Unsupervised Image Super-Resolution using Cycle-in-Cycle Generative Adversarial Networks

Yuan Yuan[12*] Siyuan Liu[134] Jiawei Zhang[1] Yongbing Zhang[3] Chao Dong[1] Liang Lin[1]
[1]Sensetime Research
[2]Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University
[3]Graduate School at Shenzhen, Tsinghua University, Shenzhen
[4]Department of Automation, Tsinghua University, Beijing

Abstract

We consider the single image super-resolution problem in a more general case that the low-/high-resolution pairs and the down-sampling process are unavailable. Different from traditional super-resolution formulation, the low-resolution input is further degraded by noises and blurring. This complicated setting makes supervised learning and accurate kernel estimation impossible. To solve this problem, we resort to unsupervised learning without paired data, inspired by the recent successful image-to-image translation applications. With generative adversarial networks (GAN) as the basic component, we propose a Cycle-in-Cycle network structure to tackle the problem within three steps. First, the noisy and blurry input is mapped to a noise-free low-resolution space. Then the intermediate image is up-sampled with a pre-trained deep model. Finally, we fine-tune the two modules in an end-to-end manner to get the high-resolution output. Experiments on NTIRE2018 datasets demonstrate that the proposed unsupervised method achieves comparable results as the state-of-the-art supervised models.

1. Introduction

Recent deep learning based super-resolution (SR) methods have achieved significant improvement either on PSNR values [8, 12, 13, 16, 17, 25, 28, 30] or on visual quality [16, 20]. These methods require supervised learning on high-resolution (HR) and low-resolution (LR) image pairs. However, their common assumption that the downscaling factor is known and the input image is noise-free hinders them from practical usages. In real-world scenarios, the SR problem often have the following properties: 1) HR datasets are unavailable, 2) downscaling method is unknown, 3) input LR images are noisy and blurry. This problem is extremely difficult if the input images suffer from different kinds of degradation. For an easier case, in this study, we assume that input images are degraded with the same processing which is complex and unavailable.

Under the above circumstances, models learned from synthetic data tend to generate similar results as traditional

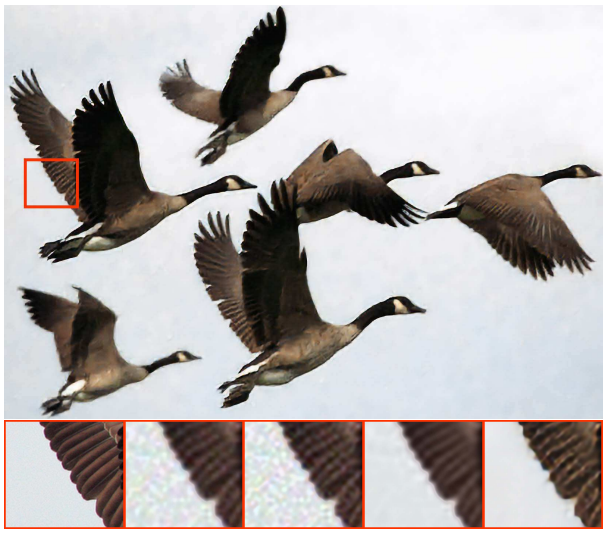


Figure 1. ×4 Super-resolution results of the proposed CinCGAN method for "0896" (DIV2K). For comparison, the sub-figures are cropped from results of existing algorithms. When the input is noisy, the results of bicubic interpolation and the EDSR [17] model both are in low quality, while CinCGAN learns to reconstruct clean result with fine details. The BM3D+EDSR method means using BM3D for denoising first and then using EDSR for super-resolution.

*Yuan Yuan and Siyuan Liu are co-first authors. This work was done when they were interns at Sensetime. Contacting email: yuanyuan@szu.edu.cn

methods [13, 30] or even simple interpolation. In Fig. 1, we show the results of bicubic interpolation and the state-of-the-art deep learning model—EDSR [17] with a noisy input. This is mainly due to the data bias between training and testing images. Detailed survey and analysis of deep learning based methods on real data can be found in [15].

As an alternative choice, blind SR [7, 19, 29] deal with the real-world data by estimating the down-sampling kernel from internal or external similar patches. However, when the input is noisy, the down-sampling kernel cannot be accurately estimated, and the inverse mapping results are accompanied by amplified noises. There are also works attempting at restoring LR images with addictive Gaussian noises [34]. But real-world noises may neither be addictive nor follow the standard Gaussian distribution, causing noise estimation infeasible. More generally, LR images may suffer from complex noises, blurry and non-uniform down-sampling kernels, which fail almost all existing blind SR methods.

Inspired by the development of unsupervised learning in image-to-image translation, such as CycleGAN [35] or WESPE [9], we intend to investigate unsupervised strategies to overcome this obstacle. In CycleGAN, images are translated between different domains with unpaired training data. They assume that the input image is of the same size as the output image, with only the difference on styles. However, in SR, output images are several times larger than the inputs, making the direct application of CycleGAN impossible. Further, using a bicubic-upsampled image as the input also could not obtain satisfactory results. SR problem is specific as it requires high quality output but not just a different style.

