Metric Learning based Interactive Modulation for Real-World Super-Resolution

Chong Mou[*1,3], Yanze Wu[3], Xintao Wang[†3], Chao Dong[4,5], Jian Zhang ✉[1,2], Ying Shan[3]

[1]Peking University Shenzhen Graduate School, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]ARC Lab, Tencent PCG
[4]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[5]Shanghai AI Laboratory
eechongm@gmail.com; {yanzewu, xintaowang, yingsshan}@tencent.com;
chao.dong@siat.ac.cn; zhangjian.sz@pku.edu.cn

Abstract. Interactive image restoration aims to restore images by adjusting several controlling coefficients, which determine the restoration strength. Existing methods are restricted in learning the controllable functions under the supervision of known degradation types and levels. They usually suffer from a severe performance drop when the real degradation is different from their assumptions. Such a limitation is due to the complexity of real-world degradations, which can not provide explicit supervision to the interactive modulation during training. However, how to realize the interactive modulation in real-world super-resolution has not yet been studied. In this work, we present a **M**etric Learning based Interactive **M**odulation for **Real**-World **S**uper-**R**esolution (**MM-RealSR**). Specifically, we propose an unsupervised degradation estimation strategy to estimate the degradation level in real-world scenarios. Instead of using known degradation levels as explicit supervision to the interactive mechanism, we propose a metric learning strategy to map the unquantifiable degradation levels in real-world scenarios to a metric space, which is trained in an unsupervised manner. Moreover, we introduce an anchor point strategy in the metric learning process to normalize the distribution of metric space. Extensive experiments demonstrate that the proposed MM-RealSR achieves excellent modulation and restoration performance in real-world super-resolution. Codes are available at https://github.com/TencentARC/MM-RealSR.

Keywords: Metric Learning, Real-World Super-Resolution, Interactive Modulation, Generative Adversarial Network

1 Introduction

Image super-resolution (SR) is the task of recovering details of a high-resolution (HR) image from its low-resolution (LR) counterpart. Most SR methods [7,9,8,24,46]

[†] Project lead.   ✉ Corresponding author.
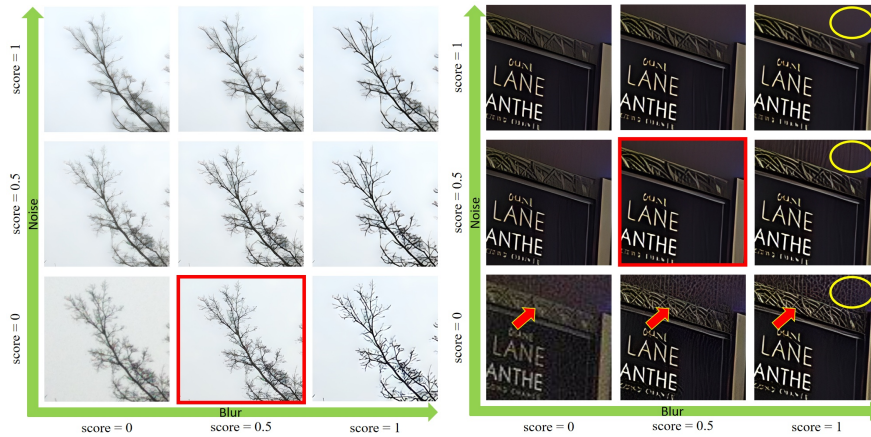[*] Chong Mou is an intern in ARC Lab, Tencent PCG.

Fig. 1. Visual results of our interactive modulation method for real-world super-resolution (×4). Best results are labeled with red boxes. **(Zoom in for best view)**

assume an ideal bicubic downsampling kernel, which is different from real-world degradations. This degradation mismatch makes those approaches unpractical in real-world scenarios. Recently, some attempts [2,43,35,27] are proposed to address real-world super-resolution (RWSR). Nevertheless, existing RWSR methods can only perform restoration with fixed one-to-one mapping. In other words, they lack the flexibility, *i.e.*, interactive modulation [12,13], to alter the outputs by adjusting different restoration strength levels.

On the other hand, several modulation-based methods [12,13,3,40] are proposed to enable controllable image restoration according to users' flavours. These modulation-based methods generally take the degradation level as a part of the network inputs, and then construct mapping between the reconstruction result and the degradation level during training. During inference, users can simply adjust the value of the input degradation level, and the network will generate reconstruction results according to the corresponding restoration strength. However, existing controllable image restoration methods can only be trained on datasets with *simple* degradation processes and *known* degradation types/levels.

