Unsupervised Image Super-Resolution with an Indirect Supervised Path

Shuaijun Chen[1†], Zhen Han[1,2†‡], Enyan Dai[1†‡], Xu Jia[1*], Ziluan Liu[3], Xing Liu[3], Xueyi Zou[1],
Chunjing Xu[1], Jianzhuang Liu[1], Qi Tian[1]
[1]Huawei Noah's Ark Lab [2]Renmin University of China
[3]Huawei Consumer Business Group
{chenshuaijun, x.jia, liuziluan, liuxing67, zouxueyi, xuchunjing, tian.qi1}@huawei.com
dey123@mail.ustc.edu.cn, 2017000828@ruc.edu.cn

Abstract

The task of single image super-resolution (SISR) aims at reconstructing a high-resolution (HR) image from a low-resolution (LR) image. Although significant progress has been made with deep learning models, they are trained on synthetic paired data in a supervised way and do not perform well on real cases. There are several attempts that directly apply unsupervised image translation models to address such a problem. However, unsupervised image translation models need to be modified to adapt to unsupervised low-level vision task which poses higher requirement on the accuracy of translation. In this work, we propose a novel framework which is composed of two stages: 1) unsupervised image translation between real LR and synthetic LR images; 2) supervised super-resolution from approximated real LR images to the paired HR images. It takes the synthetic LR images as a bridge and creates an indirect supervised path. We show that our framework is so flexible that any unsupervised translation model and deep learning based super-resolution model can be integrated into it. Besides, a collaborative training strategy is proposed to encourage the two stages collaborate with each other for better degradation learning and super-resolution performance. The proposed method achieves very good performance on datasets of NTIRE 2017, NTIRE 2018 and NTIRE 2020, even comparable with supervised methods.

1. Introduction

The task of single image super-resolution (SISR) aims at reconstructing a high-resolution (HR) image from a low-resolution (LR) image. It has broad application in tasks such as image enhancement, surveillance and medical imaging. In the SISR task, people usually assume that an LR image y is modeled as the degradation output by applying the following degradation process to an HR image x,

$$y = (x \otimes k) \downarrow_s + n, \tag{1}$$

where k denotes a blur kernel, $\downarrow_s$ denotes a downsampling operation with scaling factor s, and n denotes noise and is usually modeled as Gaussian noise with standard deviation σ. It is an ill-posed problem since there are multiple solutions that can be reconstructed from a given LR image.

Recently, data-driven methods, especially deep learning based methods [4, 25, 13, 14, 26, 27, 16, 35, 34] achieve great performance on low-level vision tasks. Trained with sampled LR-HR pairs, these learning-based methods directly learn the mapping from distribution of LR to that of HR. However, they simply take synthetically downsampled images from HR domain as the corresponding LR images. Those models are trained with clean HR and its corresponding bicubic/bilinear downsampled LR images. Thus these methods have poor generalization ability on unseen, realistic low-resolution images which suffer from other degradation

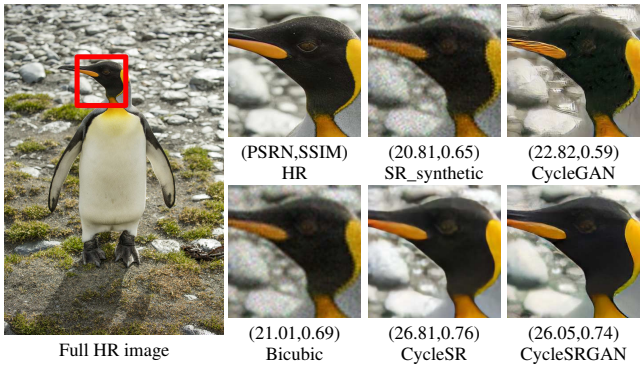


Figure 1. SR results of supervised and unsupervised methods on natural images. Ours CycleSR and CycleSRGAN perform well on realistic LR images.

[†]Equal contributions on this work
[‡]Work done during an internship at Huawei Noah's Ark Lab
[*]Corresponding author

factors such as blur and noise.

The large gap between LR domain at training and testing stage, mainly caused by the lack of HR and real LR pairs, will make such supervised-learning-based methods fail at the testing stage, as shown in Figure 1. Very recently, to reduce this gap, some researchers try to improve quality of LR images at training stage by collecting a dataset consisting of real-world LR-HR pairs [3, 33]. In this way, models trained with these real LR and HR pairs will perform well on the real LR images collected under the same setting. However, it is expensive to collect such LR-HR pairs because of complicated data collection and post-processing procedure, which makes such methods difficult to scale. Other researchers [31, 2] make attempt to reduce this gap with unsupervised learning, where cycle-consistency and adversarial losses are utilized to deal with unpaired HR and real LR images. However, it still lacks strong supervision on super-resolution results.

In this work we aim at reducing the large gap between LR domain at training and testing stage with unsupervised learning. To address the limitation with GAN-based unsupervised methods, we propose to learn the degradation process in order to generate pairs of HR and the corresponding real-LR-like images, on which we can further enjoy the advantages of supervised training. The key issue is how to minimize the domain gap between HR and real LR, and how to use approximate pairs for training an image super-resolution model.