After exploring several training strategies, we find an effective Cycle-in-Cycle structure, named CinCGAN, which could achieve superior results. The whole pipeline consists of two CycleGANs, while the second GAN covers the first one (See Fig. 2). The first CycleGAN maps the LR image to the clean and bicubic-downsampled LR space. This module ensures that the LR input is fairly denoised/deblurred. We then stack another well-trained deep model with bicubic-downsampling assumption to up-sample the intermediate result to the desired size. Finally, we fine-tune the whole network using adversarial learning in an end-to-end manner. We conduct experiments on the **NTIRE2018 Super-Resolution Challenge**[1] dataset, and show that the proposed Cycle-in-Cycle structure is much stable at training and achieves competitive performance as supervised deep learning methods.

The contributions of this work are three-folds: 1) We study a more general super-resolution problem, where the high-resolution ground truth, down-sampling kernel and degradation function are unavailable. 2) We explore several unsupervised training strategies under the above assumption, and show that super-resolution task is different from conventional image-to-image translation. 3) We propose a Cycle-in-Cycle structure that could achieve comparable results as supervised CNN networks.

[1] https://competitions.codalab.org/competitions/18024

2. Related work

2.1. Image Super-Resolution

Single image super-resolution (SISR) has been widely studied for decades. Early approaches either rely on natural image statistics [33] [13] or pre-defined models [10] [5] [26]. Later, mapping functions between LR images and HR images are investigated, such as sparse coding based SR methods [30] [32].

Recently, deep convolution neural networks (CNN) have shown explosive popularity and powerful capability to improve the quality of SR results. Ever since Dong [3] first proposed using CNN for SR and achieved the state-of-the-art performance, plenty of CNN architectures have been studied for SISR. Inspired by the VGG [24] networks used for ImageNet classification, Kim et al. [12] present a very deep network (VDSR) that learns a residual image. For accelerating the speed of SR, FSRCNN [4] and ESPCN [23] extract feature maps at the low-resolution space and upsample the image at the last layer by transposed convolution and sub-pixel convolution, respectively. All the above mentioned CNN based SR methods aim at minimizing the mean-square error (MSE) between the reconstructed HR image and the ground truth. Based on the observation that minimizing MSE will make the SR results overly smooth, SRGAN [16] combines an adversarial loss [6] and a perceptual loss [24] [11] as the final objective function, and generates visually pleasing images which contain more high frequency details than the MSE-loss based methods. The champion of NTIRE2017 Super-Resolution Challenge [27], EDSR [17], employs deeper and wider networks to achieve the state-of-the-art performance by removing the unnecessary modules in SRResNet [16].

2.2. Blind Image Super-Resolution

Although a lot of works focus on SR problems with known degradation/downsamping kernels, little works try to solve blind SR—the degradation operation from HR images to LR images are unavailable. Estimating the degradation/blur kernel is an essential step for blind SR. Wang et al. [29] propose a probabilistic framework combined with the image co-occurrence prior to estimate the unknown point spread function (PSF) parameters. According to the property that small image patches will re-appear in natu-

ral images, Michaeli and Irani [19] present a method that is able to estimate the optimal blur kernel. Another relevant work [21] introduces a convolution consistency constraint and *bi*-l_0-l_2-norm regularization [22] to guide the blur kernel estimation process, achieving state-of-the-art blind SR performance.

In this work, we investigate how deep learning can be beneficial for addressing blind SR problems.

2.3. Unsupervised Learning

Existing supervised deep learning methods cannot handle blind SR without LR-HR image pairs. In real-world scenarios, where paired data is unavailable, it is essential to find a way to realize unsupervised learning. Recent work on GAN [6] provides a feasible solution, which includes a generator and a discriminator. The generator tries to generate fake images to fool the discriminator, while the discriminator aims at distinguishing the generated results from real data. GAN is widely used to solve the unsupervised learning problems. DualGAN [31] and CycleGAN [35] are two works about image-to-image translation using unsupervised learning, and both of them present an interesting network structure that contains a pair of forward and inverse generators. The forward generator maps domain X to domain Y, while the inverse generator maps the output back to domain X to maintain cycle consistency. Ignatov et al. [9] use the similar architecture to design a weakly supervised photo enhancer (WESPE) that translates ordinary photos to DSLR-quality images.

Different from the proposed method, both DualGAN [31] and CycleGAN [35] deal with input and output images of the same size, while SR requires the output images several times larger than the inputs. Utilizing the property of cycle consistency, we present a Cycle-in-Cycle GAN (CinCGAN) to super-resolve the LR images of which the degradation operators are unknown. Our method achieves a comparable performance with the state-of-the-art *supervised* CNN based algorithms [4, 16, 17].

3. Proposed Method

Problem formulation The conventional formulation of SISR [30] is $x = SHz + n$, where x and z denote LR and HR image respectively, SH represents the down-sampling and blurring matrix, and n is the addictive noise. Blind SR [19,29] follow the same assumption, only with unknown SH. In this work, we study a more general formulation as $x = f_n(f_d(z)) + n$, where f_d is the down-sampling process, f_n is a degradation function that may introduce complex noises, shift and blur. Here, we assume that f_d, f_n and the paired HR-LR training data are unavailable. Nevertheless, we can obtain a set of LR images that can be used for analysis and unsupervised training.

