Unsupervised Single Image Super-Resolution Network (USISResNet) for Real-World Data Using Generative Adversarial Network

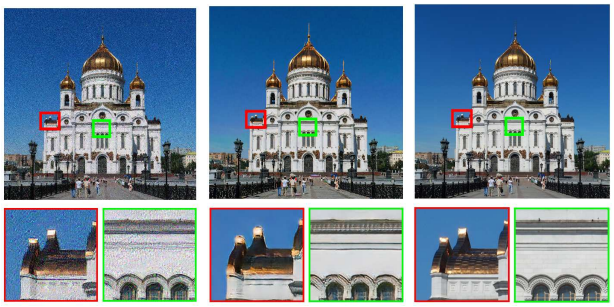
Kalpesh Prajapati[1], Vishal Chudasama[1], Heena Patel[1], Kishor Upla[1, 2],
Raghavendra Ramachandra[2], Kiran Raja[2], Christoph Busch[2]
[1]Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India.
[2]Norwegian University of Science and Technology (NTNU), Gjøvik, Norway.
{kalpesh.jp89, vishalchudasama2188, hpatel1323, kishorupla}@gmail.com,
{raghavendra.ramachandra, kiran.raja, christoph.busch}@ntnu.no

Abstract

Current state-of-the-art Single Image Super-Resolution (SISR) techniques rely largely on supervised learning where Low-Resolution (LR) images are synthetically generated with known degradation (e.g., bicubic downsampling). The deep learning models trained with such synthetic dataset generalize poorly on the real-world or natural *data where the degradation characteristics cannot be fully modelled. As an implication, the super-resolved images obtained for real LR images do not produce optimal Super-Resolution (SR) images. We propose a new SR approach to mitigate such an issue using* unsupervised *learning in Generative Adversarial Network (GAN) framework - USISResNet. In an attempt to provide high quality SR image for perceptual inspection, we also introduce a new loss function based on the Mean Opinion Score (MOS). The effectiveness of the proposed architecture is validated with extensive experiments on NTIRE-2020 Real-world SR Challenge validation (Track-1) set along with testing datasets (Track-1 and Track-2). We demonstrate the generalizable nature of proposed network by evaluating real-world images as against other state-of-the-art methods which employ synthetically downsampled LR images. The proposed network has further been evaluated on NTIRE 2020 Real-world SR Challenge dataset where the approach has achieved reliable accuracy.*

1. Introduction

The High-Resolution (HR) images consist of detailed information of the scene which helps in many use-cases such as in perceptual applications. While it is preferred to have HR images in many applications, a High Definition (HD) camera is an imperative requirement. Given the cost of HD cameras, not all real-world applications can afford to integrate the HD cameras. As an alternative, the software based approaches such as *Super-Resolution (SR)* techniques are employed widely in order to increase the resolution of the given Low-Resolution (LR) image. Super-resolving an image is a *classical* problem in the computer-vision community and despite the extensive amount of work, the problem is yet an open research problem due to number of covariates in the problem such as ill-posed nature of solutions, complexity and unavailability of true perceptual quantitative measures amongst many others [2].



(a) ESRGAN [47] (b) Proposed (c) Ground-truth

Figure 1: The SR results obtained using (a) ESRGAN [47] and (b) proposed methods. (c) The ground-truth original HR image.

The super-resolving of the LR images has been achieved using Single-Image approaches, dictionary based approaches and lately using the deep learning approaches such as Convolutional Neural Networks (CNNs). Combining the approaches of Single-Image SR and the CNNs, recent works have obtained state-of-the-art performance, both in terms of quantitative as well as qualitative assessments [47, 53, 2]. However, the proposed CNN models so far proposed use supervised learning by employing a dataset of LR-HR pair of images. In the pursuit of supervision for such models, the training process employs a single HR dataset while the corresponding LR images are synthesized through known degradation such as bicubic downsampling.

Despite such a mechanism of creating the LR-HR pairs, it can be noted that synthetically generated images from true HR image do not fully represent the true statistical modelling of LR images with a wide variety of factors. For instance, the synthesized LR images may not represent the *natural* image characteristics such as inevitable sensor noise, artifacts which are always present in the *natural or real* image [32]. As a direct impact of such data being used with synthetic degradation, the CNN models trained are not able to generalize for real LR images captured from camera. Such limitations hinder in achieving reliable accuracy as they fail to scale when real data is presented for SR tasks[33].

In order to solve above limitation, it is necessary to train the given CNN model on a dataset of true LR-HR pair. Due to practical limitations, there are not many datasets which provide true LR-HR pairs. The hardware changes to capture such LR-HR pair need varying resolutions through the spatial resolutions of the camera itself [7, 52]. Not only is such design cumbersome process but also needs extra hardware. To this end, another way to tackle this problem is to use unsupervised learning in SISR tasks which is similar to *blind* SR. In this paper, we propose a CNN based model to obtain SR of given LR observation through unsupervised learning using a Single Image Super-Resolution approach which we refer as *USISResNet*. The proposed architecture is trained with unpaired LR-HR images on the dataset of NTIRE 2020 Real-world SR Challenge [34]. We employ a Generative Adversarial Network (GAN) in unsupervised manner to obtain perceptually superior SR image. Additionally, we also introduce a novel Mean Opinion Score (MOS) loss within GAN framework in order to further enhance the SR results.