In real-world scenarios, corrupted images usually contain mixture and complex degradations (*e.g.*, blur [29], noise [42] and JPEG compression). Such a complex real-world degradation process can be simulated by a random shuffle order [43] or a high-order degradation process [35]. Due to the mixtures of different degradation levels and complicated degradation types in high-order degradations, the explicit degradation levels can not reflect the actual degradation effects in the corrupted images. For example, a corrupted image is degraded by a sequence of Gaussian blur with σ_1, Gaussian noise, and Gaussian blur with σ_2. We cannot know the final equivalent blur level on this degraded image. This case becomes more complicated when the real-world degradation process involves more degradation types and more complex degradation combinations.

Therefore, existing modulation-based methods trained with known degradation types and levels can not get effective supervision in real-world settings. How to extend the interactive modulation to RWSR is worth investigating and has not yet been studied.

In this work, we present a **M**etric Learning based Interactive **M**odulation method for **Real**-World **S**uper-**R**esolution (MM-RealSR). Specifically, we propose a metric learning scheme to map the unquantifiable degradation level to a ranking score (named *degradation score* in this work) in a metric space. The degradation score generated from the metric space can reflect the relative strength of the degradation level and provide pseudo supervision to the interactive restoration mechanism. A brief illustration and comparison to the existing interactive modulation strategy for image restoration are presented in Fig. 2. To restrict the learned degradation scores to a reasonable range and normalize the distribution of metric space, we further introduce an anchor point strategy in the metric learning process. Specifically, when the degradation score is zero, the network almost learns an identity mapping, while the network has the strongest restoration ability when the degradation score is one. Equipped with such pseudo estimations for real-world degradations, our MM-RealSR can learn the controllable restoration by mapping the degradation scores to different restoration strengths. Concretely, a condition network is used to generate controllable condition vectors to adjust the restoration quality. A base network is used to perform RWSR based on the input LR image and condition vectors.

We summarize our contributions as follows. **1)** We propose a metric learning strategy to map unquantifiable degradation levels in real-world scenarios to a metric space in an unsupervised manner. We further introduce an anchor point strategy to normalize the distribution of metric space. **2)** Our proposed MM-RealSR is the first work to investigate the interactive modulation for RWSR. **3)** Extensive experiments show that the proposed MM-RealSR achieves excellent modulation (*e.g.*, Fig. 1) and restoration performance in RWSR.

2   Related Work

2.1   Image Super-Resolution

As a pioneer work, a three-convolutional-layer network is used in SRCNN [7] to learn the LR-HR mapping for single image SR. Since then, motivated by the promising performance of CNN, several CNN-based SR methods have been proposed. Realizing the importance of the network depth, [17] designed a 20-layer network with residual blocks. [49] then combined residual learning and dense connection to extend the network depth to 100 layers. To enlarge the receptive field, [23,48,30] proposed to utilize non-local self-similarity in SR. Recently, channel attention and second-order channel attention were further introduced by RCAN [47] and SAN [5] to exploit feature correlation for improved performance. Some GAN-based methods [38] were proposed to produce higher perceptual quality. Among existing methods, setting downsampling operation as a known prior is the most popular choice. However, they have poor generalization ability due
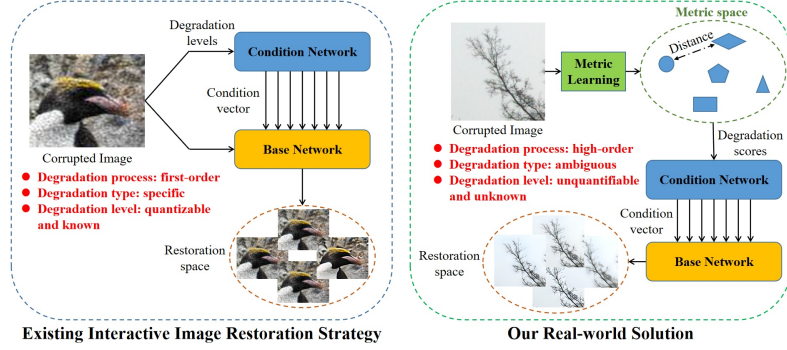
Fig. 2. Illustration of the existing interactive modulation strategy and our solution. Existing methods are confined to the sample with simple (*e.g.*, the first-order) and known degradations. Our method is an unsupervised solution to solve the more challenging real-world problem.

to the domain gap between real-world LR images and clean-LR images used for training (typically clean bicubically downsampled images).

There have been several attempts in blind SR [22,15] or real-world SR. [2,44] used a separate degradation estimation network to guide the blind SR process. Such methods usually consider simple synthetic degradations, and the accuracy of degradation representation largely affects the restoration quality. [4,39] capture real-world training pairs through specific cameras followed by tedious alignments. However, it is expensive to get such real-world training pairs, and it only works with specific imaging devices. [27,41] directly learned from unpaired data with cycle consistency loss. Nevertheless, learning fine-grained degradations with unpaired data is challenging and usually unsatisfactory. [43,35,50,14] generated training pairs as close to real data as possible, including various degradation factors and random/high-order degradation processes.