Motivation 1) Why applying unsupervised training? As the down-sampling and degradation functions are complex and coupled, it is hard to perform accurate estimation like traditional blind SR methods [19, 29]. The unavailability of HR images in practise also makes supervised training with simulated paired data impractical. This drives us to explore unsupervised learning strategies. 2) What is the difference between SR and image-to-image translation? SR accepts an LR image and outputs a HR image with much larger resolution. Further, SR requires the output to be of high quality, not just a different style. If we directly apply the image-to-image translation methods, we need to up-sample the LR image first by interpolation, which will also enlarge the noisy patterns. Directly applying existing methods like CycleGAN cannot remove such amplified noises, and training becomes very unstable. Experiments (in Sec. 4.4) also show that when the degradation function varies from image to image, it is difficult to deal with all kinds of images in a single forward pass.

Solution pipeline Our solution pipeline consists of three steps. First, we learn a mapping from an LR image set X to a "clean" LR image set Y, where images are noise-free and down-sampled from HR images Z with bicubic kernel. In other words, we deblur and denoise the input images at low resolution. Second, we adopt an existing SR model to super-resolve the intermediate results to the desired resolution. In the end, we combine and fine-tune these two models simultaneously to get the final HR images.

Under the guidance of the above pipeline, we propose a Cycle-in-Cycle structure named CinCGAN as shown in Fig. 2. To be specific, we adopt two coupled CycleGANs to learn the mapping from X to Y and Y to Z, respectively. Unpaired images $x_i \in X$, $y_j \in Y$ and $z_j \in Z$ are used for training[2], where y_j is down-sampled from z_j with bicubic kernel. Details are given in the following.

3.1. LR Image Restoration

The framework of the first CycleGAN that maps an LR image x to a clean LR image y is shown as *LR→clean LR* in Fig. 2. Given an input image x, the generator G_1 learns to generate an image $\tilde{y}$ that looks similar to the clean LR y, so as to fool the discriminator D_1. Meanwhile, D_1 learns to distinguish the generated sample $G_1(x)$ from the real sample y. To stabilize the training procedure, we use the least square loss [18] instead of the negative log-likelihood used in [6]. The generator-adversarial loss is:

$$\mathcal{L}_{GAN}^{LR} = \frac{1}{N}\sum_{i}^{N} ||D_1(G_1(x_i)) - 1||_2, \quad (1)$$

[2]For simplicity, we omit the subscript i and j in the following.

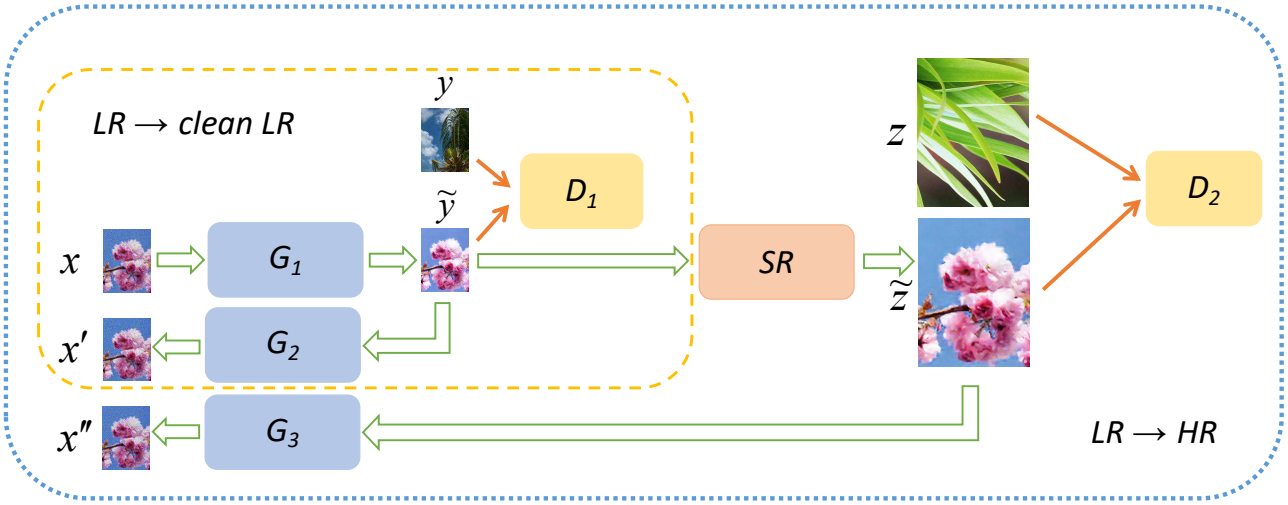


Figure 2. The framework of the proposed CinCGAN, where G_1, G_2 and G_3 are generators and *SR* is a super-resolution network. D_1 and D_2 are discriminators. The G_1, G_2 and D_1 compose the first *LR→clean LR* CycleGAN model, mapping the degrade LR images to clean LR images. The G_1, *SR*, G_3 and D_2 compose the second *LR→HR* CycleGAN model, mapping the LR images to HR images.

where N is the number of training samples. To maintain consistency between input x and output y, we add a network G_2 and let $x' = G_2(G_1(x))$ be identical to the input x. Hence, we also use a cycle consistency loss as:

$$\mathcal{L}_{cyc}^{LR} = \frac{1}{N}\sum_{i}^{N} ||G_2(G_1(x_i)) - x_i||_2. \tag{2}$$

In the previous work [35], the authors introduce an identity loss to preserve color composition between input and output images when they work on painting generation. They claim that the identity loss can help preserve the color of input images. In image SR, we also need to avoid color variation among different iterations, thus we add an identity loss

$$\mathcal{L}_{idt}^{LR} = \frac{1}{N}\sum_{i}^{N} ||G_1(y_i) - y_i||_1. \tag{3}$$

In addition, we add a total variation (TV) loss to impose spatial smoothness

$$\mathcal{L}_{TV}^{LR} = \frac{1}{N}\sum_{i}^{N}(||\nabla_h G_1(x_i)||_2 + ||\nabla_w G_1(x_i)||_2), \tag{4}$$

where ∇_h and ∇_w are functions to compute the horizontal and vertical gradient of $G_1(x_i)$.