We take synthetic LR images as an intermediate bridge, and learn the degradation from synthetic LR to real LR images instead of directly learning that from HR to real LR images. In other words, we model the degradation process from HR to real LR domain in two steps, *i.e*., synthetic degradation from HR to synthetic LR domain and mapping from synthetic LR to real LR domain. Furthermore, extra methods can be employed after downsampling, such as adding noise, to further minimize the domain gap between synthetic domain and real LR domain. This will reduce the difficulty of following translation task. For the second step, we adopt an image-to-image translation model with cycle-consistency and take its degradation direction branch to get real-LR-like images. An SR module is equipped after image translation model to super-resolve a real-LR-like image to an HR image. Hence, with an image translation model and an SR module together, we are able to train a model that super-resolves real LR images to HR images with an indirect supervised path.

Our contributions can be summarized as follows:

- We propose a novel framework designed for unsupervised SR learning, which first models the real degradation as a combination of synthetic degradation and image translation, and then train an SR module with pairs of HR and the corresponding degraded LR images.

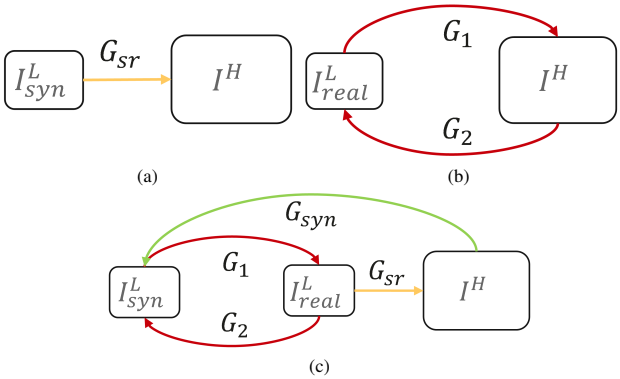


Figure 2. Image super-resolution training pipelines. (a) Typical training pipeline on synthetic paired data, (b) Typical training pipeline on unpaired data, (c) our training pipeline on unpaired data.

- The proposed framework is flexible enough to integrate any image translation model and SR model, and to be trained with either fidelity loss or perceptual quality oriented loss.

- A joint training strategy is proposed to train our framework in an end-to-end manner, which benefits both image degradation and super-resolution process.

2. Related Work

Supervised deep learning based image super-resolution. Recently, a lot of works are proposed to address the task of SISR based on Convolutional Neural Networks (CNNs). The first one is proposed by Dong *et al*. [4], which implicitly learns a mapping between LR and HR images using a shallow fully-convolutional network.

In [13, 14], Kim *et al*. borrowed the idea of residual connection from ResNet [7] and designed a very deep network to improve SISR performance. Further works [16, 17, 19, 26, 35, 29] mainly focus on design of network architecture for performace improvement. In addition, there are several researchers [17, 24, 29] working on improving perceptual quality of SISR results by combining fidelity loss with an adversarial loss [6] and a perceptual loss [10]. However, they are all trained on HR and synthetic LR pairs as shown in Figure 2 (a). [3, 33] take a further step to capture LR-HR image pairs under realistic setting, that is, tuning focal length of DSLR cameras to collect images of different resolution. However, models trained with such data may not generalize well to LR images captured by other devices such as smartphones which may contain different level of noise and blur, and it is expensive to scale.

Unsupervised deep learning based image super-resolution. Given unsatisfactory performance of most supervised methods on real data, some recent works [22]

resort to unsupervised learning to address this issue. Image super-resolution can be considered as a special image translation task, *i.e*., translating images from LR domain to HR domain. [31] introduces two cycles, with one between real LR and synthetic LR images, and another one between real LR and HR images. This method works in a two-stage fashion at both training and test stage, *i.e*., they first clean the real LR images and then apply the super-resolution on it, while our method can directly apply the super-resolution on real LR images at test stage. In [36], the model is actually a variant of CycleGAN model for translation between HR and real LR domains, with an additional perceptual loss on real LR domain and ignoring the adversarial loss on HR domain. Lugmayr *et al*. [21] also proposed to combine an unsupervised image translation model and an image restoration model, but they are trained separately. Fritsche *et al*. [5] further improved this method by separating the low and high frequency components and treat them differently.

The most similar work is [2] which proposes to directly learn the degradation from HR images to real LR images through a single high-to-low network, and a low-to-high network trained with HR and the estimated degraded LR images. This work shares a similar idea with our method in the motivation to learn degradation from real LR images. However, in that work the degradation learning process is conducted by directly translating from HR space to real LR space which is supervised with only on real LR space. The model [2] is mainly supervised with an adversarial loss and a pixel-wise loss is used to speed up convergence. While in our work the degradtion learning is decomposed into two steps, *i.e.*, synthetic degradation and translation between synthetic space and real LR space. The synthetic degradation can alleviate difficulty of direct translation from origin HR space to real LR space. And the translation is supervised with both adversarial losses and cycle-consistency losses for robustness. Both [36] and [2] share similar pipelines as described in Figure 2 (b), and difference lies in the supervision on LR and HR domains. Our method can falls into another pipeline as shown in Figure 2 (c). In addition, with a collaborative training strategy, translation model and SR model in our framework are able to help each other to obtain both good degradation and super-resolution performance.

