28.82/0.8581	29.17/0.8840
Urban	-	23.56/0.7366	23.91/0.7467



(a)

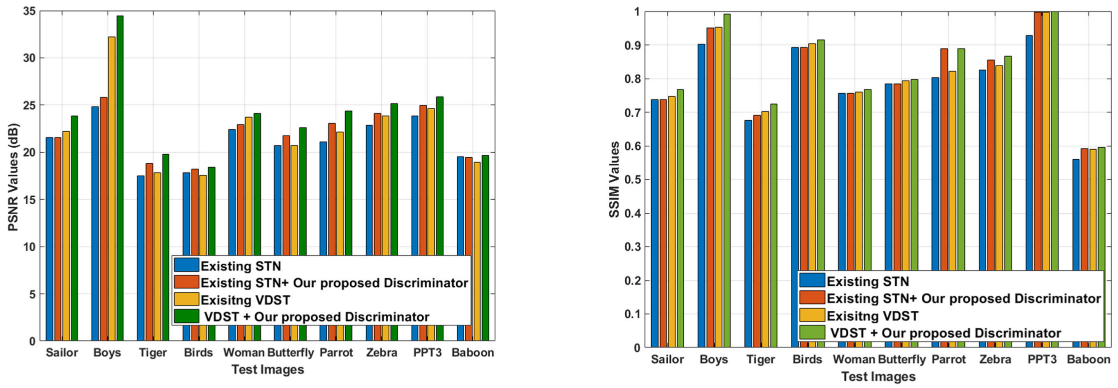
(b)

FIGURE 10. **PSNR and SSIM values of our proposed framework and various benchmarks trained with generative loss function: The original images are transformed by a combination of rotation, translation and scaling effects. The PSNR values are shown in (a), and SSIM values are shown in (b).**

discriminator using five testing datasets. All the results are illustrated in Table 1. From the results shown in Table 1, we can see that our proposed outperforms the existing SRGAN.

C. EXPERIMENTS ON EFFECTIVENESS AND ROBUSTNESS OF OUR PROPOSED DISCRIMINATOR

To validate the effectiveness and the accuracy of the proposed framework using our proposed discriminator design, we have carried out a range of experiments upon natural images to estimate the affine transformation parameters and mitigate the geometric transformation effects. Firstly, we compare our proposed with five benchmarks, including the existing STN [40], the existing VDSR [19], the existing VDST [26], STN based GAN network with our proposed discriminator design, and VDST based GAN network with existing discriminator design over the number of modelling parameters. Secondly, we validate the effectiveness of the proposed framework by experiments on recovery of transformed images, illustrating the advantage that the proposed framework can be turned into a robust end-to-end deep network against the effect of not only geometric transformations



(a)

(b)

FIGURE 11. PSNR and SSIM values of our proposed framework and various benchmarks trained with our proposed loss function: The original images are rotated by 20 degrees in clockwise direction. The PSNR values are shown in (a), and SSIM values are shown in (b).

TABLE 2. Structural comparisons between the proposed framework and the existing benchmarks.

Layers	STN +VDSR [26]	No. of Para.	VDST+VDSR [26]	No. of Para.	Our proposed framework	No. of Para.
Preprossing	Max Pooling (2,2)	/	Max Pooling (2,2)	/	Max Pooling (2,2)	/
Feature Extraction	1x Cov(5,200,3) ReLU MaxPooling(2,2)	15200	1x Cov(3,64,3) ReLU	1792	1x Cov(3,64,3) ReLU	1792
	1x Cov(5,300,200) ReLU MaxPooling(2,2)	15180300	18 x Cov(3,64,64) ReLU	664407	18 x Cov(3,64,64) ReLU	664407
			1x Cov(3,64,3) ReLU	1731	1x Cov(3,64,3) ReLU	1731
Fully Connected layer	Linear (2700,200)	540000	Linear (432,30)	12960	Linear (432,30)	12960
Classifier	Linear (200,6)	1200	Linear (30,6)	1800	Linear (30,6)	1800
Super Resolution Module	1x Cov(3,64,3) ReLU	1792	1x Cov(3,64,3) ReLU	1792	/	/
	18x Cov(64,64,3) ReLU	664407	18x Cov(64,64,3) ReLU	664407	/	/
	1x Cov(3,64,3) ReLU	1731	1x Cov(3,64,3) ReLU	1731	/	/
Total no. of parameters		16944630		1350917		682987

but also recovering high details information due to corruptions. As a result, our proposed is a powerful learning tool, which is able to simultaneously handle geometric transformations and recovery of high details information for corrupted images.