In summary, the final objective loss for the *LR→clean LR* model is a weighted sum of the four losses:

$$\mathcal{L}_{total}^{LR} = \mathcal{L}_{GAN}^{LR} + w_1\mathcal{L}_{cyc}^{LR} + w_2\mathcal{L}_{idt}^{LR} + w_3\mathcal{L}_{TV}^{LR} \tag{5}$$

where w_1, w_2, w_3 are the weights of different losses.

3.2. Jointly Restoration and Super-Resolution

We then investigate how to super-resolve the intermediate image $\tilde{y}$ to the desired size. Recently, the enhanced deep residual network – EDSR [17] has won the first prize in the NTIRE 2017 challenge on single image super-resolution [1]. For simplicity, we directly adopt EDSR as the SR network stacked after G_1. Similarly, we use a discriminator D_2 for adversarial training both G_1 and SR networks. We also utilize another generator G_3 to ensure cycle consistency between x and the reconstructed x''. The GAN loss, cycle loss and TV loss for the *LR→HR* network are formulated as follows:

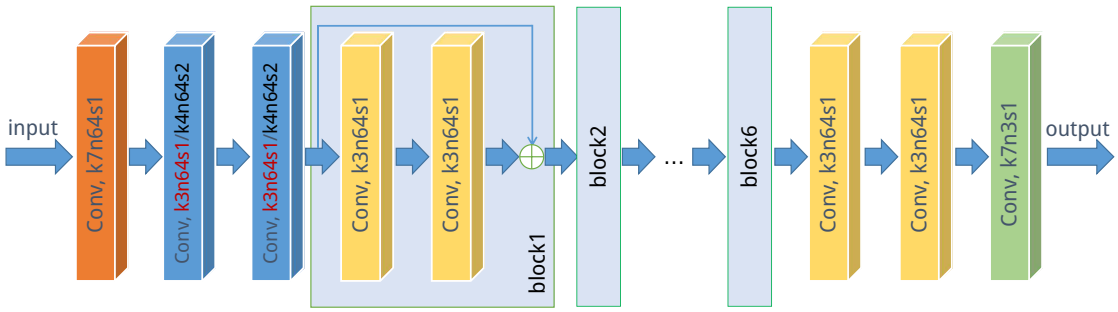
$$\mathcal{L}_{GAN}^{HR} = \frac{1}{N}\sum_{i}^{N} ||D_2(SR(G_1(x_i))) - 1||_2, \tag{6}$$

$$\mathcal{L}_{cyc}^{HR} = \frac{1}{N}\sum_{i}^{N} ||G_3(SR(G_1(x_i))) - x_i||_2, \tag{7}$$

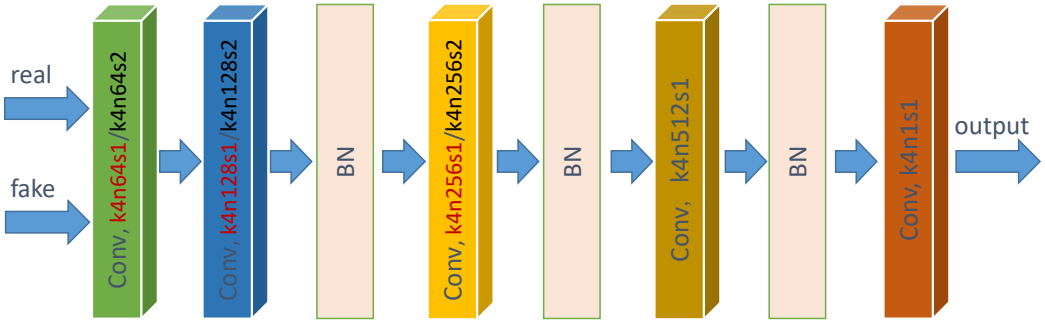
$$\mathcal{L}_{TV}^{HR} = \frac{1}{N}\sum_{i}^{N}(||\nabla_h SR(G_1(x_i))||_2 + ||\nabla_w SR(G_1(x_i))||_2). \tag{8}$$

For the identity loss, instead of maintaining the tint consistency between input and output, we consider ensuring the SR network can generate adequate quality of super-resolved images. We define a new identity loss as:

$$\mathcal{L}_{idt}^{HR} = \sum_{i} ||SR(z') - z||_2. \tag{9}$$



(a) Generator



(b) Discriminator

Figure 3. The generators G_1, G_2 and G_3 share the same framework as (a) and the discriminators D_1 and D_2 share the same framework as (b). For the 2-nd and 3-rd convolution layers in generator (a), **k3n64s1** is for G_1 and G_2, while **k4n64s2** is for G_3. For the first three convolution layers in discriminator (b), **k4n64s1**, **k4n128s1**, and **k4n256s1** are for D_1 and **k4n64s2**, **k4n128s2**, and **k4n256s2** are for D_2. Please see text for details.

where z' is down-sampled from z with bicubic kernel. This $\mathcal{L}_{idt}^{HR}$ makes the SR network does not betray its original ambition, such that the produced $\tilde{z}$ can be reasonable SR results.