The proposed approach is benchmarked with other state-of-the-art methods including the winner of PIRM2018 challenge - ESRGAN [23, 47] SR method to demonstrate the applicability of the method. The superior results of the proposed approach can be seen in the Fig. 1 alongside the SR results obtained using the ESRGAN [47]. It can be noticed from the Fig. 1 that the proposed method obtains better SR image in terms of preserving high-frequency details when compared to similar approach using ESRGAN method [47]. The key contributions of the work can therefore be summarized as:

- Many of unsupervised SR methods based on deep learning [32, 50] adopt cycle consistency loss [55] in order to obtain the real-world distributions but do not fully capture possible variations.
- The proposed GAN based method is trained in an end-to-end manner without additional networks to estimate the real-world data distribution in contrast to earlier work [32] in a cycle GAN [55] trained in a sequential manner.
- A new loss function using the Mean Opinion Score (MOS) is additionally introduced into the GAN framework to improve the superiority of the perceptual quality. The effectiveness of new loss is also verified experimentally showing committed improvement in the SR results when compared to same approach without the custom loss in the proposed SR method.

In the rest of the paper, Section 2, reviews a set of existing deep learning based SR methods. The detailed description of the proposed method and the rationale is presented in Section 3. The experimental results are presented in Section 4 following the conclusions in Section 5.

2. Related Work

A number of traditional SR methods [13, 38, 18, 10, 27, 49] have been proposed after first SR approach introduced by Tsai and Huang [46]. Given in many cases there is only single image available for a specific task, many works have focused on the Single Image based Super-Resoulution (SISR). In this direction, the success of deep learning, specifically Convolutional Neural Networks (CNN) has been exploited in SISR to obtain superior SR images by modelling the relationship between LR and HR images learnt from available large datasets. We therefore provide a breif overview of relevant SR approaches proposed in the recent years.

The first CNN based framework for SR was proposed by Dong et al. [10] which was termed as *Super-Resolution Convolutional Neural Network (SRCNN)*. The approach consisted of shallow CNN architecture with 3 convolutional layers. Later on, the same architecture was modified in order to obtain better SR image by increasing the depth from 3 to 20 layers in VDSR method [24]. Both the works [10, 24] used a bicubically interpolated LR input image for learning the missing high-frequency details in the SR image using deep models and the approaches are referred as pre-upsampling based SR methods. The other representative works based on such concept are reported in [25, 44]. In contrast, a set of CNN based SR methods adopt post-upsampling strategy in order to obtain SR methods [11, 40]. In addition to that, some SR works are also based on progressive upsampling idea in order to obtain SR image for higher upscaling factor [27, 20]. Recently, many SR approaches such as SRFeat-M [36], MSRN [29], EDSR [30] and RCAN [53] have obtained state-of-the-art performance in SISR task using CNN models. These methods use L_1 or L_2 loss in order to achieve better quantitative metrics such as PSNR and SSIM of the SR results.

Furthermore, the use of adversarial training [19] is also exploited in many SR approaches in order to improve the perceptual quality of SR image. Ledig et al. [28] proposed single image SR using GAN termed as *SRGAN* by using a deep residual network (ResNet) with skip connection [22]. Following this work, many successful GAN based models have been proposed in the literature [37, 47] to further en-

hance the quality of SR image.

As mentioned earlier, all the aforementioned deep learning based SR approaches are trained in highly supervised manner in which LR images are synthesized from bicubic downsampling and hence they result in domain shift issue [32] simply by modelling LR-HR domain shift. Alternatively, true LR-HR pairs can be used to train the deep models for SR. In this direction, Cai et al. [6] introduced RealSR dataset of true LR-HR images and also introduced the challenge on above dataset in NTIRE 2019 [5]. Many representative works on RealSR dataset have been reported recently [41, 9, 14, 16, 12, 48, 26]. However, creating large scale true LR-HR pairs is a cumbersome task which requires specially designed hardware to acquire same scene with varying resolutions and with zero registration error. Further, such datsets are limited with selected upscaling factors and hence it is very expensive to scale for varying scaling factors.

To circumvent this problem, Lugmayr et al. [32] introduced unsupervised way to train the CNN model for SISR task. In their method, first they obtain distribution of real-world images (e.g., sensor noise and other artifacts) in bicubic downsampled LR images by using cycle consistency loss [55]. Later, such LR along with it's true HR image were used to prepare datset which is then used to train the SR network to obtain final SR result. In order to reduce the severe over-fitting and poor generalization problems, they trained their cycle GAN frameworks in sequential manner. Similarly, authors in [50] use the concept of unsupervised learning based on cycle GAN [55] to obtain SR images. Here, authors used two sets of cycle GANs in order to train the SR network. First cyele GAN is used to obtain noise-free LR observation and in the second step another cycle GAN network is used to obtain final SR image. To further develop such novel idea of unsupervised learning for SR, Lugmayr et al. [32] introduced real-world SR challenge based on above concept in ICCV 2019 (*called AIM 2019 Challenge [33]*) and in conjunction with CVPR 2020 (*called as NTIRE 2020 Real-world SR Challange [34]*). Further, Fritsche et al. [15] use the ESRGAN [47] model to learn features related to low and high frequencies separately in the LR image and proposed an SR approach for Real-world data which was a winner in AIM 2019 SR challenge [33]. As a special case, authors in [4] learn the downsampling process in order to better generalize for face recognition alone by using real face images and the face priors to super-resolve it.