2.2   Interactive Modulation for Image Restoration

Most existing image restoration methods can only restore a specific image to a fixed result. Controllable image restoration methods aim to solve this problem by allowing users to adjust the restoration strength. Some pioneer works (*e.g.*, DNI [37] and AdaFM [12]) find that the learned filters of restoration models trained with different restoration levels are pretty similar at visual patterns. Based on this observation, DNI directly interpolated kernels to attain a smooth control of diverse imagery effects. AdaFM adopts a more efficient way by inserting AdaFM layers to change the statistics of filters. The work in CFSNet [34] adaptively learns the interpolation coefficients and uses them to couple intermediate features from the main branch and tuning branch. Recently, CResMD [13] distinguished different degradation types in the modulation process and proposed a framework that accepts both corrupted images and their degradation
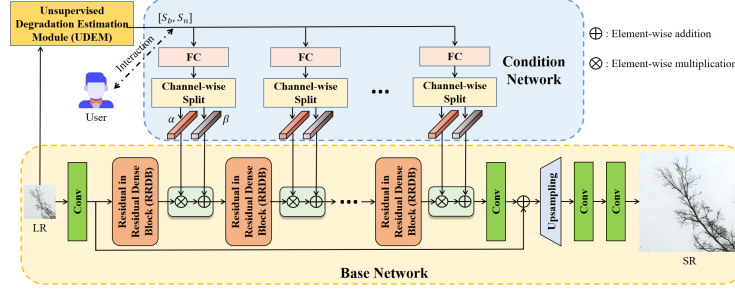
Fig. 3. Overview of our proposed metric learning based interactive modulation method for real-world super-resolution (MM-RealSR). It is composed of a base network, a condition network, and an unsupervised degradation estimation module.

information as input to realize multi-dimension modulation. Following this strategy, [3] proposed a GAN-based image restoration framework to produce higher perceptual quality with interactive modulation. However, existing modulation strategies can only be trained in a supervised manner to deal with simple and first-order degradations, which have limitations in real-world applications.

2.3 Metric Learning

Metric learning [21] aims to measure the distance among samples while using an optimal distance metric for learning tasks. Instead of providing an explicit label, metric learning compares the distance among inputs to construct a metric space. In the last few years, deep learning and metric learning have been brought together to introduce the concept of deep metric learning [25]. Most of the existing deep metric learning methods can be roughly categorized based on loss functions: 1) contrastive loss [11]; 2) triplet loss [32]; 3) margin ranking loss [19,20]. Contrastive loss and triplet loss have a similar formulation to close the distance between semantic-similar samples and enlarge the distance between dissimilar samples. Similar to the hinge loss in support vector machine (SVM) [31], margin ranking loss is designed for capturing the ranking relationship between input training samples. In this paper, we utilize the margin ranking loss to construct the metric space of degradation levels, which drives our model to predict degradation levels in an unsupervised manner.

3 Methodology

3.1 Overview

The overview of our proposed method (MM-RealSR) is presented in Fig. 3. It is composed of a base network, a condition network, and an unsupervised degradation estimation module. The condition network takes the degradation scores as inputs to generate a condition vector. During inference, we can simply
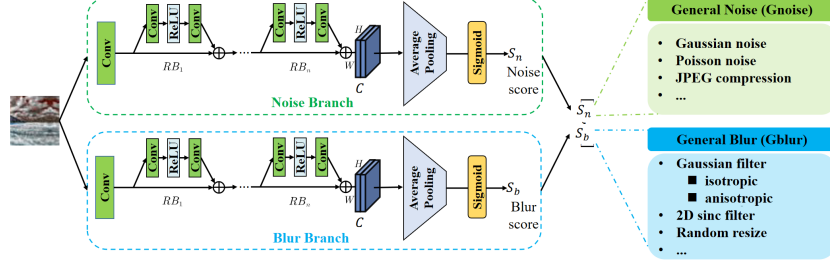
Fig. 4. The architecture of our proposed unsupervised degradation estimation module (UDEM). It is composed of two branches to estimate the degradation score of the general noise (Gnoise) and general blur (Gblur), respectively.
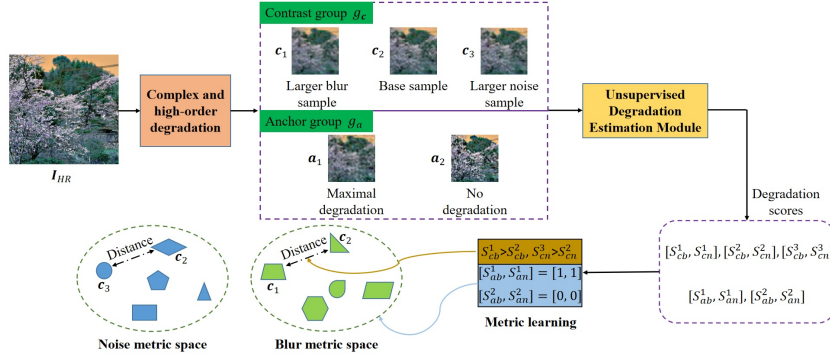

