Blind Image Super-Resolution with Degradation-Aware Adaptation*

Yue Wang[1], Jiawen Ming[1], Xu Jia[1**], James H. Elder[2], and Huchuan Lu[1**]

[1] Dalian University of Technology, Dalian, 116024, China
[2] York University, Toronto, M3J 1P3, Canada

Abstract. Most existing super-resolution (SR) methods are designed to restore high resolution (HR) images from certain low resolution (LR) images with a simple degradation, *e.g.* bicubic downsampling. Their generalization capability to real-world degradation is limited because it often couples several degradation factors such as noise and blur. To solve this problem, existing blind SR methods rely on either explicit degradation estimation or translation to bicubicly downsampled LR images, where inaccurate estimation or translation would severely deteriorate the SR performance. In this paper, we propose a plug-and-play module, which could be applied to any existing image super-resolution model for feature-level adaptation to improve the generalization ability to real-world degraded images. Specifically, a degradation encoder is proposed to compute an implicit degradation representation with a ranking loss based on the degradation level as supervision. The degradation representation then works as a kind of condition and is applied to the existing image super-resolution model pretrained on bicubicly downsampled LR images through the proposed region-aware modulation. With the proposed method, the base super-resolution model could be fine-tuned to adapt to the condition of degradation representation for further improvement. Experimental results on both synthetic and real-world datasets show that the proposed image SR method with compact model size performs favorably against state-of-the-art methods. Our source code is publicly available at https://github.com/wangyue7777/blindsr_daa.

Keywords: Blind super-resolution · multiple unknown degradations · feature-level adaptation · ranking loss · region-aware modulation.

1 Introduction

Single Image Super-Resolution (SISR) aims at predicting high-resolution (HR) images with high-frequency details from their corresponding low-resolution (LR) images. Inspired by the success of deep learning, numerous existing SR methods [2,14,21] apply CNN-based models to effectively restore the HR image based

* Partially supported by the Natural Science Foundation of China, No. 62106036, and the Fundamental Research Funds for the Central University of China, DUT21RC(3)026.
** Corresponding authors. Email address: {xjia, lhchuan}@dlut.edu.cn

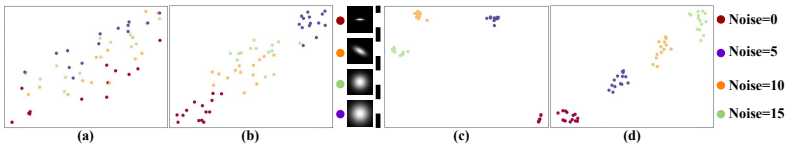

Fig. 1: An illustration of the degradation representation supervised by contrastive loss in [28] (a, c), and the proposed ranking loss (b, d) with different blur kernels (left) and noise levels (right) on Set14. Ranking loss can not only separate different degradations, but also provide information of degradation level.

on a fixed and known degradation (*e.g.* bicubic downsampling). However, these methods may have limited generalization to real-world situation where multiple degradations with unknown blur, downsampler, and noise are coupled together.

To address the SR problem with multiple degradations, several non-blind and blind SR approaches have been proposed. Most non-blind methods [3, 23, 31, 34] usually require both LR image and its explicit ground-truth degradation as inputs to predict the corresponding HR image. While most blind methods [13, 19, 22] conduct the explicit degradation estimation first and then combine with non-blind SR methods to restore HR images. However, when dealing with unknown degradation at inference stage, it may be difficult for these non-blind and blind approaches to give reasonable performance. Inaccurate explicit degradation would seriously deteriorate the performance of HR image restoration.

Instead of explicitly estimating degradation of an LR image, a new attempt called DASR [28] has been made to implicitly learn degradation representation in feature space with a degradation encoder following contrastive learning fashion [6, 12] for blind SR. Such degradation representation can distinguish various degradations and lead to a degradation-aware SR model. However, the degradation encoder is only taught to distinguish one from the other without being aware of whether its degradation is heavier than the other, which is more important in adjusting an SR model to adapt to certain degradation.

On the other hand, designing and training powerful SR models for bicubic downsampling degradation has already cost a lot of human and computation resources. However, data distribution gap between bicubicly downsampled images and images with more practical degradations, prevents those existing pretrained SR models from being generalized well to LR images with multiple degradations. There have been several works [15, 24] that take advantages of these existing SR models pretrained on bicubicly downsampled images to promote the development of SR models for real-world degradations. Image translation techniques [9, 16, 38] are employed to convert images of interesting degradations to bicubicly downsampled ones and then those converted images are fed to existing pretrained SR models for HR restoration. However, the imperfect translation would cause performance drop in the process of restoring HR image and produce many artifacts.

In this work, we propose a plug-and-play module for blind SR, which can be applied to any existing SR models pretrained on bicubicly downsampled

data. It provides degradation-aware feature-level adaptation to improve the generalization ability to real-world degraded images. It consists of three components: the pretrained SR model, a degradation encoder followed by a ranker, and a degradation-aware modulation module. Specifically, the degradation encoder computes a latent degradation representation such that the SR model can be adapted to various degradation. Ranking loss is imposed on top to make correct decision on estimating the degree of degradation in an image as illustrated in Fig. 1. To make an existing SR model adapt to various degradations, the degradation representation works as a condition to apply the proposed degradation-aware modulation to the intermediate features of SR model. Even if the degradation is spatially-invariant, different types of textures may have different sensitivity to the degradation. Hence, the modulation is designed to be region-aware and sample-specific. It allows a region in an LR image to be super-resolved adaptively not only to different degradations but also to content of the image. To further improve its performance, the pretrained SR model is fine-tuned together with the degradation encoder and modulation module. With an existing pretrained light-weight SR model as our SR model, we can obtain a compact SR model that performs favorably against existing blind SR models of larger model size.