Many SISR approaches [35, 39, 54, 21] are also based on the concept of *blind* SR in which information related to degradation process are not provided. They estimate the blur kernel in traditional way using probabilistic framework [35, 39] and also by using deep models [54, 21] recently. Lastly, a CNN based method to train the model from LR image itself was proposed later in [42]. However, the approach requires downsampled LR image of known kernel.

2.1. Constraints Noted from Related Works

Based on the review of current unsupervised SR methods for SISR, we note the following constraints with existing works:

- Many of unsupervised SR methods based on deep learning [32, 50] adopt cycle consistency loss [55] in order to obtain the real-world distributions.
- The blind SR methods in [54, 21] rely on the estimation of blur kernel in order to generate LR image of real-world distribution.
- The SR approach in [42] is based on training from given LR itself; however, it requires known downsampling kernel in order to generate faithful SR results.

3. Proposed Method

To overcome above constraints of the existing research works in unsupervised SR, we propose a GAN based deep learning architecture. In Fig. 2(a), we depict the architecture design of the proposed model which basically consists three networks: Generator, Discriminator and Quality Assessment (QA) networks. The generator network generates super-resolved images with desired upscaling factor (fixed to 4 in this work). The unpaired SR and HR images are provided to the discriminator network. While the aim of discriminator network is to discriminate the content of HR images from the SR images, the standard loss involved in GAN helps generator network to force the content of LR image to match it with that of HR image. Additionally, the QA network is used to assess the quality of SR image in terms of newly introduced Mean Opinion Score (MOS) based loss and hence it also helps generator network further to realize better perceptual quality of SR image. Inspired by [4], we use bicubically upsampled LR image with the SR image to obtain the content loss (i.e., L_1 norm). Such loss helps the SR image to preserve the content of LR observation in the final SR image. The individual description of each network is elaborated in the following paragraphs.

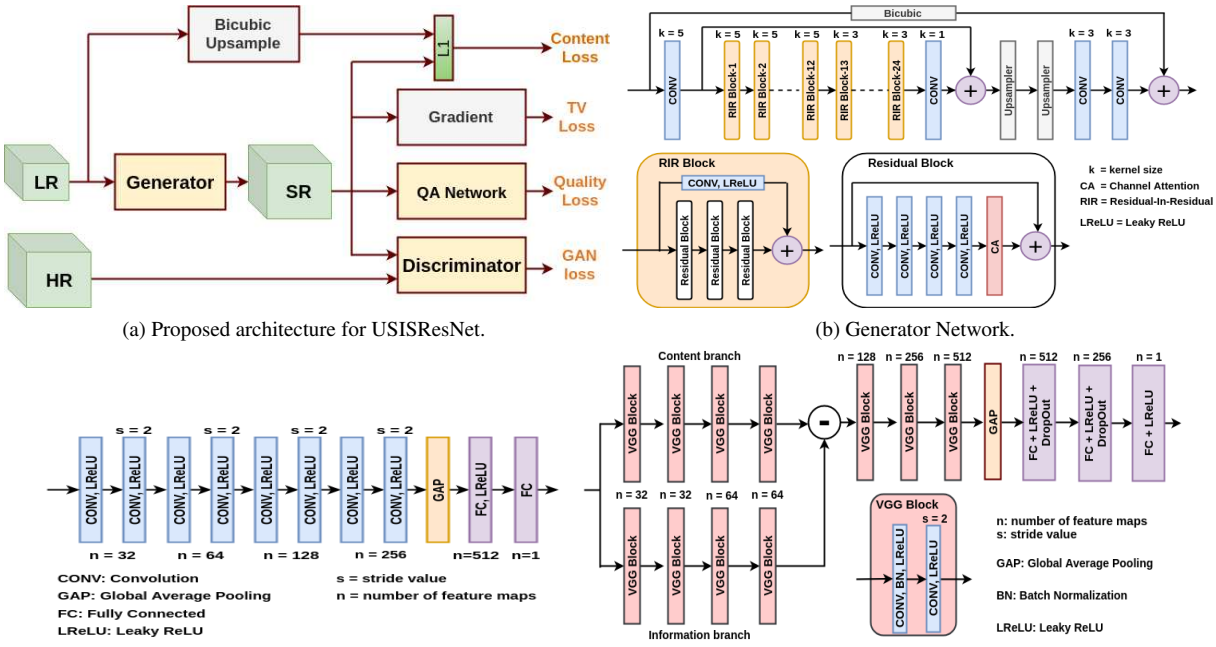
Generator Network (G):

The architecture design of the generator network is displayed in Fig. 2(b). The design of the generator network is categorised into three modules: Low-Frequency Feature Extraction (LFFE), High-Frequency Feature Extraction (HFFE) and Reconstruction (REC) modules. The LFFE module consist one convolution layer with kernel size of 5 and feature maps of size 32 with stride value of 1. The LFFE module extracts the low-frequency details from the LR image (I^{LR}) as,

$$I_{low-freq} = f_{LFFE}(I^{LR}), \tag{1}$$

where, f_{LFFE} denotes the operation of the LFFE module.