Fig. 5. The illustration of the optimization pipeline of our proposed unsupervised degradation estimation module. We utilize metric learning to construct a metric space for each degradation. We also apply two anchor points to restrict the learned degradation scores to a reasonable range and normalize the distribution of metric space.

change the degradation scores of each degradation factor to adjust the restoration strength. The base network takes condition vectors and degraded images as inputs to restore a clean result controlled by the condition vectors. The unsupervised degradation estimation module is used to map the unquantifiable degradation, which involve complex and real-world degradations, into a metric space. Each degraded image can find a degradation score in this metric space by the unsupervised degradation estimation module. Note that the degradation scores reflect the relative strength in the metric space instead of the absolute values of the degradation level. The estimated degradation scores can further provide pseudo supervision to train the condition network.

3.2 Unsupervised Degradation Estimation Module (UDEM)

Generalized degradation factors for controllable dimensions. Before we delve into the details of the unsupervised degradation estimation module

(UDEM), we first present the controllable dimensions for real-world interactive modulation. Recall that we want to investigate the interactive modulation under real-world degradations. We adopt the high-order degradation process [35,43] to simulate real-world degradations. In this setting, there are different degradation types (*e.g.*, Gaussian noise, Poisson noise, anisotropic Gaussian blur) and thus we cannot simply determine the controllable dimensions according to their types. Furthermore, due to the mixture of different degradation levels in high-order degradations, the explicit degradation levels can not reflect the actual degradation effects in the corrupted images. For example, a corrupted image is degraded by a sequence of Gaussian blur with σ_1, Gaussian noise, and Gaussian blur with σ_2. We cannot know the final equivalent blur levels on this degraded image.

Therefore, we need to re-define the controllable dimensions for real-world interactive modulation. In this paper, we propose to adopt the *general noise* (Gnoise) and *general blur* (Gblur) as the two controllable dimensions, as users usually focus on the visual effects brought by those two factors. Specifically, general noise includes Gaussian noise, Poisson noise, and JPEG compression, as they all introduce unpleasant "artifacts" into images. Note that we classify JPEG compression artifacts as Gnoise, as the JPEG blocking artifacts are visually more like noise. The general blur includes Gaussian filter (isotropic/anisotropic), 2D sinc filter [35], and random resize operation, as they all weaken image details.

Module architecture. Our proposed UDEM is used to construct two metric spaces, corresponding to Gblur and Gnoise, respectively. Each metric space can convert the unquantifiable degradation level to a degradation score. A brief illustration of our UDEM is presented in Fig. 4. Specifically, there are two prediction branches to predict the Gnoise score S_n and Gblur score S_b, respectively. Each branch comprises several residual blocks end with an average pooling operation.

Metric learning pipeline. To construct the metric space with UDEM, we propose a metric learning pipeline, as illustrated in Fig. 5. Given the HR image $\mathbf{I}_{HR}$, we first corrupt it with complex and high-order degradations as [35] to simulate the real-world degraded images. To train the metric space, we generate two groups of samples – the contrast group $\mathbf{g}_c$ and anchor group $\mathbf{g}_a$. $\mathbf{g}_c$ is used to model the distance in the metric space by comparing the sample distance among this group, and $\mathbf{g}_a$ is to normalize the distribution of the metric space. Concretely, there are three samples in the contrast group $\mathbf{g}_c = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$, in which $\mathbf{c}_1$ has larger Gblur degradation than $\mathbf{c}_2$, and $\mathbf{c}_3$ has larger Gnoise degradation than $\mathbf{c}_2$. There are two samples in the anchor group $\mathbf{g}_a = \{\mathbf{a}_1, \mathbf{a}_2\}$. $\mathbf{a}_1$ has the maximal Gnoise degradation and Gblur degradation, and $\mathbf{a}_2$ does not have any Gnoise degradation and Gblur degradation. Correspondingly, there are two groups of degradation scores generated as:

$$\{[S_{cn}^1, S_{cb}^1], [S_{cn}^2, S_{cb}^2], [S_{cn}^3, S_{cb}^3]\} = \mathcal{F}_{UDEM}(\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}) \qquad (1)$$

$$\{[S_{an}^1, S_{ab}^1], [S_{an}^2, S_{ab}^2]\} = \mathcal{F}_{UDEM}(\{\mathbf{a}_1, \mathbf{a}_2\}), \qquad (2)$$

where $\mathcal{F}_{UDEM}(\cdot)$ represents the function of UDEM.