To sum up, the total loss for fine-tuning the LR to HR networks is

$$\mathcal{L}_{total}^{HR} = \mathcal{L}_{GAN}^{HR} + \lambda_1 \mathcal{L}_{cyc}^{HR} + \lambda_2 \mathcal{L}_{idt}^{HR} + \lambda_3 \mathcal{L}_{TV}^{HR} \quad (10)$$

where $\lambda_1, \lambda_2, \lambda_3$, for $i = 1, 2, 3$, are weights of each loss.

3.3. Network Architecture

The architecture of generators G_1, G_2, G_3 and discriminators D_1, D_2 are shown in Fig. 3. We adapt similar architecture as the work of Zhu et al. [35], which has shown impressive results for unpaired image-to-image translation. Here, "conv" means convolution layer, where a Leaky ReLU layer with negative slope 0.2 is added right after except for the last convolution layer (we omit it for simplicity). "BN" means a batch normalization layer. The number after symbols k, n and s represents kernel size, number of filters and stride size, respectively. For example, **k3n64s1** refers to the convolution layer that contains 64 filters, of which the spatial size is 3 and stride is 1.

For the generators G_1 and G_2, we use 3 convolution layers at the head and tail, and 6 residual blocks in the middle. The generator G_3 shares the same architecture as G_1 and G_2, except for the 2-nd and 3-rd convolution layers, where the stride is set to 2 to perform down-sampling. As to the discriminator, we use a 70×70 PatchGAN for D_2. Since we up-sample LR images with a scale of $\times 4$, the size of input images is usually less than 70 (we use 32×32 LR images and 128×128 HR images for training). Hence, we modify the stride of the first three convolution layers as 1 for discriminator D_1, such that the respective field of D_1 is reduced to 16×16.

4. Experiments

In this section, we first introduce the dataset and details we used for training. We then evaluate the performance of the proposed CinCGAN model by comparing with several state-of-the-art SISR methods. Finally, we perform ablation study to validate the advantages of CinCGAN.

4.1. Training data

We take the track 2 dataset from the NTIRE2018 Super-Resolution Challenge for training. The challenge aims to restore a HR image given a degraded LR image. They provide a high-quality image dataset, DIV2K [1], which contains 800 training images and 100 validation images. The DIV2K dataset contains almost all kinds of natural scenarios: buildings (indoor and outdoor), forest, lakes, animals, people, etc. The track 2 dataset is degraded from DIV2K dataset, with down-sampling, blurring, pixel shifting and noises. Although the parameters of the degradation operators are fixed for all images, the blur kernels are randomly generated and their resulting pixel shifts vary from image to image. Hence, the degradation kernels of images in the track 2 dataset are unknown and diverse.

Since our purpose is to unsupervised train a network without paired LR-HR data, we take the first 400 images (numbered from 1 to 400) from the training LR set as input images X, and the other 400 images (numbered from 401 to 800) from the HR set as demanding HR images Z. The intermediate clean LR images Y are directly bicubic down-sampled from Z. Similar to [4] [24], we augment data with 90 degree rotation and flipping. Our experiments are performed with a scaling factor of $\times 4$. We randomly crop X and Y with size 32×32 and crop Z with size 128×128. We conduct testing on the provided 100 validation images. Note that, although DIV2K contains paired training dataset, we do not use paired data for supervised training.

4.2. Training details

We divide our training process into two steps. We first train the model G_1, G_2 and D_1 for mapping LR images to clean LR images (shown as *LR→clean LR* in Fig. 2). The three parameters in (5) are set to be $w_1 = 10, w_2 = 5$ and $w_3 = 0.5$, respectively. We train our model with Adam optimizer [14] by setting $\beta_1 = 0.5$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$, without weight decay. Learning rate is initialized as 2×10^{-4} and then decreased by a factor of 2 every 40000 iterations. The weights of filters in each layer are initialized using a normal distribution and the batch size is set as 16. We train the model over 400000 iterations, until it converges.

We then jointly fine-tune the LR to HR model (shown as *LR→HR* in Fig. 2). We initialize our *SR* network by publicly available EDSR model[3]. We set parameters in (10) as $\lambda_1 = 10, \lambda_2 = 5$ and $\lambda_3 = 2$. The optimizer is set almost the same as training the *LR→clean LR* model, except for we initialize learning rate with 10^{-4}. As to the weight of identity loss $\mathcal{L}_{idt}^{LR}$ in (5), we set $w_2 = 1$. At each iteration, we update (5) and (10) in turn. We first train G_1 and G_2 to update the *LR→clean LR* network. We then train G_1, SR and G_3 simultaneously to update the *LR→HR* network.

We implement the proposed networks with PyTorch and train them on a Nvidia Tesla K80 GPU. It takes about 1 day to pre-train the *LR→clean LR* model and about 2 days to jointly fine-tune the *LR→HR* model.