In order to obtain high-frequency details pertaining to edges and structures from the feature maps obtained from



(a) Proposed architecture for USISResNet.

(b) Generator Network.

(c) Discriminator Network.

(d) Quality Assessment (QA) network.

Figure 2: The architecture design of the proposed network USISResNet for SISR.

LFFE module, the feature maps of LFFE module are passed through the HFFE module. The HFFE module consists of $m = 24$ number of Residual-in-Residual (RIR) blocks and one long skip connection. As depicted in Fig. 2(b), each RIR block consists individual residual network whose design has several convolution layers, short skip connection and one Channel Attention (CA) module. Along with long skip connection, the short skip connection in RIR block is utilized to reduce the vanishing and exploding gradient problems. The CA module is employed to re-scale the channel-wise features adaptively for better generalization of the proposed generator network which is inspired from [53]. The output feature maps of the HFFE module are represented as,

$$I_{HFFE} = f_{HFFE}(I_{low-freq}). \tag{2}$$

Here, the f_{HFFE} denotes the function of the HFFE module. Finally, the SR image is reconstructed through the reconstruction module (REC). Specifically, this module has two upsampler blocks followed by two convolution layers to obtain the residual SR image as indicated by

$$I^{SR}_{residual} = f_{REC}(I_{HFFE}), \tag{3}$$

where, f_{REC} indicates the reconstruction function of the REC module. Additionally, we also implement the global residual path in which input LR observation (I^{LR}) is passed through a bicubic interpolation layer with upscaling factor 4 which produces the super-resolved image (I^{SR}_{GRL}). Such Global Residual Learning (GRL) helps the network to learn the identity function for I^{LR} and it also stabilizes the training process. Finally, the network generates the SR image (I^{SR}) at upscaling factor of 4 as defined by,

$$I^{SR} = I^{SR}_{residual} + I^{SR}_{GRL}. \tag{4}$$

Discriminator Network (D):

The design of the Discriminator network is displayed in Fig. 2(c). Here, we follow the guidelines suggested by Radford et al. [37] to design it's architecture. It consists of eight convolution layers followed by Global Average Pooling (GAP) layer and two Fully Connected (FC) layers. The strided convolution layers are used with the stride value of 2 whenever the number of feature maps are doubled. The Leaky ReLU activation function with leaky constant variable of 0.2 is used in all convolution layers except in FC last layer. The use of GAP pooling layer instead of the flattening layer helps to reduce the number of training parameters. The discriminator network takes unpaired SR and HR images as input and gives probability value of the corresponding image as output. Finally, this probability value is used in standard GAN based adversarial loss function and update the discriminator and generator networks to improve the generalization of both the networks.

Quality Assessment (QA) network:

In the proposed method, we incorporate a novel loss based on Mean Opinion Score (MOS) in order to improve the perceptual quality of SR image. Simultaneously, it also helps to suppress the unpleasant noise in the generated SR images through the proposed Quality Assessment (QA) CNN

network whose design is depicted in Fig. 2(d). It has several VGG blocks followed by three FC layers and the VGG blocks are made up of two convolution layers. As indicated in figure, in initial two FC layers we use dropout in order to overcome with the problem of over-fitting of weights during training process. In QA network, there are two branches at the beginning of the QA network called as *content* and *information* branches. Both branches consists of four VGG blocks. In order to extract the noise details from the input image, the output of the content branch is subtracted from the output of the information branch. Instead of flattening layer, the adaptive GAP layer is utilized to reduce the trainable parameters of the QA network from $30M$ to $5M$. This network is trained using KADID-10K dataset [31] which consists different noisy images with corresponding MOS values and hence, such design of QA network helps to improve the MOS value of the SR image. The trained QA network takes the generated SR image as input and gives the quality score based on MOS as output. Then, this score is used in the proposed GAN model as quality loss function to further improve the fidelity of of the generated SR image.

Loss functions

The proposed model is trained using different loss functions. The overall loss function is a weighted combination of pixel-wise content loss (i.e., $l_{content}$), Total Variation (TV) loss (i.e., l_{tv}), the proposed Quality Assessment (QA) loss (i.e., l_{qa}) and standard GAN losses (i.e., l_{gen} and l_{disc}) which can be represented as,

$$Loss = \lambda_1\, l_{content} + \lambda_2\, l_{tv} + \lambda_3\, l_{qa} + \lambda_4\, l_{gen} + l_{disc}, \quad (5)$$

where, $\lambda_1, \lambda_2, \lambda_3$ and λ_4 indicate associated weighting factors of each loss. The details of these loss functions are as follows:

Content loss ($l_{content}$): We use L_1 pixel wise content loss between the output of the generator network (i.e., I^{SR}) and the LR image after passing through the bicubic upsampling operation (i.e, I^{up}). We impose such an idea from [4] in order to enforce the output of the generator network to preserve the similar content with the LR image in unsupervised learning which is mathematically defined over N number of training samples as,

$$l_{content} = \sum_{N} \sum_{i,j} \|I^{SR}_{i,j} - I^{up}_{i,j}\|_1. \quad (6)$$