We then adopt the margin ranking loss [19,20] to construct the Gnoise metric space and Gblur metric space. The margin ranking loss is defined as:

$$\mathcal{L}_{ML} = \frac{1}{2N} \sum_{i,j}^{N} \texttt{max}(0, \gamma - \delta(\mathbf{s}_i, \mathbf{s}_j) \cdot (\hat{\mathbf{s}}_i - \hat{\mathbf{s}}_j)) \tag{3}$$

$$\delta(\mathbf{s}_i, \mathbf{s}_j) = \begin{cases} 1, & \mathbf{s}_i \geq \mathbf{s}_j \\ -1, & \mathbf{s}_i < \mathbf{s}_j, \end{cases} \tag{4}$$

where $\mathbf{s}_i$ and $\mathbf{s}_j$ are the ground truth scores. $\hat{\mathbf{s}}_i$ and $\hat{\mathbf{s}}_j$ are predicted scores. γ is the margin parameter to constrain the distance between two samples. N refers to the number of training samples. Note that during optimization, we do not need to know the explicit values of $\mathbf{s}_i$ and $\mathbf{s}_j$ but their relative values. In our cases, two margin ranking losses are applied:

$$\mathcal{L}_{ML}^{n} = \frac{1}{N} \sum_{i}^{N} \texttt{max}(0, \gamma - (S_{cn}^{1 \circledast i} - S_{cn}^{2 \circledast i})) \tag{5}$$

$$\mathcal{L}_{ML}^{b} = \frac{1}{N} \sum_{i}^{N} \texttt{max}(0, \gamma - (S_{cb}^{3 \circledast i} - S_{cb}^{2 \circledast i})), \tag{6}$$

where $\circledast i$ represents the i-th training sample. Note that since we choose the larger Gblur and Gnoise samples for $\mathbf{c}_1$ and $\mathbf{c}_3$, we already know their relative values, thus omitting the $\delta(\cdot, \cdot)$ function.

In order to further restrict the learned degradation scores to a reasonable range and normalize the distribution of metric space, we introduce an anchor point strategy in the metric learning process. Specifically, when the degradation score is zero, the network is enforced to learns an identity mapping, while the network has the strongest restoration ability when the degradation score is one. The introduced anchor loss ($\mathcal{L}_{AC}$) is defined as:

$$\mathcal{L}_{AC} = \mathcal{L}_{AC}^{u} + \mathcal{L}_{AC}^{l} \tag{7}$$

$$\begin{cases} \mathcal{L}_{AC}^{u} = \frac{1}{N} \sum_{i}^{N} (||S_{an}^{1 \circledast i} - 1||_2^2 + ||S_{ab}^{1 \circledast i} - 1||_2^2) \\ \mathcal{L}_{AC}^{l} = \frac{1}{N} \sum_{i}^{N} (||S_{an}^{2 \circledast i}||_2^2 + ||S_{ab}^{2 \circledast i}||_2^2), \end{cases} \tag{8}$$

where $|| \cdot ||_2^2$ represents the ℓ_2 norm. $\mathcal{L}_{AC}^{u}$ and $\mathcal{L}_{AC}^{l}$ constraint the upper bound and lower bound of degradation scores.

3.3   Base Network

Our proposed MM-RealSR is a GAN-based solution for interactive RWSR. Thus, the base network consists of a generator G and a discriminator D. The generator

G accepts the input image and outputs the restored result, while the discriminator D aims to discriminate the restored result and the ground truth.

We adopt the same generator as ESRGAN [38] and Real-ESRGAN [35]. Specifically, it comprises several residual-in-residual dense blocks (RRDB), as shown in Fig. 3. At the end of the network, a pixel-shuffle upsampling and two convolution layers are applied for reconstruction. For discriminator, we apply the same architecture as Real-ESRGAN, *i.e.*, a U-Net design to better consider the realness values for each pixel. Note that the restoration performance of Real-ESRGAN is the theoretical upper bound of our proposed MM-RealSR, as MM-RealSR needs to learn another objective for interactive modulation.