D. COMPUTATIONAL COMPLEXITY EVALUATION

We utilize the number of parameters as the evaluating criteria to compare our proposed framework with various benchmarks, the comparative results between our proposed framework and the existing STN + VDSR, and existing VDST + VDSR are shown in Table 2. The VDST + VDSR model as described in [26] contains two parts, which includes the deep spatial transformer (VDST) to alleviate the effect of spatial transforms for corrupted LR images, and the super resolution module (i.e. VDSR network) to refine the output of VDST to generate the HR image similar to the target image. In addition, the authors in [26], compare their work with the existing STN + VDSR. However, in our proposed framework, we train VDST without the super-resolution module not only for mitigating the transformation effects but also for generating HR images. In addition, inspired by the existing work [55], which is a typical application using the STN, we follow their design to produce 200 feature maps as the first convolutional layer and 300 feature maps as the second layer, as shown in Table 2.

In Table 2, the convolution layer is represented by Conv(k_i, n_i, c_i), where the variables k_i, n_i, c_i represent the filter size, the number of filters and the number of feature maps, respectively, and the linear layer is represented by Linear (m_i, o_i), where the variables m_i, o_i represent the size of the input vector and the size of the output vector, respectively.

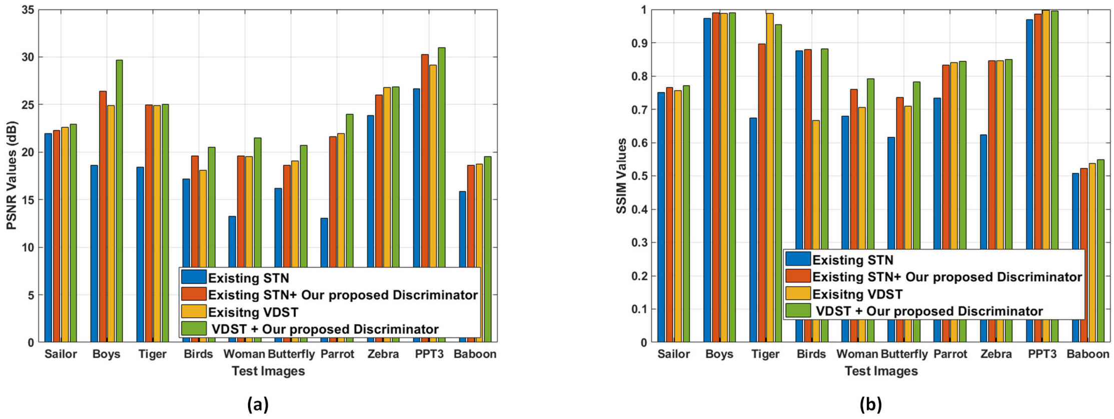


FIGURE 12. PSNR and SSIM values of our proposed framework and various benchmarks trained with proposed loss function: The original images are transformed by a combination of rotation and scaling effects. The PSNR values are shown in (a), and SSIM values are shown in (b).

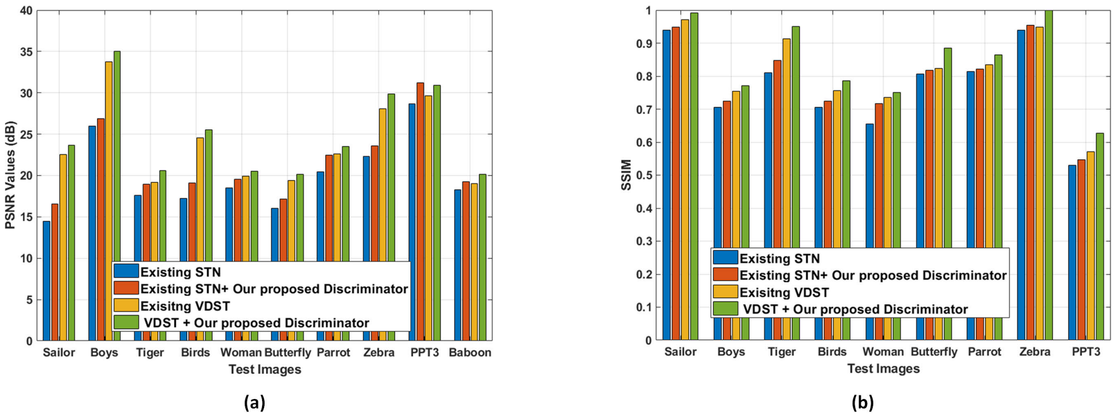


FIGURE 13. PSNR and SSIM values of our proposed framework and various benchmarks trained with proposed loss function: The original images are translated in X, and Y direction by 5 pixels. The PSNR values are shown in (a), and SSIM values are shown in (b).