Total variation (TV) loss (l_{tv}): In order to eliminate the noise from the generated SR image, we also use the Total Variation (TV) loss function [8] along with other losses. The TV loss is the sum of the absolute differences for neighboring pixel-values in the input images. It can be represented mathematically as,

$$l_{tv} = \sum_{N} \sum_{i,j} \|I^{SR}_{i+1,j} - I^{SR}_{i,j}\|_1 + \|I^{SR}_{i,j+1} - I^{SR}_{i,j}\|_1. \quad (7)$$

Quality Assessment (QA) loss (l_{qa}): As mentioned earlier, to improve the perceptual quality by reducing the unpleasant noise details from the generated SR image, QA loss is proposed in a manner to improve the quality of SR image. The QA network is trained on KADID-10K dataset [31] having MOS values of different noisy images within the scale of $1-5$ where maximum value indicates best visualization. These MOS values are used as label in the proposed quality loss function. In the proposed method, the predicted MOS value is maximized in order to improve the MOS score on the SR images. The proposed QA loss function (i.e., l_{qa}) is defined on the generated SR image as to minimize overall loss in Equation (5) as,

$$l_{qa} = \sum_{N} (5 - Q(I_{SR})), \quad (8)$$

where, $Q(I_{SR})$ represents MOS value of SR image obtained from the proposed QA network.

Standard GAN losses (l_{gen} and l_{disc}): In order to improve the perceptual quality of the generated SR image and match the it's quality with clean HR image (I^{HR}), we use the standard GAN loss functions (i.e., l_{gen} and l_{disc}). In the adversarial GAN training, the generator network is trained to learn the mapping between LR to HR image space. The generator loss function based on N number of training samples is given as,

$$l_{gen} = \sum_{N} -\log(D(G(I^{LR}))). \quad (9)$$

Simultaneously, the discriminator network is also trained as,

$$l_{disc} = \sum_{N} \big[-\log(D(I^{HR})) - \log(1 - D(G(I^{LR}))) \big]. \quad (10)$$

Here, $D(I^{HR})$ is the probability value of clean HR image and $D(G(I^{LR}))$ is the probability of the reconstructed SR image from the LR observation.

4. Experimental Analysis

To validate the effectiveness of the proposed model, various experiments[1] have been carried out on NTIRE-2020 Real-world SR Challenge validation (Track-1) as well as testing datasets (Track-1 and Track-2) [34] and it's detailed description is presented in this section. In this challenge, to perform unsupervised training, the unpaired LR and HR images are provided from two different datasets of Flickr2k [30, 45] and DIV2K [1], respectively. The total number of images in above datasets are 2650 and 800 which are used for training of the model. Additionally, the LR images from Flickr2k dataset [30, 45] are given with unknown degradation in order to apply real-world data degradation. However, such degradation remains unknown to the model during training and hence it provides practical scenario of un-

[1] All the experiments are performed on a computer with Intel Xeon(R) CPU with 128GB RAM and NVIDIA Quadro P5000 GPU with 16GB memory.

supervised learning with real-world images. For validation, 100 number of LR-HR image pairs from DIV2K dataset [1] had been employed as provided by organizers of NTIRE-2020 Real-world SR Challenge [34].

4.1. Hyperparameter Tuning

In the proposed model, number of filters are set to the value of 32 in Generator, Discriminator and QA networks. Further, it is optimized with the weighted combination of different loss functions as given by Equation (5) in which weighting factors such as $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are set empirically to 0.1, 5×10^{-11}, 2×10^{-6} and 10^{-3} values, respectively. During training, the LR images are augmented with random rotation of 0° or 90°, random horizontal flipping and random cropping operations. The proposed model is optimized using the Adam optimizer with the learning rate of 10^{-4} and it is trained upto $1,20,000$ number of iterations with batch size of 32. The number of trainable parameters in Generator and Discriminator networks of the proposed model are 5.1M and 1.3M, respectively

Additionally, the proposed QA network is trained separately on KADID-10K dataset [31] which is the largest dataset for image quality assessment with 25 different kinds of artificial noise degradation. This dataset consists 10124 total number of images which has been divided into 7124, 1000 and 2000 number of images for training, validation and testing, respectively. The QA network is trained upto $50,000$ number of iterations with batch size of 16 and it is optimized with Adam optimizer. The learning rate in this training is set to 2×10^{-4} which is decayed by half at 20%, 40%, 60% and 80% of total number of iterations. During the training process, the use of dropout in the architecture of the QA network helps to overcome with the problem of over-fitting of the weights. Once the QA network is trained, then it is used as pre-trained model in the proposed model in order to find quality loss for the generated SR image. The number of training parameters of the proposed QA network is 5.35M.

4.2. Result Analysis

Here, we present the qualitative and quantitative measurements of the proposed and other CNN based state-of-the-art exiting SISR methods. First, the effectiveness of the QA network and TV loss in the proposed model are discussed and later, result comparisons of the proposed method with other state-of-the-art methods is depicted.