Unsupervised image-to-image translation. There have been several approaches to address unsupervised image translation. Zhu *et al*. [37] proposed CycleGAN by adding cycle consistency constraint on top of pix2pix [9]. Cycle consistency enforces each image to be correctly reconstructed after translating from a source domain to a target domain and translating back to the source domain. Similar approaches are also proposed in DiscoGAN [15] and [30]. Another kind of approaches [20] assume that images from source domain and target domain share a common latent space. Once an image is projected to the shared latent space, a decoder can be used to either reconstruct the image in source domain or produce an image in target domain. Huang *et al*. [8] and Lee *et al*. [18] further proposed to decompose an image into a content-related space and a style-related space to achieve many-to-many image translation such that an image can be generated by combining arbitrary content and style representations. Since our framework is flexible to include any unsupervised image translation model, any advance in unsupervised image translation will benefit the proposed framework for image super-resolution.

3. Method

3.1. Overview

Notation We denote HR images as I^H, real LR images as I^L_{real}, synthetic degraded LR images as I^L_{syn}. After translation, we use $\hat{I}^L_{real}$ and $\hat{I}^L_{syn}$ to respectively represent the approximated real LR and synthetic LR, and use $\hat{I}^H$ to represent recovered HR images.

As shown in Figure 3, the proposed framework is composed of two stages: 1) unsupervised image translation between real LR images and synthetic LR images; 2) supervised super-resolution from approximated real LR images to HR images. Given unpaired images I^H and I^L_{real}, we first apply bicubic downsampling to HR images I^H. Taking it as a bridge, we are able to generate LR images $\hat{I}^L_{real}$ with similar noise and blur in real LR space through an unsupervised image translation model. The HR images and the approximated real LR images $\hat{I}^L_{real}$ compose paired training data such that an SR model can be indirectly trained in an supervised way. We show that the SR model in the second stage can enjoy many advantages of supervised training such as various losses to balance distortion and perceptual quality. We also show that the proposed framework is flexible enough to include any unsupervised image translation model and SR model.

3.2. Unsupervised translation for image degradation

In order to train a SR model on unpaired data, we propose to learn the degradation process as the first stage. It can generate the corresponding real-LR-like images $\hat{I}^L_{real}$ for HR images I^H, which compose paired data to train a SR model. We decompose degradation process into two steps, *i.e.*, synthetic degradation and translation from synthetic space to real LR space. With synthetic LR images as a bridge, the gap between two domains is reduced such that image translation model can focus on simulating the degradation such as noise and blur.

Different from [2] which uses a single direction model to learn the degradation directly from HR to real LR domain, we propose to use a bi-directional image translation model

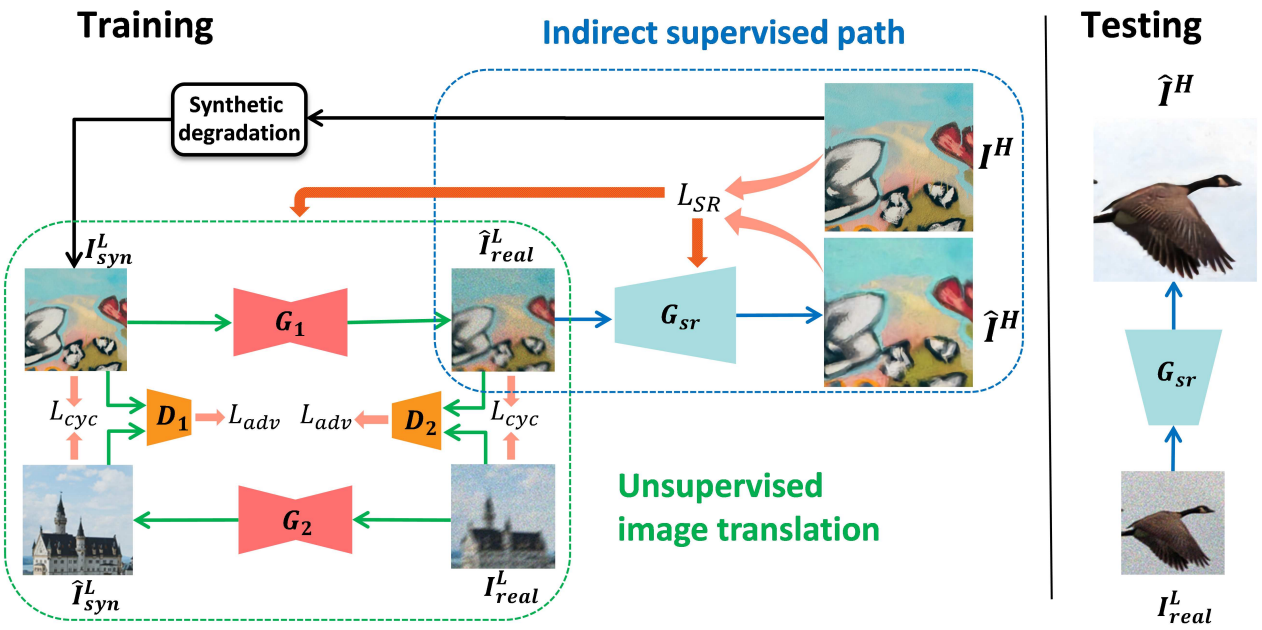


Figure 3. Pipeline of the proposed framework for unsupervised image super-resolution.