As seen, the results given in Table 2 indicate that our proposed is powerful in structure, cost-effective in learning, and compared wit the existing benchmarks, leading to the improved performances.

E. EXPERIMENTS ON EFFECTIVENESS AND ROBUSTNESS OF OUR PROPOSED FRAMEWORK

To test the effectiveness and the accuracy of our proposed framework, we have carried out a range of experiments upon natural images to estimate the affine transformation parameters and mitigate the geometric transformation effects. Fig.6 illustrates seven samples of such natural images adopted as the test images in our experiments.

To evaluate the performance of our proposed, we compare our proposed framework with different benchmarks which are mentioned in section IV-A.3. In addition, we train our proposed framework using generative loss function and our proposed loss function. Moreover, we transformed the original images using the various transformation effects as mentioned in section IV-A.2. Then, we explore the robustness of our proposed framework against geometric transformations compared with the various benchmarks.

Fig. 7, 8, 9 and 10 show the objective evaluation results in terms of PSNR/SSIM when the original images are transformed using various transformation effects with generative loss function. Specifically, In Fig. 7, we rotated the original images by 20 degrees and then we trained the proposed framework and GAN based benchmarks with the existing generative loss function. In addition, we compare our proposed framework with pixel-wise benchmarks. From Fig.7, we can notice that our framework achieves higher PSNR/SSIM values over the existing benchmarks. To evaluate the perforamnce of our proposed framework with different transformation effects, we have transformed the original images including: 1) the combinational effect of rotation, and scaling effects 2) the translation effect, and the experimental results are illustrated in Fig. 8 and 9, respectively. As seen, our proposed framework still achieves superior performance over the various benchmarks.

To test the robustness of our proposed against stronger transformation effects, we apply the combinational effect of rotation, translation and scaling, and the corresponding experimental results are illustrated in Fig. 10. As seen, the simulation results indicate that our proposed spatial transformer

TABLE 3. Experimental results (PSNR/SSIM) achieved by the existing various benchmarks compared with our proposed framework with the scaling factor ×4 and generative loss function.

Transformation type	Test Dataset	VDSR	Existing STN	Existing STN With our proposed Discriminator	Existing VDST	Existing VDST With old Discriminator	VDST With our proposed Discriminator
Rotation (R)	BSD	21.69/0.7148	21.96/0.7635	22.68/0.7783	22.41/0.8051	23.02/0.7952	**24.66/0.8173**
	Manga	20.59/0.7444	23.38/0.8137	24.59/0.8348	22.31/0.8609	23.23/0.8365	**24.72/0.8612**
	Set5	22.29/0.7596	23.39/0.8247	24.47/0.8824	24.53/0.8832	26.10/0.8828	**26.32/0.8866**
	Set14	21.68/0.7580	22.34/0.8432	24.46/0.8551	22.48/0.8450	24.44/0.8327	**24.54/0.8476**
	Urban	20.56/0.7163	21.23/0.7532	23.22/0.7751	21.89/0.8033	23.10/0.7988	**23.66/0.8057**
Rotation and scaling (RS)	BSD	21.49/0.7070	21.80/0.7722	22.68/0.7783	21.93/0.7783	21.20/0.7720	**22.92/0.7914**
	Manga	20.45/0.7680	20.56/0.7786	21.38/0.8137	21.65/0.8195	21.40/0.8123	**22.75/0.8343**
	Set5	21.09/0.7623	21.59/0.7985	22.47/0.8247	22.60/0.8245	21.36/0.8323	**23.21/0.8519**
	Set14	21.56/0.7612	21.70/0.8017	22.26/0.8125	22.12/0.8121	21.56/0.7921	**22.59/0.8460**
	Urban	20.34/0.7112	20.55/0.7563	21.10/0.7611	20.99/0.7608	22.16/0.7541	**22.99/0.7989**
Translation (T)	BSD	16.99/0.6501	19.32/0.7463	20.23/0.7670	23.07/0.7960	23.50/0.8039	**24.92/0.7962**
	Manga	14.95/0.6670	18.66/0.7879	19.66/0.7998	22.87/0.8506	22.90/0.8686	**24.51/0.850**
	Set5	17.26/0.6930	18.64/0.7905	19.20/0.8069	23.07/0.8181	23.78/0.8209	**24.87/0.803**
	Set14	16.27/0.6691	19.41/0.7735	20.39/0.7943	23.73/0.8358	24.16/0.8421	**25.52/0.8355**
	Urban	16.51/0.6691	19.04/0.7239	18.93/0.7246	22.44/0.7900	22.79/0.7898	**24.34/0.7909**
RTS	BSD	18.00/0.6855	18.30/0.7487	19.72/0.7606	20.48/0.7652	21.68/0.7783	**22.02/0.7976**
	Manga	16.32/0.7023	18.26/0.7906	19.18/0.8073	20.46/0.8089	20.99/0.8270	**22.67/0.8293**
	Set5	18.35/0.7256	18.40/0.7781	19.66/0.7922	21.60/0.8060	21.43/0.8149	**22.39/0.8269**
	Set14	17.12/0.7161	18.13/0.7562	19.07/0.7795	21.41/0.8031	21.76/0.8115	**22.80/0.8313**
	Urban	17.45/0.6703	18.71/0.7263	19.53/0.7494	20.60/0.7485	21..20/0.7676	**22.82/0.7704**