4.2.1 Effectiveness of the QA Network

In the proposed model, we introduce a novel QA network in order to improve the perceptual quality of SR images with simultaneous elimination of unpleasant noise details in the SR image. In this subsection, we justify it's effectiveness by conducting experiments on the testing dataset of the Kadid-10K [31] and validation set of NTIRE 2020 Real-world SR Challenge [34]. As mentioned earlier, the proposed QA network is trained on Kadid-10K [31] dataset and then the obtained pre-trained QA network is used in training of the proposed GAN model via quality loss function. The QA network takes an image and gives the value of MOS score of that image. In Fig. 3, we display the graph between the predicted and actual MOS scores on Kadid-10K [31] testing dataset. Ideally, this relation must be linear with unit slope when it exactly fits to data with quantitative measure Spearman's Rank Correlation Coefficient (SROCC) [43] value of 1. However, it can be observed from the Fig. 3 that close linearity with SROCC value of 0.89 (very close to ideal) is obtained using the proposed QA network.

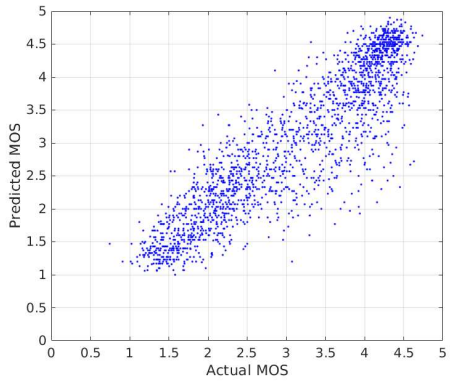


Figure 3: The performance of QA network in terms of actual and predicted MOS score values on KADID-10K [31] testing dataset.


(a) w/o QA Net (b) w/o TV loss (c) proposed (d) Ground-truth

Figure 4: The SR results obtained using the proposed method (a) without QA network, (b) without TV loss and (c) proposed. (d) The ground-truth HR image. (better visualization in zoomed images)

Furthermore, to understand the role of the QA network in the proposed model, we conduct an additional experiment on NTIRE 2020 Real-world SR Challenge validation dataset in which the proposed model is trained without using QA network and hence without it's quality loss function in the GAN framework. We present the qualitative and quantitative analysis of this experiment in Fig. 4(a) and in Table 1, respectively. One can notice from Fig. 4(a) that SR images obtained using proposed method without QA net-

Table 1: The quantitative comparison to validate the effectiveness of the proposed QA network.

Proposed Model	PSNR ↑	SSIM ↑	LPIPS ↓
Without QA Network	20.77	0.5333	0.425
Proposed	**21.71**	**0.5895**	**0.375**

work have more structural degradation in addition to noise details when compared to that of the SR results obtained with QA network i.e., proposed method (see Fig. 4(c)). Furthermore, Table 1 shows the quantitative comparison in terms of PSNR, SSIM and LPIPS measures obtained from the SR results of the proposed model trained with and without QA network. Similarly, it can be observed here that the proposed model trained using the QA network achieves better quantitative measurements as compared to that of the proposed model trained without the QA network.

4.2.2 Effectiveness of the TV loss function

We further demonstrate the impact of TV loss in the proposed method by conducting an additional experiment on the validation set of NTIRE-2020 Real-world SR challenge and it's results are depicted in Fig. 4(b). As noted from earlier works, TV loss is used to remove the unwanted noisy pixels from the SR image [8] and our experiments verify the same as shown in from Fig. 4(b). The SR images obtained using the proposed method without TV loss are relatively noisier than that of obtained with the proposed method (i.e., trained with TV loss function, see Fig. 4(c)) and hence we use the same in the proposed method.

4.2.3 Comparison with State-of-the-art Methods

In order to see the visual improvement obtained using the proposed method, we compare SR results obtained using the proposed model with three different state-of-the-art SR models named as ESRGAN [47], RCAN [53] and MsDNN [17]. We choose these methods based on following criteria: Currently, the ESRGAN model [47] is a GAN based state-of-the-art SISR method which is proposed to improve the perceptual quality of the SR image and it is also a winner in PIRM 2018 [23]. Apart from GAN framework, the RCAN model [53] is CNN based state-of-the-art SISR method. The MsDNN model [17] is the state-of-the-art method for RealSR dataset [6] proposed recently in CVPR 2019 workshops. The SR results of these methods are obtained using their pre-trained model. Regarding training of these methods, the ESRGAN [47] and RCAN [53] have been trained in their pre-trained model with single HR dataset in which LR image is prepared from the bicubic downsampling as referred in [32]. Similarly, the MsCNN pre-trained SR method [17] is trained on RealSR dataset [6].

The Quantitative Evaluation

We further compare the SR results in terms of distortion measures (i.e., PSNR and SSIM). As noted from earlier works, the distortion measures have not been more effective when considering perception based SR comparison [23, 3, 34]. We therefore compare these methods in terms of perception based metric called LPIPS [51]. The lower value of LPIPS represents better perceptual quality of SR image. Table 2 shows such quantitative comparison of the SR results obtained using the proposed along with other state-of-the-art methods experimented on NTIRE-2020 Real-world SR challenge Track-1 validation dataset. It can be observed from the Table 2 that the proposed model attains 0.375 LPIPS value on validation dataset which indicates that the proposed method performs better when compared to other existing SR methods.