Main contributions of this work are three-fold.

- We come up with a novel plug-and-play module for blind SR. It can provide degradation-aware feature-level adaptation for any existing SR network pretrained with the degradation of only bicubic downsampling to improve the generalization ability to various degraded images.

- We propose a ranking loss for extracting degradation representation with information of the degradation degree, and a dynamic region-aware modulation for adaptation on intermediate features within the pretrained SR network.

- Our method has relatively compact model size and performs favorably against the state-of-the-art SR methods on both synthetic and real-world datasets.

2 Related Work

SR with a Simple Degradation. Early SR methods focus on LR images with a simple degradation, *e.g.* bicubic downsampling. Since [7], numerous SR methods apply CNN-based networks to improve performance. [21, 36] enhance the results by utilizing deep residual networks with excessive convolutions layers. [14,18] design lightweight networks for SR to save the computational costs, while preserving good SR quality. However, these methods can not generalize well on real-world LR images which couples multiple unknown degradations.

SR with Various Degradations. To address this problem, several non-blind methods [3, 23, 31, 34] have been proposed to use the explicit ground-truth degradation as inputs for HR restoration. However, they have limited applications since the explicit ground-truth degradation may be unknown during inference. Recently, blind SR methods have been investigated to avoid requiring the explicit ground-truth degradation as input during inference. Several methods [13, 19, 20, 22] apply the explicit degradation estimation for assisting the HR

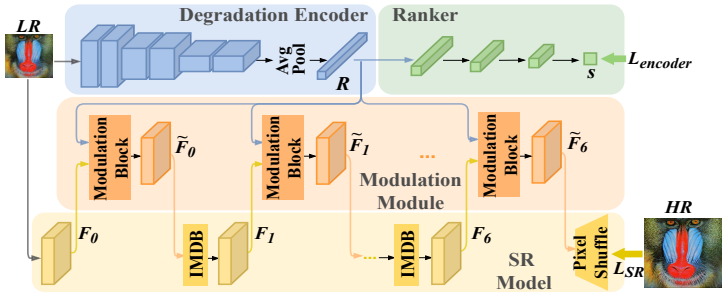

Fig. 2: The structure of our overall method. It contains a degradation encoder with a ranker; an SR model pretrained on bicubicly downsampled LR images; and a degradation-aware modulation module with several modulation blocks.

restoration. [19] first predicts the kernel from LR image, and then applies a non-blind SR method using both the estimated kernel and LR image as inputs. [13] adopts a deep alternating network where the kernel estimation and HR restoration can be alternately and repeatedly optimized. Meanwhile, [15] estimates the correction filter and [24] uses Generative Adversarial Network (GAN) for image translation to transfer the LR image to look like a bicubicly downsampled one, which can then be fed to any existing SR model pretrained on bicubicly downsampled LR images for HR restoration. However, these non-blind methods are sensitive to the degradation estimation or image translation so that any errors happened in these processes would severely deteriorate the SR performance. A novel strategy for blind SR is to learn an implicit degradation representation [28] with contrastive learning, and build a fixed SR structure with dozens of modulation blocks for generating HR image with specific information of degradation.

Meanwhile, blind SR methods with other alternative ways have been proposed. [29, 32, 37] try to improve the generalization capability of any SR models by generating a large number of synthetic data for training. Specially, [29, 32] design practical degradation models by considering the real-life degradation processes, while [37] uses GAN [10] to generate realistic blur-kernels. The self-supervised internal learning can also be used for blind SR [25, 26]. [25] applies zero-shot learning by training a small image-specific CNN at test time, while [26] further applies meta-transfer learning [8] to decrease the gradient steps.

In this paper, we propose a novel flexible plug-and-play module for blind SR. It can be applied on any existing SR networks pretrained with the degradation of only bicubic downsampling to improve its generalization ability to real-world degraded images. Our method requires implicit degradation representation learning for degradation-aware modulation with two improvements. Firstly, we use a ranking loss instead of contrastive learning, which can provide the degree level of a degradation. Secondly, we propose a region-aware modulation instead of uniformly modulating features on all spatial positions.

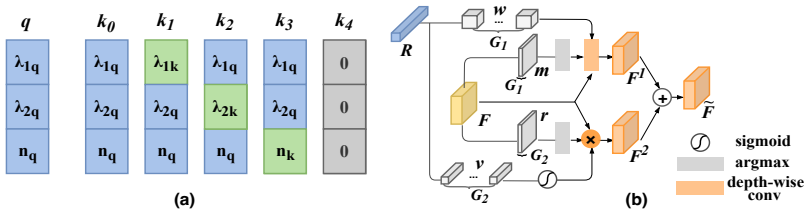

Fig. 3: Details of the proposed method. (a) Data preparation for training degradation encoder. Two kinds of patches are cropped from an HR image to apply various degradations and form q and k_i. (b) The structure of degradation-aware modulation block. It consists of two types of dynamic region-aware modulation: a depth-wise convolution and a channel-wise attention.

3 Method

3.1 Problem Formulation

In this work, we focus on blind super-resolution with various unknown degradations. The degradation model can be formulated as:

$$I^{LR} = (I^{HR} \otimes kernel) \downarrow_{sf} + noise, \tag{1}$$

where I^{HR} is the original HR image, I^{LR} is the degraded LR image. $kernel$, $noise$, $\downarrow_{sf}$ and $\otimes$ denote the blur kernel, noise, downsampling operation with scaling factor of sf and convolution operation. Following [28], we apply bicubic downsampling, the anisotropic Gaussian kernels and Additive White Gaussian Noise to synthesize LR images for training. The anisotropic Gaussian kernel we use is characterized by a Gaussian probability density function $N(0, \Sigma)$, which means it has zero mean and varying covariance matrix Σ. Therefore, the blur kernel can be determined by two eigenvalues λ_1, λ_2, and rotation angle θ of its Σ. The Additive White Gaussian Noise can be determined by its noise level n.