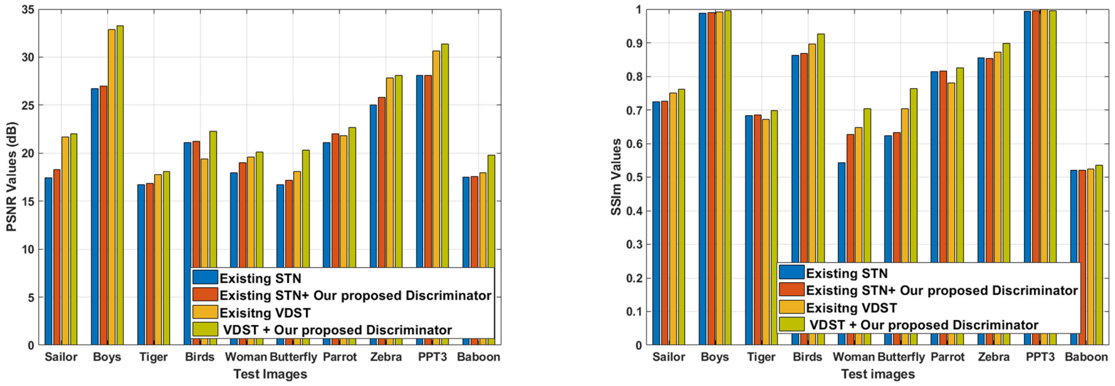
TABLE 4. Experimental results (PSNR/SSIM) achieved by the existing various benchmarks compared with our proposed framework with the scaling factor ×4 and our proposed loss function.

Transformation type	Test Dataset	VDSR	Existing STN	Existing STN With Our Proposed Discriminator	Existing VDST	VDST With Our Proposed Discriminator
Rotation (R)	BSD	21.69/0.7148	21.96/0.7635	22.55/0.7735	22.41/0.8051	**23.65/0.8153**
	Manga	20.59/0.7444	23.38/0.8137	23.44/0.8145	22.31/0.8609	**23.70/0.8560**
	Set5	22.29/0.7596	23.39/0.8824	25.95/0.8834	24.53/0.8832	**26.44/0.8860**
	Set14	21.68/0.7580	22.34/0.8432	24.28/0.8421	22.48/0.8450	**24.54/0.8475**
	Urban	20.56/0.7163	21.23/0.7532	22.92/0.7995	21.89/0.8033	**23.68/0.8046**
Rotation and scaling (RS)	BSD	21.49/0.707	21.80/0.7722	22.38/0.7855	21.93/0.7783	**22.80/0.7911**
	Manga	20.45/0.7680	20.56/0.7786	22.05/0.7956	21.65/0.8195	**22.56/0.8341**
	Set5	21.09/0.7623	21.59/0.7985	22.10/0.8035	22.60/0.8245	**23.16/0.8531**
	Set14	21.56/0.7612	21.70/0.8017	22.26/0.8176	22.12/0.8121	**23.56/0.8949**
	Urban	20.34/0.7112	20.55/0.7563	21.33/0.7856	20.99/0.7608	**22.80/0.8474**
Translation (T)	BSD	16.99/0.6501	19.32/0.7463	20.32/0.7565	23.07/0.7960	**24.87/0.8571**
	Manga	14.95/0.6670	18.66/0.7879	19.62/0.7974	22.87/0.8506	**23.46/0.8871**
	Set5	17.26/0.6930	18.64/0.7905	19.54/0.8006	23.07/0.8181	**24.78/0.8933**
	Set14	16.27/0.6691	19.41/0.7735	20.48/0.7937	23.73/0.8358	**24.95/0.8657**
	Urban	16.51/0.6691	19.04/0.7239	20.09/0.7445	22.44/0.7900	**24.77/0.8271**
RTS	BSD	18.00/0.6855	18.30/0.7487	18.87/0.7490	20.48/0.7652	**21.01/0.7338**
	Manga	16.32/0.7023	18.26/0.7906	18.46/0.8102	20.46/0.8089	**21.78/0.8295**
	Set5	18.35/0.7256	18.40/0.7781	18.60/0.7768	21.60/0.8060	**21.87/0.7882**
	Set14	17.12/0.7161	18.13/0.7562	19.19/0.7767	21.41/0.8031	**22.50/0.8138**
	Urban	17.45/0.6703	18.71/0.7263	19.10/0.7470	20.60/0.7485	**21.93/0.7699**