[3]https://github.com/thstkdgus35/EDSR-PyTorch

4.3. Results

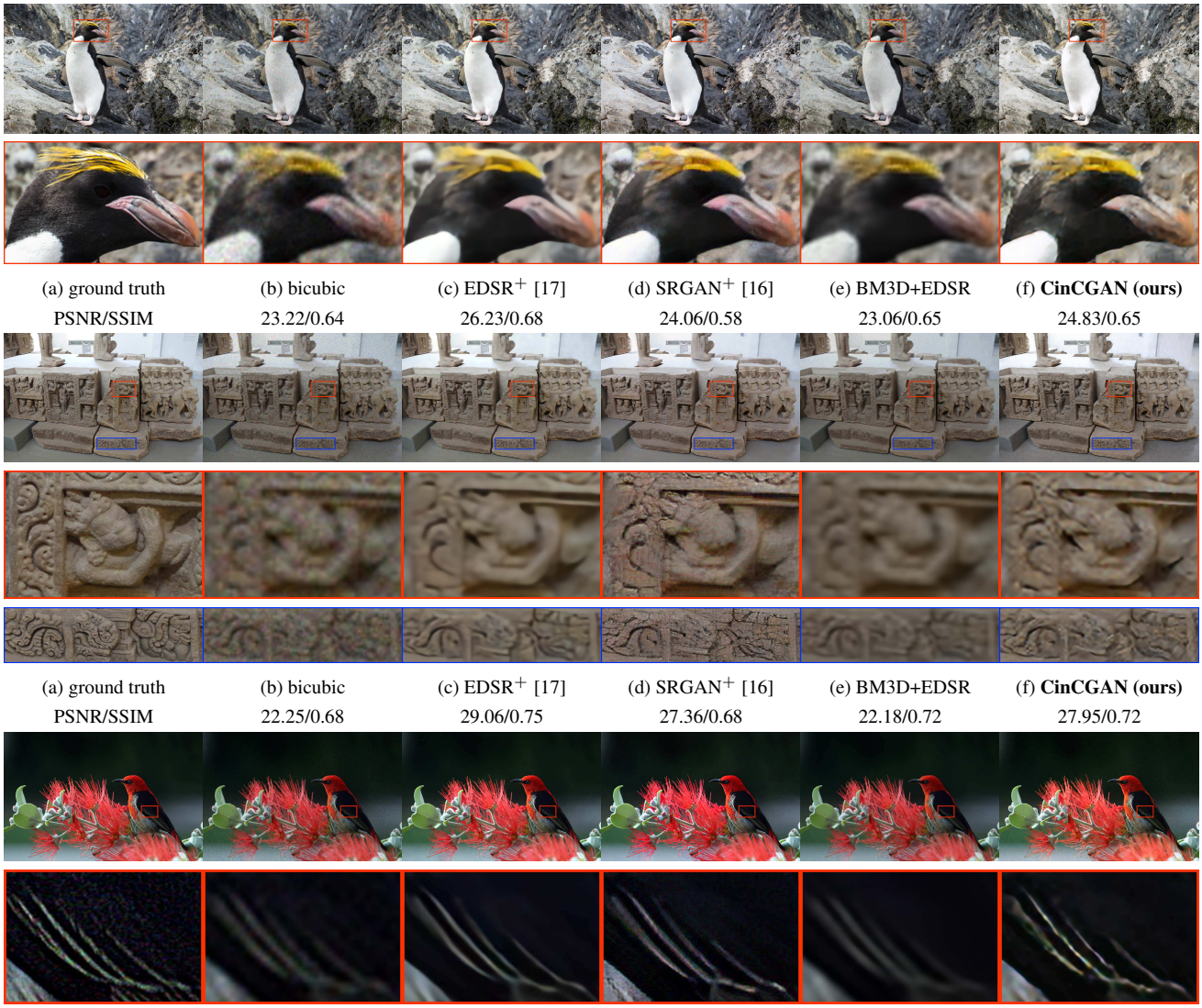
We compare the performance of the proposed CinCGAN model with several state-of-the-art SISR methods: FSRCNN [4], EDSR [17] and SRGAN [16]. We use the publicly available FSRCNN and EDSR models which are trained with paired LR and HR images, where the inputs are clean LR images down-sampled from HR images. To make the results more comparable, we also fine-tune EDSR and SRGAN (labelled as EDSR$^+$ and SRGAN$^+$ respectively) with the paired track 2 dataset. To emphasize the effectiveness of CinCGAN structure, we also try to first denoise the input LR images and then super-resolve the denoised images for comparison. BM3D [2] is one of the state-of-the-art image denoising approach, which is an efficient and powerful denoiser. Hence, we pre-process the test LR images with BM3D first, and then super-resolve it using EDSR (labelled as BM3D+EDSR).

Table 1 shows the average PSNR and SSIM values of the restored test images. It shows that FSRCNN and EDSR cannot work well if the blur and noises are unknown in the training process. After fine-tuning by paired track 2 dataset, EDSR$^+$ and SRGAN$^+$ improve their results and our method can work comparably against SRGAN$^+$ in terms of PSNR and SSIM without paired training data. Although BM3D can remove noise, it also over-smooth the input images. The PSNR and SSIM values of BM3D+EDSR are lower than the proposed method. Several subjective results are illustrated in Fig. 4.

4.4. Ablation Study

To validate the advantages of the proposed CinCGAN model for the unsupervised SISR problem, we design some other network structures for comparison.