3.4   Condition Network

As shown in Fig. 3, the condition network is composed of several fully-connected (FC) layers, placed at the end of each RRDB. Each FC takes degradation scores $[S_n, S_b]$ as input and generates a condition vector $Z_i \in \mathbb{R}^{B \times 2C}$, where B and C represent the batch size and the number of channels. i represents the modulation operation in the i-th RRDB. Then, a channel-wise splitting operation divides Z_i into two affine transformation parameters α_i, $\beta_i \in \mathbb{R}^{B \times C}$. The modulation operation is performed through an affine transformation with α_i and β_i:

$$\widetilde{\mathbf{F}}_i = \mathbf{F}_i \cdot \alpha_i + \beta_i, \tag{9}$$

where $\mathbf{F}_i$ and $\widetilde{\mathbf{F}}_i$ refer to the output feature in the i-th RRDB and the transformation results, respectively.

3.5   Loss Function

The proposed MM-RealSR constructs a conditional mapping in real-world super-resolution. The training objective is to drive our MM-RealSR to match this mapping relation, leading to two specific goals, *i.e.*, interactive modulation mechanism and GAN-based image restoration. The interactive modulation mechanism is optimized with the metric learning strategy, as illustrated in Sec. 3.2, including two margin ranking losses ($\mathcal{L}_{ML}^n$, $\mathcal{L}_{ML}^b$) and two anchor losses ($\mathcal{L}_{AC}^u$, $\mathcal{L}_{AC}^l$). The GAN-based image restoration is optimized with $\mathcal{L}_1$ loss, perceptual loss $\mathcal{L}_{per}$ [16], and GAN loss $\mathcal{L}_{GAN}$ [10]. The total training objective is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{GAN} + \lambda_2 \mathcal{L}_{per} + \lambda_3 \mathcal{L}_1 + \lambda_4 (\mathcal{L}_{ML}^n + \mathcal{L}_{ML}^b + \mathcal{L}_{AC}^u + \mathcal{L}_{AC}^l), \tag{10}$$

where $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are hyper parameters to balance these loss items. In our implementation, we set these four weights as $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{0.1, 1, 1, 0.05\}$.

4   Experiment

4.1   Experimental Settings

Datasets. Similar to ESRGAN [38] and Real-ESRGAN [35], we adopt DIV2K [1], Flickr2K [33] and OutdoorSceneTraining [36] datasets for training. For eval-

uation, we use the dataset provided in the challenge of Real-World Super-Resolution: AIM19 [26], and we also use a test set: RealSRSet [4] modeling DLSR camera corruptions. In all cases, we use $\times 4$ upsampling.

Training Details. Our training process consists of two stages. First, we pre-train the PSNR-oriented modulation model through Eq. 10, without perceptual loss $\mathcal{L}_{per}$ and GAN loss $\mathcal{L}_{GAN}$. In this stage, We train our system for $1000K$ iterations with the learning rate 2×10^{-4}. The well-trained model is served as the starting point for the next GAN training. In the second stage, our system is optimized with Eq. 10 for $400K$ iterations. The learning rate is fixed as 1×10^{-4}. For optimization, we use Adam [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. In both two stages, we set the batch size to 48, with the input patch size being 64.

Evaluation Framework. The main contribution of this paper is extending the interactive modulation mechanism to RWSR with the help of metric learning. The evaluation includes three aspects. **First**, we evaluate the estimation ability of our UDEM to present that the proposed metric learning strategy can well perform degradation estimation in an unsupervised manner. **Second**, we evaluate the restoration performance without manually adjusting the degradation score estimated by UDEM. This aims to present that our MM-RealSR can achieve state-of-the-art RWSR performance. **Third**, we compare our MM-RealSR with the recent modulation SR methods to present that existing modulation methods can not deal with real-world scenarios. Since AIM19 and RealSRSet datasets provide a paired validation set, we compute the learned perceptual image patch similarity (LPIPS) [45] and deep image structure and texture similarity (DISTS) [6] as the evaluation items. We also adopt a non-reference image quality assessment (IQA) metric (*i.e.*, NIQE [28]) for quantitative evaluation.

4.2   Effectiveness of the Unsupervised Degradation Estimation

As mentioned above, our degradation estimation module is optimized in an unsupervised manner to adapt the real-world degradation settings. To verify the effectiveness of this module, we first evaluate it on the synthesized data, corrupted by noise, blur, and JPEG compression degradations, respectively. For the noise degradation, we choose the Gaussian noise with the noise level range [1, 30]. For the blur degradation, we choose the Gaussian blur, with the σ range $[0.2, 3]$. For the JPEG compression, we set the quality factor ranging from 30 to 95. We divide the degradation range of these three kinds of degradations equally into 20 points and add them to natural images, respectively. The evaluation results are presented in Fig. 6, including the estimation performance of our proposed MM-RealSR and the same model trained without anchor loss (w/o anchor loss). One can see that with the increase of degradation level, the estimated degradation score increases monotonously. Note that for JPEG compression, the degradation artifact is weakened with the increase of the quality factor.