still outperforms the existing STN in estimating the affine parameters and mitigating the transformation effects.

To evaluate our proposed loss (i.e MSE, adversarial, and DCT losses), we repeat the previous experiment using the same testing images in Fig. 6. Specifically, we have transformed the original images using the same transformation and then we train the proposed framework and GAN based benchmarks using our proposed loss function. All experimental results are given in Fig. 11, 12, 13, and 14. From these figures, we can see that our proposed framework still produce better have higher PSNR/SSIM values over the numerous benchmarks.

We test the robustness of our proposed framework against the different geometric transformations, including rotation (R), rotation and scaling (RS), translation (T), and combination of rotation, scaling and translation (RTS), we carried out four experiments in total to evaluate the performances of our proposed in comparison with the existing state of the



(a)

(b)

FIGURE 14. **PSNR and SSIM values of our proposed framework and various benchmarks trained with proposed loss function: The original images are transformed by a combination of rotation, translation and scaling effects. The PSNR values are shown in (a), and SSIM values are shown in (b).**

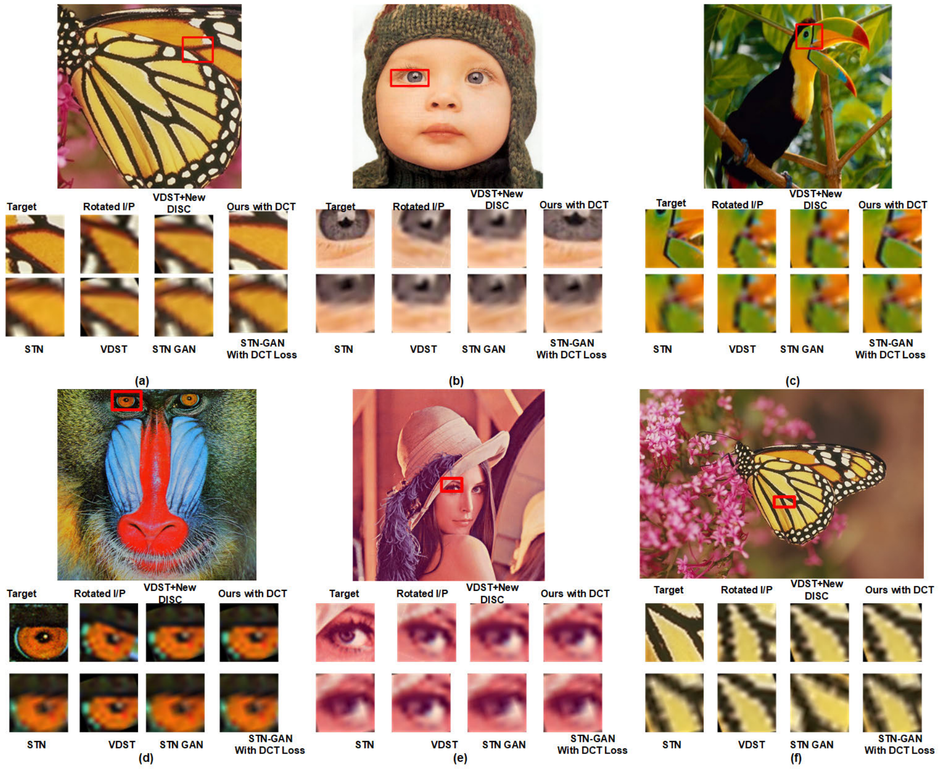


FIGURE 15. **Visual comparison samples of our proposed framework and various benchmarks: Top and third rows: original image. Under each row: The desired output, rotated input patch, our proposed framework output with generative loss function (PSNR/SSIM), our proposed framework output with our loss function output (PSNR/SSIM), existing STN, existing VDST, STN based GAN network with generative loss function, STN based GAN network with our proposed loss function output (PSNR/SSIM), respectively.**