Structure 1 The first frame structure is to restore LR images X to HR images Z using only one CycleGAN, *i.e.* denoise, deblur and super-resolve the LR images at the same time. The structure of the model is shown in Fig. 5(a), where we set an LR image x as input to the *SR* network directly. Correspondingly, we only minimize the total loss $\mathcal{L}_{total}^{HR}$ (with replacing $SR(G_1(\cdot))$ as $SR(\cdot)$ in Eq. (6)(7)(8)). However, during the training procedure, we found that the result $\tilde{z}$ are always unstable and there are a lot of undesired artifacts, as shown in Fig. 6(a). It is hard for a single network to simultaneously denoise, deblur and up-sample the degraded



(a) ground truth	(b) bicubic	(c) EDSR$^+$ [17]	(d) SRGAN$^+$ [16]	(e) BM3D+EDSR	(f) **CinCGAN (ours)**
PSNR/SSIM	23.22/0.64	26.23/0.68	24.06/0.58	23.06/0.65	24.83/0.65
(a) ground truth	(b) bicubic	(c) EDSR$^+$ [17]	(d) SRGAN$^+$ [16]	(e) BM3D+EDSR	(f) **CinCGAN (ours)**
PSNR/SSIM	22.25/0.68	29.06/0.75	27.36/0.68	22.18/0.72	27.95/0.72
(a) ground truth	(b) bicubic	(c) EDSR$^+$	(d) SRGAN$^+$ [16]	(e) BM3D+EDSR	(f) **CinCGAN (ours)**
PSNR/SSIM	26.81/0.83	30.28/0.88	29.05/0.85	26.84/0.86	28.26/0.84

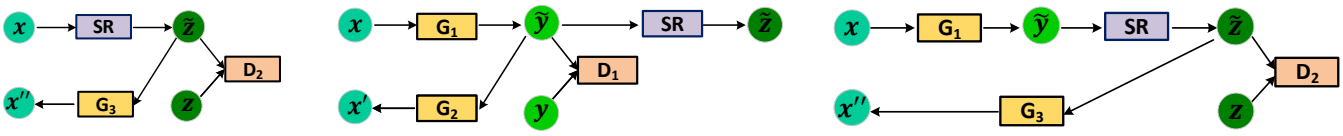
Figure 4. Super-resolution results of "0801", "0816" and "0853" (DIV2K) with scale factor $\times 4$. EDSR$^+$ and SRGAN$^+$ are trained on paired NTIRE2018 track 2 dataset. BM3D+EDSR means using BM3D for denoising first and then using EDSR for super-resolution. The proposed CinCGAN model shows comparable results with SRGAN$^+$ and is better than BM3D+EDSR method.

Table 1. Quantitative evaluation on NTIRE 2018 track 2 dataset of the proposed CinCGAN model, in terms of PSNR and SSIM.

method	bicubic	FSRCNN [4]	EDSR [17]	EDSR$^+$	SRGAN$^+$ [16]	BM3D+EDSR	**CinCGAN (ours)**
PSNR	22.85	22.79	22.67	25.77	24.33	22.88	24.33
SSIM	0.65	0.61	0.62	0.71	0.67	0.68	0.69

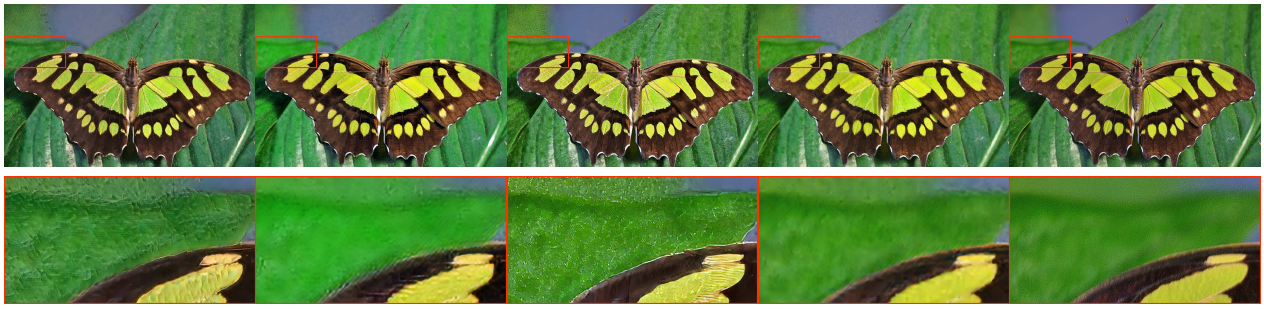
images, especially when the degradation kernels are different from image to image and with unsupervised learning.

Structure 2 We remove D_2 and G_3 from the proposed CinCGAN model for our second experiment. We map the input LR images to a set of clean LR images using the same *LR→clean LR* networks shown in Fig. 2; we then super-resolve the converted LR images directly using the SR network. The whole structure is shown in Fig. 5(b). The corresponding result is illustrated in Fig. 6(b). As we can see, some negligible noise in the resulted clean LR images is



(a) Structure 1 (b) Structure 2 (c) Structure 3

Figure 5. Experiments for validating the advantages of the proposed structure. (a) Structure 1: transform the LR images x to HR images z directly with one CycleGAN model; (b) Structure 2: remove D_2 and G_3 from the proposed CinCGAN model; (c) Structure 3: remove D_1 and G_2 from the proposed CinCGAN model.