Given an existing SR model pretrained on LR-HR pairs with only bicubic downsampling as degradation, the goal of our method is to adapt this pretrained SR model to work on images with various unknown degradations. To achieve this purpose, we introduce a framework which consists of three parts: a pretrained base SR model, a light degradation encoder with ranker, and a degradation-aware modulation module with several light modulation blocks. The structure of this end-to-end system is shown in Fig 2.

3.2 Weakly Supervised Degradation Encoder

Instead of explicitly estimating parameters for various degradations, the goal of a degradation encoder is to distinguish various degradations and to make estimation about degradation levels. A very recent work called DASR [28] makes initial attempt in computing degradation representation by means of contrastive

learning. However, the degradation representation can only distinguish various degradations among images but can not tell which image is more severely degraded than the other. To address this issue, we propose to append a ranker to the end of degradation encoder in the training process.

Both degradation encoder and ranker are light-weighted with only six and four convolution layers. To make the degradation representation encoded by degradation encoder of high discriminative capability, the ranker after degradation encoder is taught to give higher scores to images with heavier degradation and lower scores to the ones with lighter degradation. Although the model does not need exact degradation parameters as supervision, it requires order relation between a pair of images on level of degradation. Therefore, we need to prepare images with various degradation levels to train degradation encoder and ranker.

Data Preparation. According to Eq. 1, when generating I^{LR} from I^{HR}, the degradation is determined by four parameters: λ_1, λ_2, θ for blur, and n for noise. Here we use λ_1, λ_2 and n to determine the ranking order of different degradations since degradation with larger values of them would have higher degree level. While θ only affects the rotation, not the degree of degradation.

For each I^{HR}, we randomly extract two types of patches to apply various degradations. The first kind of patch is used to generate a query LR image q, and the second kind of patch is used to generate five key LR images as k_i for calculating the loss for encoder, where $i \in \{0, 1 \dots 4\}$. Two sets of degradation parameters are randomly selected for query patches and key patches, which are indicated as $P_q = \{\lambda_{1q}, \lambda_{2q}, n_q\}$ and $P_k = \{\lambda_{1k}, \lambda_{2k}, n_k\}$. P_q is used as the degradation parameters for generating q and k_0 separately for two patches so that these two LR images contain different contents but the same degradation. Then, we generate degraded key patches k_i ($i \in \{1, 2, 3\}$) by P_{k_i}, where $P_{k_i}[i] = P_k[i]$ and the other two parameters remain the same as P_q, so that only one parameter of k_i is different from q. Finally, we generate k_4 by using only bicubic downsampling as degradation. It is used as a baseline for other LR images with degradation indicated as $P_{k_4} = \{0, 0, 0\}$ since it does not have any blur and noise. The degradation parameters of q and all k_i are shown in Fig.3(a).

Ranking-based Supervision. The degradation encoder and ranker produce both degradation representations and ranking scores for the generated LR images. The degradation representations are indicated as $R_q \in \mathbb{R}^C$ and $R_{k_i} \in \mathbb{R}^C$ which are used for modulation, while the ranking scores s_q and s_{k_i} are just numbers for calculating the loss for encoder.

First of all, we calculate the ranking loss by q, k_1, k_2 and k_3 in a pair-wise manner. Given q and each k_i, here $i \in \{1, 2, 3\}$, only the value of i-th parameter for the degradation is different. So it is easy to decide the ground-truth ranking order for these two images as:

$$\begin{cases} s_q < s_{k_i} & if \quad P_q[i] < P_{k_i}[i] \\ s_q > s_{k_i} & if \quad P_q[i] > P_{k_i}[i] \end{cases} \tag{2}$$

Therefore, we train our degradation encoder by a margin-ranking loss [35] to guide the output ranking scores to have the right order:

$$\begin{aligned} L_{rank_1} &= \sum_{i=1}^{3} max(0, (s_q - s_{k_i}) * \gamma + \varepsilon) \\ where \quad & \begin{cases} \gamma = 1 & if \quad P_q[i] < P_{k_i}[i] \\ \gamma = -1 & if \quad P_q[i] > P_{k_i}[i] \end{cases} \end{aligned} \tag{3}$$

where γ indicates the ground-truth order between q and k_i, ε is the margin which controls the distance between two scores. By forcing the ranking scores to have the right order, it encourages the degradation representations to encode information about degradation and its level for later degradation-aware modulation. With an appropriate margin ε, our ranking loss can also encourage distinguishing degradation representations between LR images degraded in different ways.

Meanwhile, we also force LR images with the same degradation to have the same ranking score. It is achieved by using L1 loss ($L1$) to supervise the ranking score s_q and s_{k_0} of q and k_0 to have the same value:

$$L_{eq} = L1(s_{k_0}, s_q) \tag{4}$$

We then use k_4 as a baseline which has lower scores than q and all other k_i, and has no difference on degradation with bicubicly downsampling LR images:

$$\begin{aligned} L_{rank_2} = &\sum_{i=0}^{3} max(0, (s_{k_i} - s_{k_4}) * -1 + \varepsilon) \\ &+ max(0, (s_q - s_{k_4}) * -1 + \varepsilon) \end{aligned} \tag{5}$$

$$L_{k_4} = L1(s_{k_4}, 0). \tag{6}$$

The overall loss for degradation encoder includes the above mentioned losses:

$$L_{encoder} = L_{rank_1} + \beta_1 L_{rank_2} + \beta_2 L_{k_4} + \beta_3 L_{eq} \tag{7}$$

where β_1, β_2, β_3 are hyper-parameters to combine these losses.