art benchmarks. In addition, we test our proposed framework performance utilizing the existing generative loss and our proposed loss functions. Then, we compare the simulation results of our proposed framework obtained with the different benchmarks.

Table 3 shows the comparative experimental results between the existing benchmarks and our proposed framework under a number of the transformation effects with the generative loss function, and the scaling factors ×4. From the simulation results given in Table 3, we can notice that our proposed framework significantly outperforms different benchmarks in terms of both PSNR and SSIM. Specifically, we can seen that our proposed framework have 1 2 dB increment in PSNR values. Our explanation for this improvement is that our proposed framework is able to remove the geometric and background noises. Consequently, the discriminator is able to differentiate between the generated and real image is improved. Then, the generator tries generate image which are

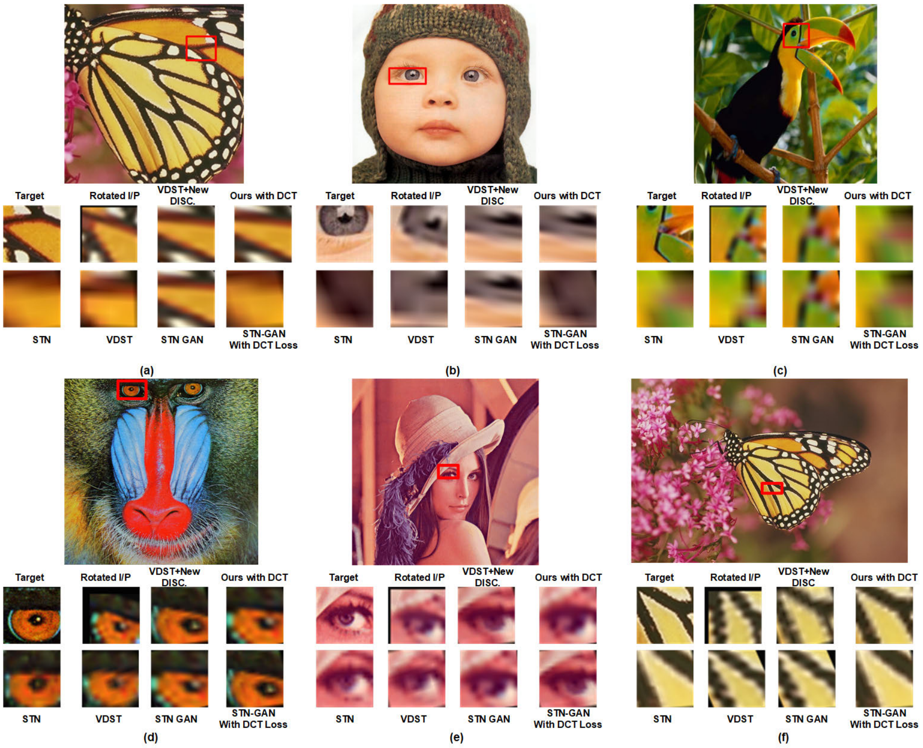


FIGURE 16. Visual comparison samples of our proposed framework and various benchmarks: Top and third rows: original image. Under each row: The desired output, RTS input patch, our proposed framework output with generative loss function (PSNR/SSIM), our proposed framework output with our loss function output (PSNR/SSIM), existing STN, existing VDST, STN based GAN network with generative loss function, STN based GAN network with our proposed loss function output (PSNR/SSIM), respectively.

more realistic by reducing the error between both generated and target images.

Table 4 show the experimental results between the existing benchmarks and our proposed framework under a number of the transformation effects with our proposed loss function, the scaling factors ×4. It can also be seen that our proposed framework outperforms the existing benchmarks in both PSNR and SSIM values. Correspondingly, it can be concluded that our proposed successfully provides a well-validated solution for tackling the effects of geometric transformations, and achieve a robust single image super-resolution.