(a) Structure 1 (b) Structure 2 (c) Structure 3 (d) **CinCGAN (ours)** (e) ground truth

Figure 6. Super-resolution results of "0829" (DIV2K) with scale factor $\times 4$, for each frame structure as described in Fig. 5.

magnified and now is visible in the super-resolved images, which affects the visual quality.

Structure 3 Our third experiment is performed by removing D_1 and G_2 from the proposed CinCGAN model, as shown in Fig. 5(c). We use one CycleGAN for the LR to HR model, where we take $G_1 + SR$ as the forward network and G_3 as the inverse network. D_2 is used for distinguishing $\tilde{z}$ from z. We load the pre-trained G_1 (in the *LR→clean LR* networks) and the downloaded EDSR models for initialization. Experimental results on Fig. 6(c) show that the resulting $\tilde{z}$ are still noisy. Since without the $\mathcal{L}_{cyc}^{LR}$ and $\mathcal{L}_{GAN}^{LR}$ constraints on G_1 network ($\mathcal{L}_{idt}^{LR}$ and $\mathcal{L}_{tv}^{LR}$ are still used for this model), G_1 is unable to deonise and deblur. The whole model becomes similar to Structure 1.

Proposed Method We then propose our final solution as shown in Fig. 2: jointly fine-tune LR to HR networks with CinCGAN. We sequentially update the *LR→ clean LR* and the *LR→HR* models. With the two constraint $\mathcal{L}_{total}^{LR}$ and $\mathcal{L}_{total}^{HR}$, the G_1 network can denoise and deblur the degraded input image x, while the *SR* network can up-sample as well as further restore the resulted intermediate image $\tilde{y}$. The final resulted SR image is shown in Fig. 6(d), which shows the best visual result comparing with other three structures.

5. Conclusions

We investigate the single image super-resolution problem with a more general assumption: the low-/high-resolution image pairs and the down-sampling process are unavailable. Inspired by the recent successful image-to-image translation applications, we resort to the unsupervised learning methods to solve this problem. Using generative adversarial networks (GAN), the proposed method contains two CycleGANs, where the second GAN covers the first one. The solution pipeline consists of three steps. First, we map the input LR images to the clean and bicubic-downsampled LR space with the first CycleGAN. We then stack another well-trained deep model with bicubic-downsampling assumption to up-sample the intermediate result to the desired size. Finally, we fine-tune the two modules in an end-to-end manner to get the high-resolution out. Experimental results demonstrate that the proposed unsupervised method achieves comparable results as the state-of-the-art supervised models.

Acknowledgement. This work is supported by SenseTime Group Limited and in part by the Projects of National Science Foundations of China (61571254), Guangdong Special Support plan (2015TQ01X16), and Shenzhen Fundamental Research fund (JCYJ20160513103916577).

References

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

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Bm3d image denoising with shape-adaptive principal component analysis. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[4] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407. Springer, 2016.

[5] R. Fattal. Image upsampling via imposed edge statistics. In *ACM transactions on graphics (TOG)*, volume 26, page 95. ACM, 2007.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] Y. He, K.-H. Yap, L. Chen, and L.-P. Chau. A soft map framework for blind super-resolution image reconstruction. *Image and Vision Computing*, 27(4):364–373, 2009.

[8] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015.

[9] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. Wespe: Weakly supervised photo enhancer for digital cameras. *arXiv preprint arXiv:1709.01118*, 2017.

[10] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991.

[11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[12] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

[13] K. I. Kim and Y. Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] T. Köhler, M. Bätz, F. Naderi, A. Kaup, A. K. Maier, and C. Riess. Benchmarking super-resolution algorithms on real data. *arXiv preprint arXiv:1709.04881*, 2017.

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2016.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, volume 1, page 3, 2017.

[18] X. Mao, Q. Li, H. Xie, R. Y. Lau, and Z. Wang. Multi-class generative adversarial networks with the l2 loss function. *CoRR, abs/1611.04076*, 2, 2016.

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

[20] M. S. Sajjadi, B. Schölkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4501–4510. IEEE, 2017.

[21] W.-Z. Shao and M. Elad. Simple, accurate, and robust non-parametric blind super-resolution. In *International Conference on Image and Graphics*, pages 333–348. Springer, 2015.

[22] W.-Z. Shao, H.-B. Li, and M. Elad. Bi-l0-l2-norm regularization for blind motion deblurring. *Journal of Visual Communication and Image Representation*, 33:42–59, 2015.

[23] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] A. Singh and N. Ahuja. Super-resolution using sub-band self-similarity. In *Asian Conference on Computer Vision*, pages 552–568. Springer, 2014.

[26] J. Sun, Z. Xu, and H.-Y. Shum. Image super-resolution using gradient profile prior. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[27] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1110–1121. IEEE, 2017.

[28] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126. Springer, 2014.

[29] Q. Wang, X. Tang, and H. Shum. Patch based blind image super resolution. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 709–716. IEEE, 2005.

[30] 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, 2010.

[31] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017.

[32] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[33] H. Zhang, J. Yang, Y. Zhang, and T. S. Huang. Non-local kernel regression for image and video restoration. In *European Conference on Computer Vision*, pages 566–579. Springer, 2010.

[34] K. Zhang, W. Zuo, and L. Zhang. Learning a single convolutional super-resolution network for multiple degradations. *arXiv preprint arXiv:1712.06116*, 2017.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.