3.3 Degradation-aware Modulation

Basic SR Model. We first introduce the basic SR model to be modulated. In this work, we take a lightweight super-resolution network IMDN [14] as the basic model. It is composed of six information multi-distillation blocks (IMDBs) and gives reasonable performance on bicubicly downsampled images. To show that the proposed method is able to work on any existing SR model, we do not make any changes on top of IMDN. For the above-mentioned generated LR images, we only use q as input for HR restoration for efficiency. Then we denote $F_l \in \mathbb{R}^{C \times H_l \times W_l}$ as the feature of q within IMDN, where $l \in \{0, 1, ...6\}$. H_l, W_l are the resolution of F_l and C is the feature dimensionality. Here, $F_l, l \in$

$\{1, 2, ...6\}$ indicates the feature output from the l-th IMDB, and F_0 indicates the feature before the first IMDB. Specifically, the pretrained IMDN is fine-tuned together with training the degradation encoder and modulation module for better performance.

Modulation Module. To adapt each feature F_l of q in SR model, we design a modulation module which consists of seven modulation blocks. These modulation blocks are separately inserted into IMDN, so that each block is used to modulate one feature of q before and after each IMDB. Meanwhile, each modulation block uses information of degradation degree from q's degradation representation R for adapting F_l.

We notice that DASR applies two kinds of modulations in one modulation block, a depth-wise convolution and a channel-wise attention. The first one takes the degradation representation R as input and predicts one convolutional kernel $w_l \in \mathbb{R}^{C\times1\times3\times3}$. w_l is then used as the parameters of a 3×3 depth-wise convolution on F_l to process F_l^1. The second one also takes R as input to predict one channel-wise modulation coefficient $v_l \in \mathbb{R}^C$. v_l is then used to rescale different channel components for all spatial positions of F_l to produce F_l^2. F_l^1 and F_l^2 are added together to form the adapted feature. Both two modulations in DASR assume the degradation equally affects all spatial positions of one image so that they only learn one set of w_l and v_l for all spatial positions.

However, even though by applying a spatially-invariant degradation for all spatial positions in an HR image, the degradation may have different impacts on different spatial positions. It is mainly because different types of textures have different sensitivity to the degradation. For example, positions which present the contour of an object would contain more high-frequency information. More information loss may occur on these positions when applying the degradation to HR image. Consider that, it would be better to design a region-wise and sample-specific modulation. Therefore, we propose an efficient region-aware modulation by modifying the modulation based on DRConv [5] (Fig.3(b)).

Region-Aware Modulation. For the first kind of modulation, instead of only learning one convolution kernel for all positions in F_l, we follow DRConv to learn G_1 filter kernels from degradation representation R and denote them as w_l^g, $g \in \{1, 2, \cdots, G_1\}$. G_1 is the number of kernels in this kind of modulation. Each filter kernel $w_l^g \in \mathbb{R}^{C\times1\times3\times3}$ is only applied to a selected number of spatial positions instead of all positions in F_l. We then learn a series of guided masks $m_l^g \in \mathbb{R}^{H_l\times W_l}$ which divide all spatial positions in F_l to G_1 groups. Each mask represents one region of F_l with a selected number of positions. Here, we learn these spatial-wise masks from F_l to focus on the characteristic of the feature to be modulated. In this way, only the g-th kernel is applied on the selected positions in g-th mask, so that different positions from F_l would adaptively select different kernels to apply. To maintain the efficiency, we also apply the idea of depth-wise convolution as DASR. The c-th dimensionality of output feature map F_l^1 can be expressed as follow where (h, w) is one point in selected positions in m_l^g:

$$\begin{aligned} {F_l^1}_{(h,w,c)} &= {F_l}_{(h,w,c)} * {w_l^g}_{(c)} \\ (h, w) &\in m_l^g \end{aligned} \tag{8}$$

The channel-wise attention modulation can also be modified in the same way to achieve a dynamic region-aware modulation. It means that we can also predict G_2 channel-wise coefficients from R and denote them as v_l^g, $g \in \{1, 2, \cdots, G_2\}$. G_2 is the number of channel-wise coefficients for this modulation. Each $v_l^g \in \mathbb{R}^C$ is also only applied to a selected number of spatial positions in F_l according to a new series of guided masks $r_l^g \in \mathbb{R}^{H_l \times W_l}$ which divide all positions in F_l to G_2 groups. The c-th dimensionality of output feature map F_l^2 is expressed as:

$$\begin{aligned} {F_l^2}_{(h,w,c)} &= {F_l}_{(h,w,c)} \times {v_l^g}_{(c)} \\ (h,w) &\in r_l^g \end{aligned} \tag{9}$$

Here (h, w) is one of the points in selected positions in r_l^g.

We obtain the adapted feature by combining the modulated features from two kinds of modulations. By modifying the original modulation to be in a region-aware way, we can achieve an efficient pixel-wise modulation. Same as DASR, we apply the two kinds of above-mentioned modulation twice in each modulation block, while for simplicity, we only show the structure of applying them once in Fig.3(b). The final adapted feature from i-th modulation block is denoted as $\widetilde{F}_i$ and is used as the input to the $(i+1)$-th IMDB in the SR model to predict the SR image. The loss function on SR is:

$$L_{SR} = L1(SR, HR) \tag{10}$$

And the overall loss function is as follow:

$$L_{overall} = L_{SR} + \alpha * L_{encoder} \tag{11}$$

With the help of modulation module, the output SR image is specifically predicted based on the degradation of the input LR image. It helps to improve the generalization of the overall structure to work for not only the multiple degradations in training set, but also any unknown degradations in testing set.