Table 2: The quantitative comparison of the proposed and other exiting SR methods on NTIRE 2020 Real-world SR Challenge validation dataset (Track-1).

Method	PSNR ↑	SSIM ↑	LPIPS ↓
MsDNN [17]	25.08	**0.7079**	0.482
RCAN [53]	**25.31**	0.6402	0.576
ESRGAN [47]	19.04	0.2422	0.755
Proposed	21.71	0.5895	**0.375**

Table 3: The quantitative measurements obtained using the proposed method on NTIRE-2020 Real-world SR Challenge Track-1 testing dataset.

	PSNR ↑	SSIM ↑	LPIPS ↓
Proposed	21.22	0.576	0.397

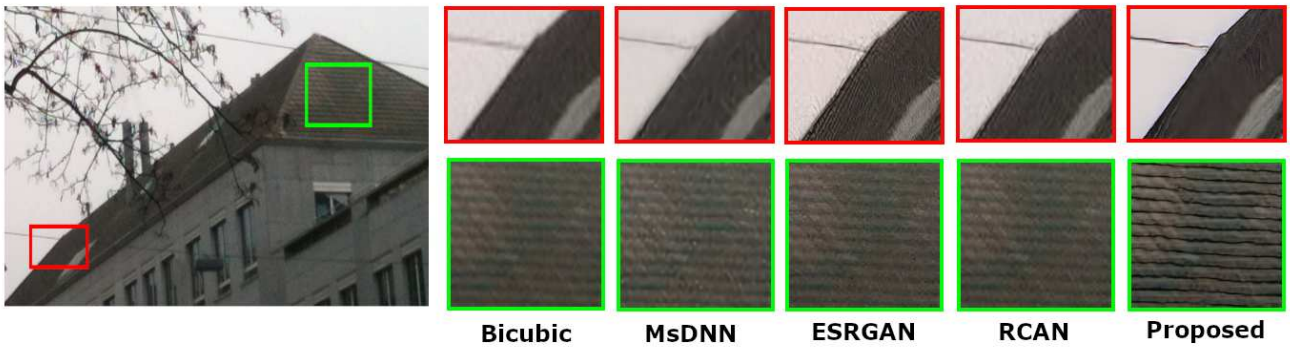
The Fidelity of SR Results

In order to verify the fidelity of the proposed SR technique, we show the visual comparison with three different datasets. We use NTIRE-2020 Real-world SR Challenge Track-1 validation dataset in which original HR images are available. Additionally, we also show SR results obtain on the testing datsets of NTIRE-2020 Real-world SR Challenge Track-1 (*called Image Processing Artifacts*) and Track-2 (*called Smartphone Images*) in which original HR images are not provided. In Fig. 5(a), we show the visual comparison of a single image from NTIRE-2020 Real-world SR challenge Track-1 validation dataset. Here, for better visualization, a zoomed-in patches of SR results of the proposed as well as MsDNN [17], ESRGN [47] and RCAN [53] models are depicted. It can be observed from the Fig. 5(a) that the proposed model exhibits better high-frequency details with less noise artifacts than the other models and it also generates SR image which is close to the original HR image.

Moreover, the SR results obtained using proposed and other models on NTIRE-2020 Real-world SR Challenge Track-1 and Track-2 testing datasets are displayed in Fig. 5(b) and Fig. 5(c), respectively. One can observe in these figures that the existing state-of-the-art methods contain the impact of additional noise and do not preserve the high-frequency details. However, the proposed model pro-

(a) NTIRE-2020 Real-world SR Challenge Track-1 validation dataset

(b) NTIRE-2020 Real-world SR Challenge Track-1 testing dataset (Image Processing Artifacts)



(c) NTIRE-2020 Real-world SR Challenge Track-2 testing dataset (Smartphone Images)

Figure 5: The comparison of SR results obtained using the proposed and other exiting SR methods on different datasets.

duces better perceptual SR images with more preservation of texture details and with reduction of unpleasant noisy pixels than that of the other models. Furthermore, in Table 3, we display the quantitative measurements in terms of PSNR, SSIM and LPIPS metrics obtained from the NTIRE-2020 Real-world SR challenge Track-1 testing dataset [2].

5. Conclusion

In order to mitigate the issue of bicubic downsampling in the supervised training, in this paper, we propose an SR approach with *unsupervised* learning in standard GAN framework. Additionally, we also introduce novel loss based on Mean Opinion Score (MOS) in order to further enhance the perceptual quality of the SR image. The proposed method effectively generalizes the characteristic of real image verified on NTIRE 2020 validation and testing datasets eliminating the need for true LR-HR pairs. The proposed approach has been validated using number of performance measures. The generalizable nature of the proposed approach can be exemplified with the results obtained on NTIRE 2020 Real-world SR Challenge datasets evaluated independently.

[2]The approach was participated to Track-1 and obtained 14th position based on LPIPS, and 19th position on the basis of PSNR and SSIM.

Acknowledgment

This work was supported by ERCIM, who kindly enabled the internship of Kishor Upla at NTNU, Gjøvik. Authors are also thankful to Science and Engineering Research Board (SERB), a statutory body of Department of Science and Technology (DST), Government of India for providing support for this research work (ECR/2017/003268).