To visually compare the experimental results between our proposed framework trained by various versions of the loss function and the existing benchmarks, we illustrate a number of samples in Fig.15 and Fig. 16 for visual inspections and subjective assessments. In Fig.15, we show visual comparisons on Set5, Set14 with a scaling factor of ×4 under the rotation(R) effect. From Fig. 15, we can conclude that our proposed framework is able to generate HR images which are more realistic and closer to the ground truth images than other benchmarks. Moreover, we can notice that our proposed framework which utilizes the loss function with the DCT loss term is able to produce more realistic images with more texture details.

To test the robustness of our proposed against stronger transformation effects, we apply the combinational effect of rotation, translation and scaling, and the corresponding experimental results are illustrated in Fig. 16. As notice again, the proposed framework still produce better performance than other benchmarks.

V. CONCLUSION

In this paper, we propose a novel deep generative adversarial network to achieve robust single image super-resolution reconstruction. Specifically, we propose an improved discriminator to enhance its capability in discriminating the fake and the original images in order to generate more realistic images. Compared with the existing state of the arts, our proposed framework is able to simultaneously perform both geometric corrections and super-resolution reconstruction. In addition, we also propose to add a new DCT loss term to improve the perceptual quality of the generated images. Extensive evaluations on widely used datasets in comparison with the existing state-of-the-art networks show that our proposed framework is able to produce more realistic and more similar images to the ground truth than those compared benchmarks.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[2] H. Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, 2008.

[3] J. S. Isaac and R. Kulkarni, "Super resolution techniques for medical image processing," in *Proc. Int. Conf. technol. Sustain. Develop. (ICTSD)*, 2015, pp. 1–6.

[4] Y. Huang, L. Shao, and A. F. Frangi, "Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6070–6079.

[5] F. Lin, C. Fookes, V. Chandran, and S. Sridharan, "Super-resolved faces for improved face recognition from surveillance video," in *Proc. Int. Conf. Biometrics*. Berlin, Germany: Springer, 2007, pp. 1–10.

[6] L. Zhang, H. Zhang, H. Shen, and P. Li, "A super-resolution reconstruction algorithm for surveillance images," *Signal Process.*, vol. 90, no. 3, pp. 848–859, 2010.

[7] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," in *Proc. Int. Conf. Articulated Motion Deformable Objects*. Cham, Switzerland: Springer, 2016, pp. 175–184.

[8] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[9] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, 1979.

[10] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP, Graph. Models Image Process.*, vol. 53, no. 3, pp. 231–239, May 1991.

[11] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, no. 2, p. 12, 2011.

[12] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[13] Z. Xiong, X. Sun, and F. Wu, "Robust Web image/video super-resolution," *IEEE Trans. Image Process.*, vol. 19, no. 8, pp. 2017–2028, Aug. 2010.

[14] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.

[15] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, p. I.

[16] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 349–356.

[17] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[18] C. Dong, C. L. Chen, K. He, and X. Tang, *Learning a Deep Convolutional Network for Image Super-Resolution*, vol. 8692. Cham, Switzerland: Springer, 2014, pp. 184–199.

[19] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[20] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1637–1645.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 136–144.

[22] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[23] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 624–632.

[24] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2414–2423.

[25] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," pp. 694–711, 2016, *arXiv:1603.08155*. [Online]. Available: https://arxiv.org/abs/1603.08155

[26] J. Jiang, H. M. Kasem, and K.-W. Hung, "A very deep spatial transformer towards robust single image super-resolution," *IEEE Access*, vol. 7, pp. 45618–45631, 2019.

[27] A. Esmaeilzehi, M. O. Ahmad, and M. N. S. Swamy, "Compnet: A new scheme for single image super resolution based on deep convolutional neural network," *IEEE Access*, vol. 6, pp. 59963–59974, 2018.

[28] P. Liu, Y. Hong, and Y. Liu, "Deep differential convolutional network for single image super-resolution," *IEEE Access*, vol. 7, pp. 37555–37564, 2019.

[29] F. Li, H. Bai, L. Zhao, and Y. Zhao, "Dual-streams edge driven encoder-decoder network for image super-resolution," *IEEE Access*, vol. 6, pp. 33421–33431, 2018.

[30] Y. Wang, L. Wang, H. Wang, and P. Li, "End-to-end image super-resolution via deep and shallow convolutional networks," *IEEE Access*, vol. 7, pp. 31959–31970, 2019.