4 Experiments

4.1 Datasets

Following [28], we use 800 training HR images in DIV2K [1] and 2650 training images in Flickr2K [27] and apply Eq. 1 to synthesize LR images, which form LR-HR pairs as our training set. Specifically, we apply the anisotropic Gaussian kernels, bicubic downsampler and additive white Gaussian noise in Eq. 1. For bicubic downsampler, the scaling factor sf is set to 4 for $\times 4$ SR. For anisotropic Gaussian kernels, the kernel size is fixed to 21×21, the ranges of eigenvalues λ_1 and λ_2 are set to $[0.2, 4.0)$, and the range of rotation angle θ is set to $[0, \pi)$. For additive white Gaussian noise, the range of noise level n is set to $[0, 25)$.

During inference, we use HR images from benchmark dataset Set14 [30] and also apply Eq. 1 to synthesize LR images as [28]. These LR-HR pairs are used as

Table 1: Quantitive results (×4 SR) on in-domain synthetic test sets. The best two results are in Red and Blue. We also present the number of parameters (Params) and Flops.

Method	Params (M)	Flops (G)	$\lambda_1/\lambda_2/\theta$ / n	2.0/0.5/0 PSNR	2.0/0.5/0 SSIM	2.0/1.0/10 PSNR	2.0/1.0/10 SSIM	3.5/1.5/30 PSNR	3.5/1.5/30 SSIM	3.5/2.0/45 PSNR	3.5/2.0/45 SSIM	3.5/2.0/90 PSNR	3.5/2.0/90 SSIM
DnCNN+IKC	6.0	24.3	10	26.00	0.6874	26.06	0.6837	24.56	0.6281	24.44	0.6190	24.49	0.6201
			15	25.53	0.6659	25.50	0.6619	24.23	0.6134	24.11	0.6049	24.12	0.6040
DnCNN+DAN	5.0	81.1	10	25.78	0.6783	25.70	0.6722	24.23	0.6174	24.05	0.6069	24.12	0.6086
			15	25.36	0.6596	25.25	0.6533	23.98	0.6055	23.81	0.5962	23.86	0.5973
DnCNN +CF+RCAN	26.3	68.1	10	25.15	0.6781	25.12	0.6747	23.85	0.6210	23.60	0.6100	23.58	0.6099
			15	24.56	0.6545	24.48	0.6492	23.40	0.6012	23.22	0.5914	23.05	0.5894
DnCNN+SRMDNF +Predictor	2.2	9.1	10	25.97	0.6843	25.78	0.6747	24.14	0.6149	23.92	0.6032	23.98	0.6046
			15	25.55	0.6656	25.38	0.6577	23.94	0.6052	23.74	0.5952	23.79	0.5965
DnCNN+MANet	10.6	40.8	10	20.20	0.5023	20.33	0.5091	21.12	0.5329	21.18	0.5332	20.92	0.5207
			15	20.23	0.5034	20.36	0.5089	21.09	0.5291	21.13	0.5284	20.86	0.5175
BSRNet	16.7	73.5	10	25.59	0.6803	25.57	0.6772	24.69	0.6430	24.61	0.6369	24.63	0.6362
			15	24.39	0.6493	24.37	0.6460	23.67	0.6156	23.55	0.6097	23.61	0.6111
DASR	6.0	13.1	10	26.58	0.7030	26.49	0.6960	25.62	0.6624	25.44	0.6538	25.43	0.6527
			15	26.04	0.6827	25.93	0.6764	25.10	0.6434	24.94	0.6350	24.91	0.6346
Ours	2.9	6.2	10	26.58	0.7008	26.47	0.6939	25.62	0.6616	25.48	0.6538	25.42	0.6523
			15	26.04	0.6815	25.94	0.6752	25.12	0.6436	24.97	0.6358	24.93	0.6347

our synthetic testing set to conduct experiments with various unknown degradations. Meanwhile, we enlarge the range of three parameters of degradation, λ_1, λ_2 and n during testing. We also use RealSR [4] testing set to further evaluate the performance of our method on real-world LR images. It contains 100 HR images and their corresponding LR images taken from real world.

4.2 Implementation Details

Our model is implemented based on the Pytorch toolbox and trained on one GTX 3090 GPU. The batch size and the patch size of LR images are set to 32 and 48×48 during training. We also use random rotation and flipping as the data augmentation technique during training to avoid overfitting. In our experiments, we set $\alpha = 0.2$, $\beta_1 = 0.2, \beta_2 = 2, \beta_3 = 5$, $\varepsilon = 0.5$, $G_1 = G_2 = 4$. For the optimizer, we adopt Adam [17]. The overall network is trained in two stages. For stage one, we only train the degradation encoder with ranker by optimizing Eq. 7 for 100 epochs. The initial learning rate is set to 10^{-4} and decreased with the power of 0.1 after 60 epochs. For stage two, we freeze the ranker while training all the rest parts by optimizing Eq. 11 for 700 epochs. The initial learning rate is set to 10^{-4} and decreased to half after every 125 epochs.

4.3 Comparison with state-of-the-art methods

To verify the effectiveness of our proposed blind SR method with degradation-aware adaptation, we compare our method with 7 state-of-art SR methods, including one non-blind SR method: SRMDNF [34]; three blind SR methods with explicit degradation estimation: IKC [11], DAN [13], MANet [19]; one blind SR with image translation: CF [15]; one blind SR method with implicit degradation representation learning: DASR [28]; and one blind SR method with enlarged synthetic training data: BSRNet [32]. Specifically, IKC, DAN, MANet and SRMDNF are designed to handle images with only blur kernel and downsampling

Table 2: Quantitive results (×4 SR) on out-domain synthetic test sets and real world test set (RealSR). The best two results are in Red and Blue.