for better robustness. As shown in Figure 3, we use a generator G_1 to learn the mapping from synthetic LR domain to real LR domain, and another generator G_2 to learn the inverse mapping. In order to address the unpaired setting, adversarial losses are applied to encourage the translated images to follow the same distribution of images in the target domain. For example, the branch for the mapping from synthetic domain to real LR domain would translate clean LR images into images with similar noise and blur as real LR images. In addition, cycle consistency, *i.e.*, $G_2(G_1(I^L_{syn})) \approx I^L_{syn}$, is added to further constrain the model training such that the content in the generated LR images $\hat{I}^L_{real}$ can be maintained, and vice versa. The loss for training such unsupervised image translation model is defined as below.

$$\begin{aligned} &L_{Trans}(G_1, D_1, G_2, D_2, I^L_{real}, I^L_{syn}) \\ &= L_{adv}(G_1, D_1, G_2, D_2, I^L_{real}, I^L_{syn}) \\ &+ \lambda_{cyc} L_{cyc}(G_1, G_2, I^L_{real}, I^L_{syn})) \\ &+ \lambda_{other} L_{other}(G_1, G_2, I^L_{real}, I^L_{syn}), \end{aligned} \tag{2}$$

where L_{adv} is the adversarial losses on both real LR domain and synthetic LR domain, L_{cyc} is the cycle-consistency loss and L_{other} is other losses in an unsupervised image translation model. Our framework is flexible such that any unsupervised image translation model can be applied here. In this work, we experiment with two popular translation models, CycleGAN [37] and UNIT [20]. They share in common on the constraints of adversarial losses and cycle-consistency losses, while CycleGAN has an additional identity loss for content preserving and UNIT has an additional KL divergence loss for constraint on latent space. In Section 4, we show that both of them perform equally well on several datasets, which demonstrates the effectiveness of the proposed framework.

3.3. Indirectly supervised learning for super-resolution

Given unpaired I^H and I^L_{real}, there is no direct pairwise supervision to train an SR model G_{sr}. However, with real-like LR images $\hat{I}^L_{real}$ which are generated from HR images I^H in two steps, we are able to construct a supervised path from I^L_{real} to I^H, where the trained SR model is expected to work for real LR images $\hat{I}^L_{real}$. With such indirect supervision the model can enjoy many advantages with supervised training. Any existed SR model which is originally proposed for the supervised methods which are trained on synthetic pairs could be used here. In addition, our whole framework brings much flexiblity for the training of SR module. Different from most unsupervised super-resolution methods [31, 2] where adversarial losses must be used to train the SR module, the SR module in our framework can be trained with either only pixel-wise distortion oriented losses or that combined with perceptual quality oriented losses such as adversarial loss and perceptual loss [11]. The loss for this stage can be fomulated as follows,

$$\begin{aligned} &L_{SR}(G_{sr}, D_h, \hat{I}^L_{real}, I^H) \\ &= \lambda_{mse} L_{mse}(G_{sr}, \hat{I}^L_{real}, I^H) \\ &+ \lambda_{advsr} L_{adv}(G_{sr}, D_h, \hat{I}^L_{real}, I^H) \\ &+ \lambda_{percep} L_{percep}(G_{sr}, \hat{I}^L_{real}, I^H), \end{aligned} \tag{3}$$

where $\hat{I}^L_{real} = G_1(I^L_{syn})$ is the generated real-like LR images, L_{mse} is reconstruction loss for little distortion, L_{percep} and L_{adv} are respectively perceptual loss and adversarial loss on HR space for good perceptual quality. We follow the ESRGAN [29] and use losses in Relativistic GAN [12] as L_{adv}. The important role of distortion oriented loss L_{mse} in the SR module training allows the SR result to benefit from pixel-wise feedback from HR images and suffer less from artifacts brought by adversarial losses.