[31] Z. Lu, Z. Yu, P. Yali, L. ShiGang, W. Xiaojun, L. Gang, and R. Yuan, "Fast single image super-resolution via dilated residual networks," *IEEE Access*, vol. 7, pp. 109729–109738, 2018.

[32] Y. Zhao, G. Li, W. Xie, W. Jia, H. Min, and X. Liu, "GUN: Gradual upsampling network for single image super-resolution," *IEEE Access*, vol. 6, pp. 39363–39374, 2018.

[33] J. S. Park, J. W. Soh, and N. I. Cho, "High dynamic range and super-resolution imaging from a single image," *IEEE Access*, vol. 6, pp. 10966–10978, 2018.

[34] X. Zhao, W. Li, Y. Zhang, and Z. Feng, "Residual super-resolution single shot network for low-resolution object detection," *IEEE Access*, vol. 6, pp. 47780–47793, 2018.

[35] L. Zhao, Q. Sun, and Z. Zhang, "Single image super-resolution based on deep learning features and dictionary model," *IEEE Access*, vol. 5, pp. 17126–17135, 2017.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[37] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: https://arxiv.org/abs/1511.06434

[38] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," 2015, *arXiv:1511.05666*. [Online]. Available: https://arxiv.org/abs/1511.05666

[39] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[40] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. NIPS*, 2015, pp. 2017–2025.

[41] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4491–4500.

[42] M. Cheon, J.-H. Kim, J.-H. Choi, and J.-S. Lee, "Generative adversarial network-based image super-resolution using perceptual content losses," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[43] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "Srfeat: Single image super-resolution with feature discrimination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 439–455.

[44] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[45] F. Li, L. Ma, and J. Cai, "Multi-discriminator generative adversarial network for high resolution gray-scale satellite image colorization," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 3489–3492.

[46] J.-H. Kim and J.-S. Lee, "Deep residual network with enhanced upscaling module for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 800–808.

[47] O.-Y. Lee, Y.-H. Shin, and J.-O. Kim, "Multi-perspective discriminators-based generative adversarial network for image super resolution," *IEEE Access*, vol. 7, pp. 136496–136510, 2019.

[48] Á. Arcos-García, J. A. Álvarez-García, and L. M. Soria-Morillo, "Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods," *Neural Netw.*, vol. 99, pp. 158–165, Mar. 2018.

[49] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[50] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[51] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Surrey, U.K., Sep. 2012, pp. 1–10.

[52] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Berlin, Germany: Springer, 2010, pp. 711–730.

[53] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5197–5206.

[54] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, 2017.

[55] D. CireçAn, U. Meier, J. Masci, and J. Schmidhuber, "2012 special issue: Multi-column deep neural network for traffic sign classification," *Neural Netw.*, vol. 32, no. 1, pp. 333–338, 2012.

HOSSAM M. KASEM received the Ph.D. degree from the Egypt-Japan University of Science and Technology (E-JUST), Egypt, in 2015. From 2015 to 2017, he was an Assistant Professor with the Faculty of Engineering, Tanta University, Egypt. Since 2017, he has been a Postdoctoral Fellow with the Research Institute for Future Media Computing, Shenzhen University, China. His research interests include deep learning applications in digital multimedia analysis, wireless communication, and signal processing application in multimedia.

KWOK-WAI HUNG received the B.Eng. and Ph.D. degrees from Hong Kong Polytechnic University, in 2009 and 2014, respectively. From 2014 to 2016, he was a Research Engineer with Huawei and ASTRI. Since 2016, he has been an Assistant Professor with the Research Institute for Future Media Computing, Shenzhen University, China. His research interests include deep learning applications in digital multimedia processing and signal processing applications in multimedia.

JIANMIN JIANG received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 1994. From 1997 to 2001, he was a Full Professor of computing with the University of Glamorgan, Pontypridd, U.K. In 2002, he joined the University of Bradford, Bradford, U.K., as the Chair Professor of Digital Media and the Director of the Digital Media and Systems Research Institute. He was a Full Professor with the University of Surrey, Guildford, U.K., from 2010 to 2014, and the Distinguished Chair Professor (1000-Plan) with Tianjin University, Tianjin, China, from 2010 to 2013. He is currently the Distinguished Chair Professor and the Director of the Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He has published around 400 refereed research articles. His current research interests include, image/video processing in compressed domain, digital video coding, medical imaging, computer graphics, machine learning and AI applications in digital media processing, and retrieval and analysis. He was a Chartered Engineer, a fellow of IEE and RSA, a member of the EPSRC College, U.K., and an EU FP-6/7 Evaluator.