Method	Params (M)	Flops (G)	$\lambda_1/\lambda_2/\theta$ / n	Out-domain 2.0/1.0/10 PSNR	SSIM	3.5/1.5/30 PSNR	SSIM	3.5/2.0/45 PSNR	SSIM	3.5/4.5/60 PSNR	SSIM	4.5/5.0/120 PSNR	SSIM	5.0/5.0/180 PSNR	SSIM	RealSR PSNR	SSIM
DnCNN +IKC	6.0	24.3	30	24.22	0.6152	23.30	0.5801	23.20	0.5732	22.29	0.5412	21.94	0.5286	21.86	0.5248	27.19	0.7833
			40	23.49	0.5941	22.79	0.5640	22.68	0.5588	21.89	0.5305	21.57	0.5189	21.51	0.5163		
			50	22.85	0.5751	22.23	0.5491	22.13	0.5446	21.43	0.5192	21.21	0.5104	21.12	0.5080		
DnCNN +DAN	5.0	81.1	30	24.16	0.6128	23.20	0.5775	23.10	0.5712	22.21	0.5400	21.84	0.5280	21.78	0.5248	27.80	0.7877
			40	23.47	0.5928	22.72	0.5630	22.61	0.5580	21.83	0.5306	21.52	0.5198	21.45	0.5173		
			50	22.85	0.5751	22.19	0.5490	22.09	0.5447	21.39	0.5199	21.17	0.5121	21.09	0.5095		
DnCNN +CF+RCAN	26.3	68.1	30	23.25	0.5973	22.49	0.5604	22.28	0.5491	21.41	0.5088	21.05	0.4902	21.03	0.4883	27.72	0.7825
			40	22.47	0.5724	21.94	0.5371	21.78	0.5313	21.10	0.4951	20.71	0.4745	20.60	0.4696		
			50	21.89	0.5500	21.34	0.5177	21.29	0.5119	20.75	0.4797	20.37	0.4643	20.27	0.4562		
DnCNN +SRMDNF +Predictor	2.2	9.1	30	24.28	0.6174	23.20	0.5799	23.11	0.5741	22.17	0.5414	21.81	0.5292	21.75	0.5259	27.62	0.7789
			40	23.54	0.5956	22.73	0.5664	22.62	0.5611	21.81	0.5331	21.49	0.5219	21.43	0.5196		
			50	22.87	0.5772	22.18	0.5516	22.08	0.5472	21.37	0.5230	21.14	0.5148	21.06	0.5122		
DnCNN +MANet	10.6	40.8	30	20.37	0.5052	20.89	0.5169	20.93	0.5170	20.80	0.5045	20.81	0.5042	20.77	0.5023	oom	oom
			40	20.37	0.5035	20.73	0.5111	20.79	0.5116	20.64	0.5005	20.62	0.4991	20.58	0.4976		
			50	20.28	0.5007	20.60	0.5059	20.58	0.5055	20.39	0.4944	20.42	0.4942	20.34	0.4920		
BSRNet	16.7	73.5	30	21.44	0.5871	21.01	0.5644	20.92	0.5609	20.51	0.5405	20.34	0.5317	20.28	0.5284	27.35	0.8071
			40	19.95	0.5519	19.77	0.5380	19.70	0.5342	19.38	0.5176	19.29	0.5114	19.22	0.5085		
			50	18.99	0.5250	18.78	0.5121	18.76	0.5107	18.58	0.4979	18.43	0.4933	18.46	0.4920		
DASR	6.0	13.1	30	24.67	0.6320	23.87	0.6011	23.81	0.5963	22.95	0.5650	22.45	0.5471	22.36	0.5430	27.80	0.7934
			40	23.65	0.5969	23.06	0.5701	23.02	0.5672	22.27	0.5396	21.88	0.5239	21.79	0.5194		
			50	22.82	0.5693	22.42	0.5480	22.36	0.5439	21.69	0.5174	21.49	0.5094	21.33	0.5045		
Ours	2.9	6.2	30	24.73	0.6335	23.95	0.6041	23.89	0.5995	22.96	0.5663	22.46	0.5483	22.36	0.5441	27.84	0.8024
			40	23.93	0.6061	23.33	0.5814	23.22	0.5760	22.42	0.5475	22.00	0.5319	21.91	0.5278		
			50	23.16	0.5719	22.62	0.5491	22.54	0.5452	21.79	0.5196	21.56	0.5117	21.42	0.5058		

as degradation without noise. We notice that DASR tests this kind of SR model by first denoising the testing LR images using DnCNN [33]. For a fair comparison, we follow the same strategy to test these four methods and Predictor of IKC is used to estimate blur kernels for SRMDNF. Meanwhile, [32] focuses on improving perceptual quality by training SR model with adversarial loss. Here, we compare with its non-GAN version (BSRNet) for fairness.

Comparison on Synthetic Data. During training, we set ranges to the four degradation parameters, λ_1, λ_2, n and θ, the same as DASR for generating LR images. While at inference time, we enlarge the ranges of λ_1, λ_2, n to generate more testing data. The testing LR images with degradation parameters within the training ranges are considered as in-domain data, and the testing LR images with degradation parameters out of the training ranges are considered as out-domain data compared to the training data. Here, we present the comparison of in-domain and out-domain data in Tab.1 and Tab.2. We also analyze the efficiency of each method by presenting the number of parameters and FLOPs.

According to Tab.1, our method can achieve comparable results on in-domain test data with DASR even though it has only half the model size and FLOPs of DASR, which indicates the effectiveness of our method. Meanwhile, both DASR and our method can perform much better than all other methods, which indicates the advantage of degradation-aware adaptation and latent degradation representation learning for blind SR.