3.4. Collaborative training

Since our framework consists of two stages, one for translation module to learn degradation and the other for super-resolution module to give the final result. It is intuitive to train these two stages separately. In that way the performance of super-resolution would be heavily dependent on the training of degradation. However, independently trained translation module might not be perfect in producing real-like LR images of correct color or degradation level, which leads to unsatisfactory results on final super-resolution performance. Therefore, we propose to train these two stages in a collaborative way. With collaborative training the SR module is encouraged to give feedback to the translation module on the degradation quality. In other words, there is an additional constraint from SR module that forces the generated real LR images to keep the original content unchanged and only change degradation of images. In turn, the SR module is trained on a more realistic real-like LR images such that it can obtain good performance on real LR images. The total loss for the collaborative training can be formulated as follows,

$$\begin{aligned} &L_{total}(G_1, G_2, D_1, D_2, G_{sr}, D_h, I^L_{real}, I^L_{syn}, I^H) \\ &= L_{Trans}(G_1, D_1, G_2, D_2, I^L_{real}, I^L_{syn}) \\ &+ L_{SR}(G_{sr}, D_h, \hat{I}^L_{real}, I^H). \end{aligned} \tag{4}$$

The total loss L_{total} can be either optimized jointly with respect to parameters of both modules simultaneusly or alternatively between the total loss with respect to translation module and SR module respectively. When L_{SR} only contains distortion oriented loss L_{mse}, both kinds of optimization methods are equivalent. But when L_{SR} also contains perceptual quality oriented losses L_{percep} and L_{adv}, joint optimization with respect to all parameters would become unstable and difficult to converge. To make the optimization flexible and consistent for both cases, we choose to alternatively optimize the modified translation loss L'_{Trans} and SR loss L_{SR} during training in this work.

$$\begin{aligned} &L'_{Trans}(G_1, G_2, D_1, D_2, I^L_{real}, I^L_{syn}, I^H) \\ &= L_{Trans}(G_1, G_2, D_1, D_2, I^L_{real}, I^L_{syn}, I^H) \\ &+ \lambda^{SR}_{Trans} L_{mse}(G_{sr}, G_1(I^L_{syn}), I^H) \end{aligned} \tag{5}$$

Synthetic LR | NTIRE 2017 T2 (unknown) | NTIRE 2018 T2 (mild) | NTIRE 2018 T4 (wild)

Figure 4. Three tracks of super-resolution challenge on DIV2K.

The modified translation loss L'_{Trans} considers the feedback from SR module but is only used to update parameters of translation module. In the alternative optimization, the SR loss $L_{SR}(\hat{I}^L_{real}, I^H)$ considers $\hat{I}^L_{real}$ as constant and is only used to update parameters of SR module. The translation module would benefit from only reconstruction loss of HR images in the process of parameter update, which helps producing more reaistic degradation and makes it easy to train. We can also easily find that when $L_{SR} = L_{mse}$ and $\lambda^{SR}_{Trans} = \lambda_{mse}$, alternative optimization is equivalent to joint optimization of the total loss with respect to all parameters simultaneusly. Unless specified otherwise we have $\lambda^{SR}_{Trans} = \lambda_{mse}$ in all implementations.

4. Experiments

4.1. Experimental Setup

Datasets DIV2K [1, 28], which contains 1,000 images with different scenes and is splitted to $\{800, 100, 100\}$ for training, validation and testing. It was collected for NTIRE2017 and NTIRE2018 Super-Resolution Challenges in order to encourage research on image super-resolution with more realistic degradation. We evaluate the proposed method on three tracks in these two challenges, *i.e.*, unknown, mild and wild degradation respectively. Specifically, LR images in the unknown track suffer from only blur, the ones in the mild track suffer from both blur and Poisson noise, and the ones in the wild are similar but the level of blur and noise varies across images. In all three tracks, the degradation is unknown. An example is shown in Figure 4.

Since HR and degraded LR images appear in pairs in all three tracks, we can quantitatively evaluate the reconstruction performance of proposed method. To experiment with unsupervised setting, the first 400 images of HR and the rest 400 of LR are selected in original training set to compose our unpaired training set. Performance on the original validation set is used for comparison. Following [28], we evaluate all comparisons with 100 validation images in same way.

Quantitative metrics We use PSNR, SSIM and LPIPS to evaluate the performance. LPIPS [32] is a learned metric to measure the perceptual quality of reconstruction. And the other two are known as classical distortion measurements directly calculated on image pairs.

Table 1. Quantitative comparison for $4\times$ SR on three datasets: average PSNR/SSIM/LPIPS for scale factor x4. † means the methods are under unsupervised setting. The arrows indicate if high↑ or low↓ values are desired. **Blue** text indicates the best and **green** text indicates the second best performance.

Methods	NTIRE17 T2			NTIRE18 T2			NTIRE18 T4		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
†Bicubic	23.976	0.644	0.487	23.196	0.563	0.547	22.579	0.543	0.572
SR_syn	23.955	0.654	0.457	23.066	0.545	0.560	22.410	0.517	0.581
SR_paired	**29.819**	**0.818**	**0.326**	24.154	**0.617**	0.515	23.706	**0.590**	**0.538**
†CycleGAN	23.213	0.648	0.459	22.901	0.517	0.513	21.685	0.466	0.550
†Bulat *et al.*	25.204	0.690	0.433	23.609	0.585	0.530	22.655	0.536	0.561
†*Bulat *et al.*	26.191	0.705	0.409	23.280	0.555	0.564	23.250	0.554	0.553
†CycleSR	**27.021**	**0.770**	**0.399**	**24.779**	**0.631**	**0.500**	**23.807**	**0.593**	**0.526**
†UNITSR	26.613	0.732	0.428	**24.894**	0.616	**0.511**	**23.819**	0.576	**0.538**