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131, 2017. 5, 6

[2] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *arXiv*, 2019. 1

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 7

[4] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 187–202, Cham, 2018. Springer International Publishing. 3, 5

[5] Jianrui Cai, Shuhang Gu, Radu Timofte, and Lei Zhang. Ntire 2019 challenge on real image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[6] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pages 3086–3095, October 2019. 3, 7

[7] C. Chen, Z. Xiong, X. Tian, Z. Zha, and F. Wu. Camera lens super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1652–1660, 2019. 2

[8] Qiang Chen, Philippe Montesinos, Quan Sen Sun, Peng Ann Heng, et al. Adaptive total variation denoising based on difference curvature. *Image and vision computing*, 28(3):298–306, 2010. 5, 7

[9] Guoan Cheng, Ai Matsune, Qiuyu Li, Leilei Zhu, Huaijuan Zang, and Shu Zhan. Encoder-decoder residual network for real super-resolution. In *CVPR Workshops*, June 2019. 3

[10] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, Feb 2016. 2

[11] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, pages 391–407, Oct 2016. 2

[12] Chen Du, He Zewei, Sun Anshun, et al. Orientation-aware deep neural network for real image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[13] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 6(12):1646–1658, Dec 1997. 2

[14] Ruicheng Feng, Jinjin Gu, Yu Qiao, and Chao Dong. Suppressing model overfitting for image super-resolution networks. In *CVPR Workshops*, June 2019. 3

[15] M. Fritsche, S. Gu, and R. Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3599–3608, 2019. 3

[16] Shangqi Gao and Xiahai Zhuang. Multi-scale deep neural networks for real image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[17] Shangqi Gao and Xiahai Zhuang. Multi-scale deep neural networks for real image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 7

[18] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 349–356, Sep. 2009. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[20] K. Grm, W. J. Scheirer, and V. Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, 29(1):2150–2165, 2020. 2

[21] J. Gu, H. Lu, W. Zuo, and C. Dong. Blind super-resolution with iterative kernel correction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1604–1613, 2019. 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[23] Andrey Ignatov, Radu Timofte, Thang Van Vu, et al. Pirm challenge on perceptual image enhancement on smartphones: Report. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 315–333, Cham, 2019. Springer International Publishing. 2, 7

[24] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE CVPR*, pages 1646–1654, June 2016. 2

[25] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, June 2016. 2

[26] Junhyung Kwak and Donghee Son. Fractal residual network and solutions for real super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[27] W. Lai, J. Huang, N. Ahuja, and M. Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2599–2613, Nov 2019. 2

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[29] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In

Proceedings of the European Conference on Computer Vision (ECCV), pages 517–532, 2018. 2

[30] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolutaion. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 2, 5

[31] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 5, 6

[32] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *ICCV Workshops*, 2019. 2, 3, 7

[33] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. Aim 2019 challenge on real-world image super-resolution: Methods and results. In *ICCV Workshops*, 2019. 2, 3

[34] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. Ntire 2020 challenge on real-world image super-resolution: Methods and results. *CVPR Workshops*, 2020. 2, 3, 5, 6, 7

[35] T. Michaeli and M. Irani. Nonparametric blind super-resolution. In *2013 IEEE International Conference on Computer Vision*, pages 945–952, 2013. 3

[36] Seong-Jin Park, Hyeongseok Son, Sunghyun Cho, Ki-Sang Hong, and Seungyong Lee. Srfeat: Single image super-resolution with feature discrimination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 439–455, 2018. 2

[37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2, 4

[38] R. R. Schultz and R. L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6):996–1011, June 1996. 2

[39] Wen-Ze Shao and Michael Elad. Simple, accurate, and robust nonparametric blind super-resolution. In Yu-Jin Zhang, editor, *Image and Graphics*, pages 333–348, Cham, 2015. Springer International Publishing. 3

[40] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, June 2016. 2

[41] Y. Shi, H. Zhong, Z. Yang, X. Yang, and L. Lin. Ddet: Dual-path dynamic enhancement network for real-world image super-resolution. *IEEE Signal Processing Letters*, 27:481–485, 2020. 3

[42] A. Shocher, N. Cohen, and M. Irani. Zero-shot super-resolution using deep internal learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 3

[43] Charles Spearman. The proof and measurement of association between two things. *In J. J. Jenkins D. G. Paterson (Eds.), Studies in individual differences: The search for intelligence, Appleton-Century-Crofts*, pages 45–58, 1961. 6

[44] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017. 2

[45] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 5

[46] Roger Y. Tsai and Thomas S. Huang. Multiframe image restoration and registration. In *Advances in computer vision and image processing*, pages 317–339, 1984. 2

[47] Xintao Wang, Ke Yu, Shixiang Wu, et al. Esrgan: Enhanced super-resolution generative adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing. 1, 2, 3, 7

[48] Xuan Xu and Xin Li. Scan: Spatial color attention networks for real single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 3

[49] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, Nov 2010. 2

[50] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 814–81409, June 2018. 2, 3

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[52] X. Zhang, Q. Chen, R. Ng, and V. Koltun. Zoom to learn, learn to zoom. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3757–3765, 2019. 2

[53] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 1, 2, 4, 7

[54] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *ICCV*, October 2019. 3

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 3