The performance of out-domain test data in Tab.2 also shows that our method as well as DASR can be superior to other methods. Meanwhile, with the help of the proposed ranking loss and region-aware modulation, our method achieves better results than DASR on these out-domain test data. It means that our method has a better generalization ability on LR images with unseen degradations, which have different kernels and noises compared to training data. We also present some qualitative examples in Fig.4 (a)(b) for both in-domain and

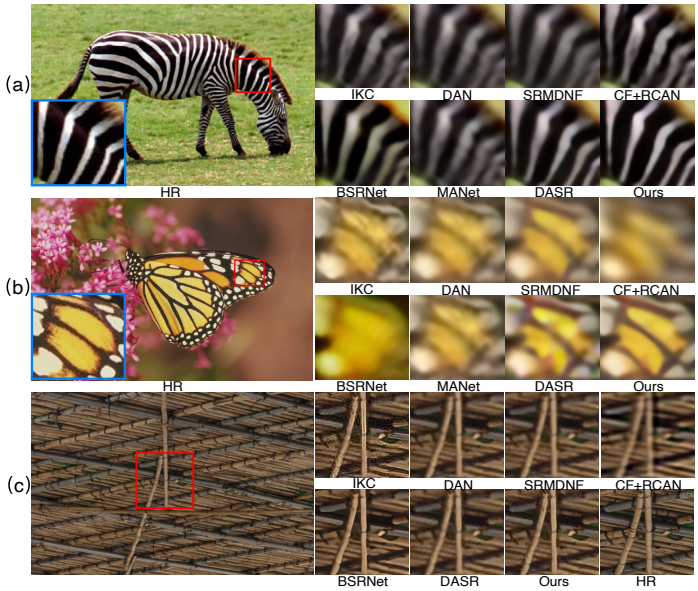

Fig. 4: Visual results (×4 SR) on (a) in-domain, (b) out-domain synthetic test data, and (c) RealSR test data. Zoom in for better visual comparison.

out-domain data. Compared to other methods, our method tends to generate clearer textures with less artifacts.

Especially, our method has a relatively compact model size and FLOPs by using IMDN as SR model, which indicates the efficiency of our method. Our FLOPs is the smallest among all methods, while our number of parameters is only slightly larger than SRMDNF, a light-weighted non-blind SR method which does not require any degradation learning. Note that CF+RCAN is larger since RCAN is a heavier SR model, CF itself has much smaller model size. However, it requires image-specific training at testing period which is different from others.

Comparison on Real-World Data. To further show the generalization ability of all methods on real-world situation, we directly test all models trained with synthetic data on RealSR without re-training or fine-tuning on any real-world data. The comparisons are presented in Tab.2. It indicates that our method can perform favorably against other methods in most cases, which proves our method can also generalize well on real-world data which is entirely different from the training data. Qualitative examples are presented in Fig.4(c).

4.4 Ablation Study

Study on Each Component. Here, we present the ablation study to show the improvement of our proposed components. It can be separated into three parts: the loss for degradation encoder; the type of modulation; and the training

Table 3: Ablation study on different components on synthetic data (2.0/5.0/90/0) and real data. The best results are in Red.

Method	Encoder		Modulation		SRNet	Set14		RealSR	
	+CL	+RL	+UM	+RM	Fix	PSNR	SSIM	PSNR	SSIM
Model-1					✓	23.47	0.6004	27.65	0.7796
Model-2	✓		✓			24.86	0.6525	27.76	0.7931
Model-3		✓	✓			24.98	0.6590	27.79	0.7991
Model-4		✓		✓	✓	24.37	0.6355	27.79	0.8001
Model-5		✓		✓		25.08	0.6602	27.84	0.8024

strategy for base SR model. Model-1 represents the original IMDN which is pre-trained on bicubicly downsampling images without using degradation encoder and modulation. For Model-2, 3, 5, we fine-tune IMDN while applying different kinds of degradation-aware modulation. We try two kinds of the losses for degradation encoder, 'CL' is the contrastive learning loss in [28] while 'RL' is the proposed ranking loss. For the type of modulation, we try 'UM', the uniform modulation in [28] where the same modulation parameter is used for features among all spatial positions, and 'RM', the proposed region-aware modulation. For Model-4, we fix the original IMDN and only train the degradation encoder as well as modulation module for degradation-aware modulation. Tab.3 shows the training strategy for each model and quantitative results on both synthetic data ($\lambda_1 = 2.0, \lambda_2 = 5.0, \theta = 90, n = 0$) and RealSR. Model-5 is our final model.

We notice that by applying different kinds of modulation, all models can achieve improvements compared to Model-1 (original IMDN). It indicates that degradation-aware modulation does improve the generalization ability of SR model. However, different kinds of degradation-aware modulation would also affect the performance. Improvements from Model-2 to Model-3 show that 'RL', ranking loss which learns the degradation degree can perform better than 'CL', which can only distinguish one degradation from the other. Meanwhile, from Model-3 to Model-5, the improvements show that 'RM', region-aware modulation which allows different regions in feature to choose different parameters for modulation is better than 'UM'. Moreover, even though Model-4 which fixes base SR model during training can achieve improvements compared to Model-1 by applying the same degradation-aware modulation as Model-5, Model-5 by fine-tuning SR model during training gains further enhancements from Model-4.

Table 4: Ablation study with different SR Net on in-domain (3.5/2.0/45/25), out-domain (4.5/5.0/120/5), and real data. The best results are in Red.