Implementation details Our framework is composed of two stages, unsupervised translation and supervised super-resolution. For the first stage, we experiment with two popular translation models, CycleGAN [37] and UNIT [20]. For the second stage, we adopt a modified VDSR for track 2 of NTIRE17 and the original SRResNet model for track 2 and track 4 of NTIRE18 respectively. In training process, we use Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is set to 2×10^{-4} and 1×10^{-4} for the translation module and SR network. $\lambda_{SRtrans}$ are set to be $\{1e3, 1e2, 1e2\}$ respectively for track 2 of NTIRE17, track 2 and track 4 of NTIRE18. Other hyper-parameters of translation part remain unchanged as they originally are. Before joint training, translation module is pre-trained with unpaired real LR and synthetic LR images, while super-resolution module is pre-trained with HR and synthetic LR pairs. Both are pre-trained for a few epochs to have a good initialization for stable training. After joint training for 100 epochs, the learning rate of SR model starts to linearly decay and stops at zero after another 100 epochs. For fair comparison with baselines, only L2 loss is taken in SR training.

4.2. Comparison on DIV2K

In this subsection, we compare the proposed method with several baseline methods on three datasets. Both quantitative and qualitative results are given to demonstrate the effectiveness of our framework. By employing CycleGAN and UNIT as translation module and include only L_{mse} in SR losses L_{SR}, we can get two instances of the proposed model, **CycleSR** and **UNITSR**. Compared methods are briefly explained as below:

Bicubic: bicubic interpolation is applied to LR images;

SR_syn: Models of the same architecture as SR module in our framework is trained on synthetic LR-HR images pairs and evaluated on the three datasets;

CycleGAN: LR images are first upsampled with bicubic interpolation and then a CycleGAN is trained on those images and HR images;

Bulat *et al.*[2]: Since their training code is not publicly available, we carefully reimplement their method and tune the hyper-parameters to train the model. For fair comparison, we take the same network as the SR module in our method as low-to-high network, and modify a branch of CycleGAN as high-to-low network. The whole framework is trained with almost the same losses as mentioned in the paper except that we use only pixel loss on low-to-high network for good PNSR/SSIM.

Bulat *et al.[2]: Further our collaborative traning strategy are adopted in their method.

SR_paired: We also train models of the same architecture as SR module in our framework on original paired data. However, on track of mild and wild, paired LR image and HR images are not exactly aligned due to severe motion blur. We first preprocess these pairs to achieve alignment.

Quantitative results. We compute PSNR, SSIM and LPIPS scores for all methods. As shown in Table 1, the method trained on synthetic pairs performs even worse than the simple bicubic upsampling on all three datasets, which implies the importance of study on image super-resolution with unpaired real data. Among all the unsupervised methods which are marked by †, our CycleSR outperform all others, showing the advantage of the proposed framework in dealing with unpaired training data. Specifically, CycleGAN gives the worst reconstruction performance among all compared unsupervised methods. That could be attributed to the lack of pixel-wise reconstruction loss on HR space, which also explains why Bulat *et al.* [2] gives better performance. With our collaborative training strategy, Bulat *et al*. [2]can gain further improvement. In spite of this, CycleSR and UNISR still perform better than that due to superiority of the proposed unsupervised super-resolution framework. Our