Additionally, the anchor loss allows the estimated score to have a larger dynamic range. Therefore, our unsupervised degradation estimation module can accurately estimate the relative value from the corresponding metric space. The anchor loss (Eq. 8) can constrain the distribution of points in metric space. In
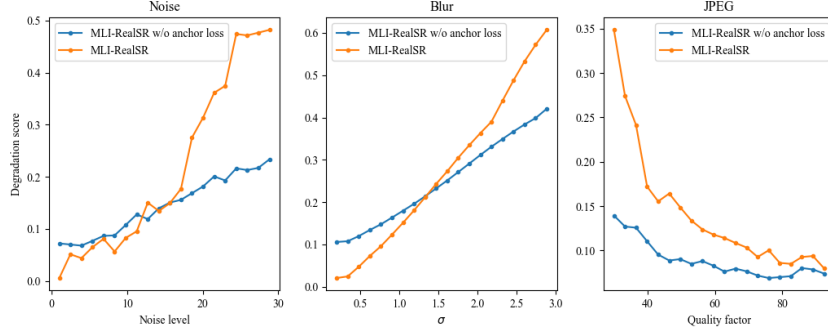
Fig. 6. The evaluation of our UDEM, including the model trained with and without the anchor loss. It shows that metric learning can generate good metric spaces to rank the degradation levels for various degradation types. The anchor loss can further normalize the space distribution and better distinguish degradation levels in a larger score range.
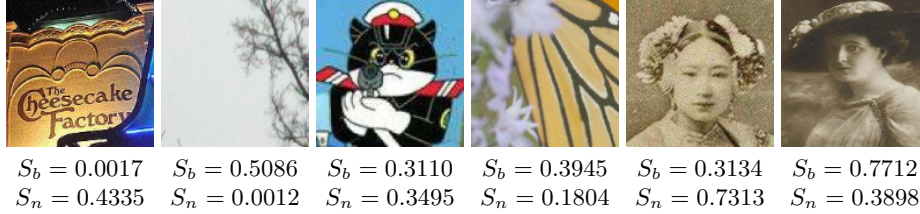

| $S_b = 0.0017$ | $S_b = 0.5086$ | $S_b = 0.3110$ | $S_b = 0.3945$ | $S_b = 0.3134$ | $S_b = 0.7712$ |
| $S_n = 0.4335$ | $S_n = 0.0012$ | $S_n = 0.3495$ | $S_n = 0.1804$ | $S_n = 0.7313$ | $S_n = 0.3898$ |

Fig. 7. Visualization of the degradation scoring ability of our unsupervised degradation estimation module (UDEM) on real-world data.

Fig. 6, there are some minor fluctuations for estimating noise and JPEG artifacts, while the curve is pretty smooth for the blur degradation. We conjecture that estimating noise is more difficult than estimating blur, as distinguishing the textures on bricks, wood, and noisy artifacts are challenging.

Apart from the quantitative evaluation, we also visualize the scoring performance on the real-world data, as presented in Fig. 7. We can find that the scoring results on the real-world data are reasonable and satisfactory. The first sample is noisy, and the second sample is blurry. We can observe that they obtain large scores on noise and blur, respectively, while they have a very low score on the other degradations. The fifth sample has large noise with some blur, and the predicted scores can also reflect such an observation. Note that the scoring for noise and blur is achieved in two separate metric spaces. Therefore, the noise score and blur score are not suitable for strict corresponding, and there might be a slight difference with visual comparison.

4.3 Evaluation for Restoration Performance

Comparison details. In this part, we fully compare our proposed MM-RealSR with recent modulation (*i.e.*, CUGAN [3]) and non-modulation methods. The

Bicubic ×4   RealSR [14]   ESRGAN [38]   BSRGAN [43]   RealESRGAN [35]   MM-RealSR
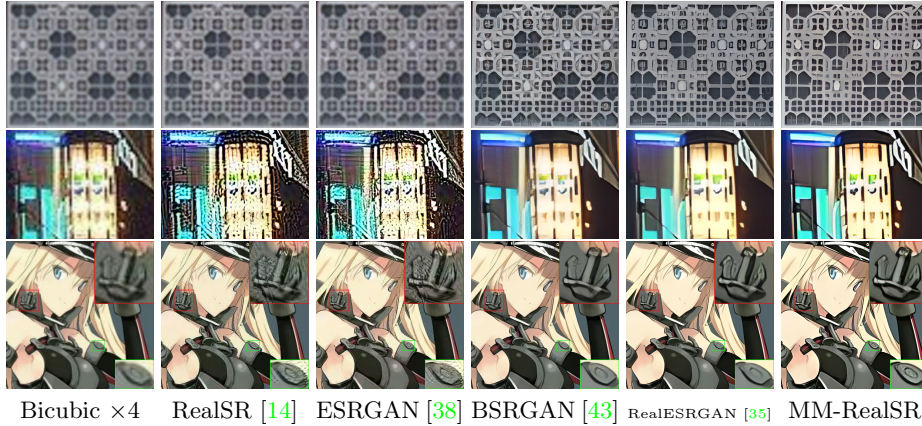
Fig. 8. Visual comparison between our proposed MM-RealSR and several recent methods on real-world super-resolution. The degradation scores are estimated by our UDEM without specific adjustment during inference.