Method	IMDN						RCAN						EDSR					
	In-domain		Out-domain		RealSR		In-domain		Out-domain		RealSR		In-domain		Out-domain		RealSR	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
-bic	20.07	0.2845	22.37	0.5156	27.65	0.7796	19.81	0.2728	22.36	0.5116	27.65	0.7797	19.64	0.2574	22.35	0.5098	27.64	0.7793
-ft	24.05	0.6047	23.61	0.5873	27.66	0.7963	24.13	0.6087	23.55	0.5832	27.70	0.7921	24.21	0.6118	23.65	0.5870	27.73	0.7925
-Ours	24.19	0.6094	23.78	0.5934	27.84	0.8024	24.28	0.6133	23.77	0.5921	27.86	0.7939	24.23	0.6117	23.84	0.5952	27.87	0.7934

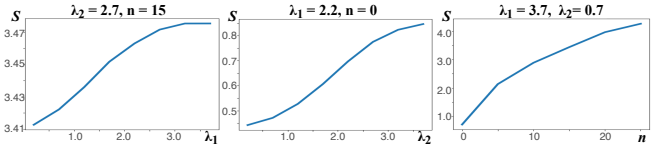

Fig. 5: Curves for ranking scores.

Study on Different SR Net. To show that the proposed structure is more flexible than DASR [28] since it can be applied to other SR models, we also try to implement the same structure to RCAN [36] and EDSR [21]. Here, the number of modulation blocks is set to 11 and 33 for 10 residual group blocks in RCAN and 32 residual blocks in EDSR. The results of in-domain, out-domain and RealSR data are shown in Tab.4. It indicates that using the proposed degradation-aware modulation on these three base SR (-Ours) gain improvements compared to their original SR models pretrained on bicubicly downsampled LR images (-bic). We also show the results of simply fine-tuning base SR models on the same training data as '-Ours' without using any degradation-aware modulation (-ft), which achieves worse results especially on out-domain data and RealSR compared to '-Ours'. It indicates that even though '-ft' uses training data with various degradations, it may still limit the generalization ability on unseen degradations without learning an informative degradation representation for applying a modulation specific to the degradation.

Study on Ranking Scores. To prove that the proposed degradation encoder with ranker can produce ranking scores with the right order. We generate a series of synthetic LR images by HR images from RealSR with degradation of using two fixed parameters while altering the third one in λ_1, λ_2 and n. We then produce their ranking scores s by our degradation encoder and ranker and draw curves as in Fig.5. It shows that by setting larger value for the unfixed parameter, the generated LR image would have larger s.

5 Conclusions

In this paper, we propose a blind SR method with degradation-aware adaptation. It applies a plug-and-play module to improve the generalization capability of an existing SR model pretrained on bicubicly downsampled LR images to real-world degradation. The proposed method consists of three components: the pretrained base SR model, a degradation encoder followed by a ranker, and a degradation-aware modulation module. The degradation encoder extracts a latent degradation representation supervised by ranking loss to estimate the degree of degradation for modulation. The degradation-aware modulation module then uses degradation representation as condition to apply a region-aware and sample-specific adaptation for the intermediate features of SR model. Our method has relatively compact model size and performs favorably against the state-of-the-art SR methods on both synthetic and real-world datasets.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European conference on computer vision (ECCV). pp. 252–268 (2018)
3. Aquilina, M., Galea, C., Abela, J., Camilleri, K.P., Farrugia, R.A.: Improving super-resolution performance using meta-attention layers. IEEE Signal Processing Letters **28**, 2082–2086 (2021)
4. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3086–3095 (2019)
5. Chen, J., Wang, X., Guo, Z., Zhang, X., Sun, J.: Dynamic region-aware convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8064–8073 (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. pp. 184–199. Springer (2014)
8. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017)
9. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. arXiv preprint arXiv:1705.06830 (2017)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
11. Gu, J., Lu, H., Zuo, W., Dong, C.: Blind super-resolution with iterative kernel correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1604–1613 (2019)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
13. Huang, Y., Li, S., Wang, L., Tan, T., et al.: Unfolding the alternating optimization for blind super resolution. Advances in Neural Information Processing Systems **33**, 5632–5643 (2020)
14. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th acm international conference on multimedia. pp. 2024–2032 (2019)
15. Hussein, S.A., Tirer, T., Giryes, R.: Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1428–1437 (2020)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)

17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: The International Conference on Learning Representations (2015)
18. Li, W., Zhou, K., Qi, L., Jiang, N., Lu, J., Jia, J.: Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. Advances in Neural Information Processing Systems **33**, 20343–20355 (2020)
19. Liang, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4096–4105 (2021)
20. Liang, J., Zhang, K., Gu, S., Van Gool, L., Timofte, R.: Flow-based kernel prior with application to blind super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10601–10610 (2021)
21. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
22. Luo, Z., Huang, Y., Li, S., Wang, L., Tan, T.: End-to-end alternating optimization for blind super resolution. arXiv preprint arXiv:2105.06878 (2021)
23. Ma, C., Tan, W., Yan, B., Zhou, S.: Prior embedding multi-degradations super resolution network. Neurocomputing (2022)
24. Rad, M.S., Yu, T., Musat, C., Ekenel, H.K., Bozorgtabar, B., Thiran, J.P.: Benefiting from bicubically down-sampled images for learning real-world image super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1590–1599 (2021)
25. Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3118–3126 (2018)
26. Soh, J.W., Cho, S., Cho, N.I.: Meta-transfer learning for zero-shot super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3516–3525 (2020)
27. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 114–125 (2017)
28. Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., Guo, Y.: Unsupervised degradation representation learning for blind super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10581–10590 (2021)
29. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1905–1914 (2021)
30. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces. pp. 711–730. Springer (2010)
31. Zhang, K., Gool, L.V., Timofte, R.: Deep unfolding network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3217–3226 (2020)
32. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4791–4800 (2021)

33. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017)
34. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3262–3271 (2018)
35. Zhang, W., Liu, Y., Dong, C., Qiao, Y.: Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3096–3105 (2019)
36. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
37. Zhou, R., Susstrunk, S.: Kernel modeling super-resolution on real low-resolution images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2433–2443 (2019)
38. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)