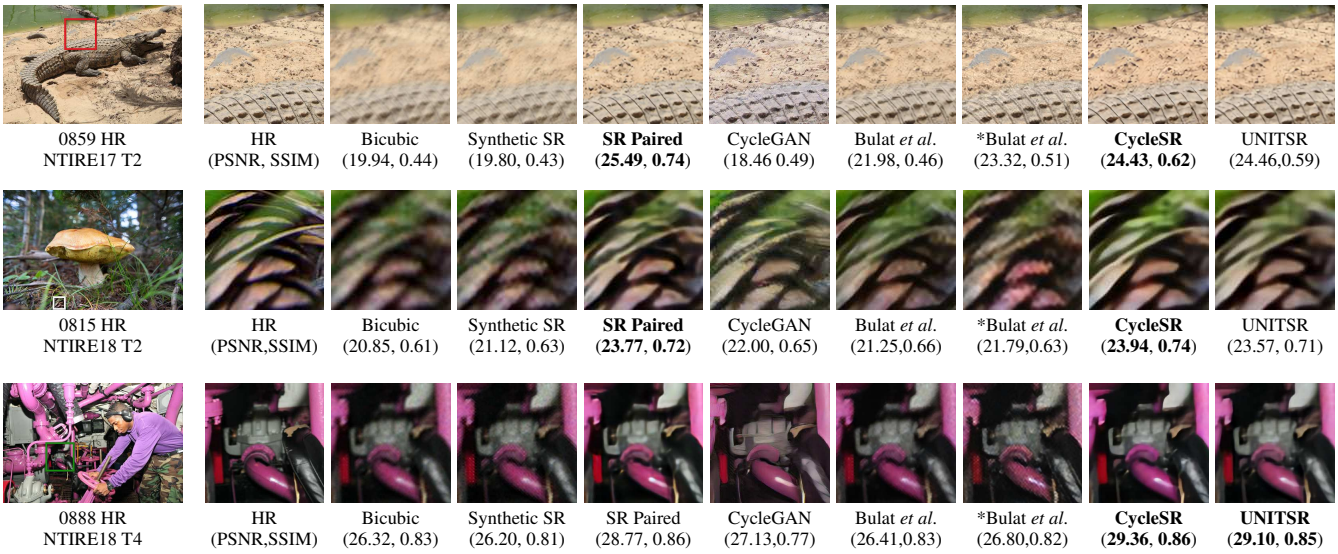


Figure 5. Visual comparison for $4\times$ SR on three datasets. *Bulat *et al*. denotes their method with our collaborative training. The best two results in terms of PSNR/SSIM are in **bold**

Table 2. Preliminary results for Track1 of NTIRE 2020 Real-World Super-Resolution Chanllenge.

Team Name	PSNR↑	SSIM↑	LPIPS↓
AITA_Noah_ExtraData	24.65 (18)	0.70 (13)	0.222 (1)
Noah_AITA_noExtraData	25.72 (11)	0.74 (9)	0.223 (2)
Impressionism	24.67 (17)	0.68 (15)	0.232 (3)
Samsung_SLSI_MSL	25.59 (13)	0.73 (10)	0.252 (4)
MSMers	23.20 (20)	0.65 (19)	0.272 (5)
BOE-IOT-AIBD	26.71 (4)	0.76 (4)	0.280 (6)

methods impose more constraint on the unsupervised degradation learning process such that it can generate more realistic LR images and the learned super-resolution model can give better performance at test stage. Compared to SR_paired method, our method is outperformed on Track2 of NTIRE 2017 but is comparable or even better on both tracks of NTIRE 2018. With severe blur and noise degradation with two tracks of NTIRE 2018, supervised methods are sensitive to the result of pre-alignment, while our methods do not require pre-alignment and are robust to such severe degradation.

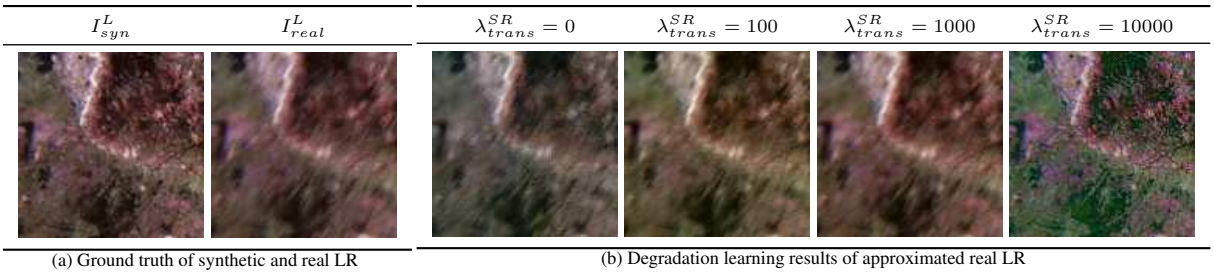
In addition, to demonstrate the effectiveness of our framework on real-world degradations, we also participated NTIRE2020 Real-World Super-Resolution [23] challenge. Our CycleSR is used by the team AITA-Noah to generate degraded images. For Track1, named image processing artifacts, CycleSR is the part of iterative data degradation framework to provide degraded images, while in Track2, named Smartphone Images, only degraded images produced by CycleSR are used. Furthermore, after getting fake-paired images via CylceSR, similar architecture based on the ESRGAN is used in both tracks to further improve the super-resolution performance. The preliminary results in Track1 are shown in Table 2 and unfortunately there is no release result in Track2 before paper submitting. From Table 2 can be seen that the method achieves superior LPIPS score compared to other approaches.

Qualitative results. We visualize the super-resolutions of compared methods in Figure 5. Model trained on clean pairs can not deal with LR images with noise and blur. CycleGAN is prone to images with much artifact and color drift. Bulat *et al*. [2] is also able to give reasonable results but still suffers from some artifacts and pollution by noise and blur. Our methods can remove blur and noise better and produce sharper images than theirs due to the robust degradation learning model and strong supervision on SR images with our framework. Our method obtains even comparable perceptual performance as the supervised method.

4.3. Ablation Study

In this section we conduct ablation study in terms of training strategy and SR losses. Experiments here are based on **CycleSR**, *i.e.*, CycleGAN is taken as the choice of the translation stage in our whole framework.

Training strategy: The first choice is to train the two stages separately, where super-resolution module is trained with a fixed pre-trained translation module. For the collaborative training, different model is obtained when $\lambda^{SR}_{Trans} = \lambda_{mse}$ and is chosen from 1e2, 1e3, 1e4. We also experiment with training with $\lambda^{SR}_{Trans} = 0$ and $\lambda_{mse} = 1$, *i.e.*, L_{Trans} and L_{SR} are optimized alternatively.



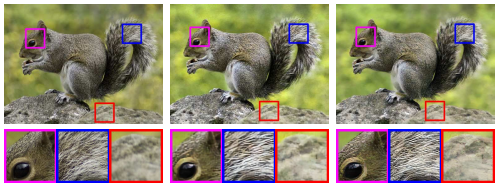
(a) Ground truth of synthetic and real LR

(b) Degradation learning results of approximated real LR

Figure 6. Visual comparison for real LR and degradation learning results of approximated real LR cases on NTIRE17 T2.

Table 3. Quantitative results of ablation study on training strategy. PSNR/SSIM/LPIPS of each strategy on NTIRE17 T2 are presented.

$\lambda_{SRtrans}$	Seperated	0	100	1000	10000
PSNR↑	24.663	25.479	25.886	**27.021**	17.352
SSIM↑	0.685	0.714	0.733	**0.770**	0.556
LPIPS↓	0.502	0.467	0.434	**0.399**	0.573



HR CycleSR CycleSRGAN

Figure 7. Visual comparison for our CycleSR and CycleSRGAN on track 2 of NTIRE 2018

As shown in Table 3, the strategy of training two stages separately perform worse than collaborative training strategies. Even for a special case of collaborative training, where the two modules are trained alternatively without feedback from SR module to translation module, it still obtains better SR result than the one trained separately. When $\lambda^{SR}_{Trans} = \lambda_{Trans}$ and is not zero, *i.e.*, there is feedback from SR module to translation module, the performance is further improved. The best result is obtained when it is set to 1000. However, as it further increases to $1e4$, it dominates the optimization of translation module. In this way, the degradation learning overfits in terms of excessively satisfying the super-resolution part.

Moreover in Figure 6 we visualize the degraded LR images $\hat{I}^L_{real}$ to show our degradation learning results. We can see that the level of degradation gets closer to the real one as feedback from SR module increases. The best result is given when $\lambda_{SR} = 1e3$. However, there is severe overfitting with further increase in λ_{SR} so that the degradation looks like bicubic downsampling.

Choice of SR loss Since our framework is flexible in choices of SR loss and the SR modules are trained with only fidelity loss L_{mse} in above experiments. Here we also include perceptual oriented losses in SR loss, and propose the CycleSRGAN for better perceptual quality. As shown in Table 4, CycleSRGAN performs worse than CycleSR in terms of PSNR and SSIM, but gives higher perceptual scores in terms of LPIPS in all three tracks. Visual comparison in Figure 7 shows that CycleSRGAN produces sharper edges and more details in the super-resolution result.

Table 4. Quantitative results of our method trained with different losses in supervised path: average PSNR/SSIM/LPIPS for scale factor x4 in all three challenges.

Challenge		CycleSR	CycleSRGAN
PSNR↑	NTIRE17 T2	**27.021**	25.978
	NTIRE18 T2	**24.779**	23.605
	NTIRE18 T4	**23.807**	22.303
SSIM↑	NTIRE17 T2	**0.770**	0.737
	NTIRE18 T2	**0.631**	0.545
	NTIRE18 T4	**0.593**	0.497
LPIPS↓	NTIRE17 T2	0.399	**0.377**
	NTIRE18 T2	0.500	**0.466**
	NTIRE18 T4	0.526	**0.489**

5. Conclusion

In this work, we present a general framework for unsupervised image super-resolution, which is closer to real scenario. Instead of directly applying unsupervised image translation to address this task, we propose a novel approach which integrates translation and supervised training into one framework and enables collaboration between two modules in training process. Synthetic LR images are taken as a bridge and creates an indirect supervised path from real LR images to HR images. We show that the proposed approach learns to super-resolve a real LR image without any corresponding HR images in the training dataset. It is flexible enough to integrate any existed deep learning based translation and super-resolution models, including those trained with either fidelity losses or perceptual oriented losses. It is evaluated on image super-resolution challenge datasets and achieves favorable performance against supervised methods.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017.

[2] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *ECCV*, 2018.

[3] 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. *CoRR*, abs/1904.00523, 2019.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2):295–307, 2016.

[5] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCV Workshops*, 2019.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[12] TAlexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *ICLR*, 2019.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016.

[15] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017.

[17] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017.

[20] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

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

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

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

[24] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017.

[25] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[26] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017.

[27] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *CVPR*, 2017.

[28] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. NTIRE 2018 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2018.

[29] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *ECCVWS*, 2018.

[30] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[31] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPRW*, 2018.

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[33] Xuaner Zhang, Qifeng Chen, Ren Ng, , and Vladlen Koltun. A novel loss (cobi) for slightly misalighned image-to-image translation. In *CVPR*, 2019.

[34] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

[35] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018.

[36] Tianyu Zhao, Wenqi Ren, Changqing Zhang, Dongwei Ren, and Qinghua Hu. Unsupervised degradation learning for single image super-resolution. *CoRR*, abs/1812.04240, 2018.

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.