Table 1. Quantitative comparisons (NIQE/LPIPS/DISTS) on RealSRSet [4] and AIM19 [26] test sets. Best results are **highlighted**. We compare our MM-RealSR with both modulation (Mod.) and non-modulation (Non-mod.) methods. Real-ESRGAN [35] is the upper-bound model of our MM-RealSR.

		RealSR [14]	ESRGAN [38]	CUGAN [3]	BSRGAN [43]	Upper bound [35]	MM-RealSR
	Setting	Non-mod.	Non-mod.	Mod.	Non-mod.	Non-mod.	For both
Real SRSet	NIQE ↓	8.31	8.23	6.17	4.62	**4.53**	4.62
	LPIPS ↓	0.4047	0.4054	0.3757	0.3647	0.3640	**0.3606**
	DISTS ↓	0.2201	0.2190	0.1764	**0.1535**	0.1545	0.1664
AIM19	NIQE ↓	**2.86**	3.60	4.77	3.92	3.60	3.61
	LPIPS ↓	0.5566	0.5558	0.4635	0.4048	0.3957	**0.3948**
	DISTS ↓	0.2325	0.2336	0.1944	0.1596	0.1545	**0.1507**

non-modulation methods include several well-known SR methods (*e.g.*, RealSR [14] and ESRGAN [38]) and some recent top-performing methods (*e.g.*, BSR-GAN [43] and Real-ESRGAN [35]). Real-ESRGAN is also the theoretical upper bound of our proposed MM-RealSR, as MM-RealSR needs to learn another objective for interactive modulation. Note that CUGAN must be provided with the degradation levels (S_b, $S_n \in [0, 1]$), which are not available in real-world test sets, *i.e.*, RealSRSet and AIM19. Thus, we divide the noise and blur degradation levels into 11 points uniformly distributed between 0 and 1, and then traversed all cases (121 cases in total). We select the best result (lowest LPIPS) for each image. Unlike the above two settings, our MM-RealSR can be used either with the degradation scores estimated by our UDEM module or with the user input scores. In this part, to make a fair comparison, the input degradation score of our MM-RealSR is provided by UDEM without manual adjustment.

Comparison results. The quantitative comparison is presented in Tab. 1. One can see that compared with other methods, our proposed MM-RealSR achieves excellent performance even with interactive modulation ability. Es-
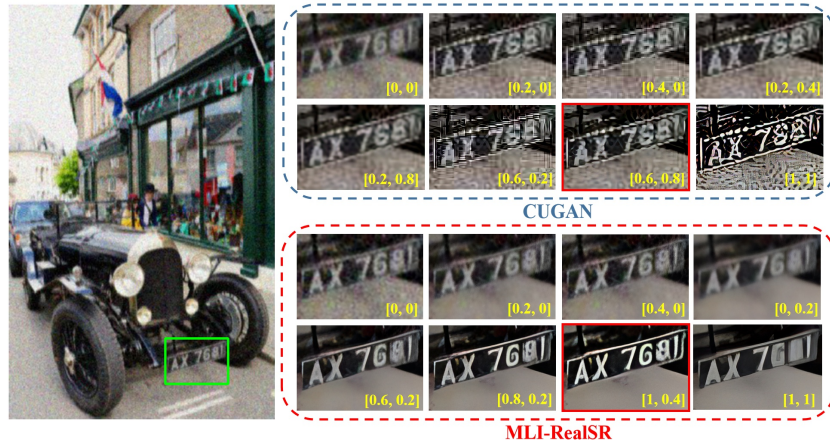
Fig. 9. The evaluation of the modulation performance on real-world super-resolution ($\times 4$). The degradation scores (*i.e.*, $[S_b,\ S_n]$) are manually selected for CUGAN [3] and our proposed MM-RealSR to present the visual quality and adjusting ability. More comparison videos are in the appendix.

pecially, our performance is comparable or even better than the upper-bound model Real-ESRGAN. Even selecting the best result for each image, the performance of CUGAN is still inferior to our proposed MM-RealSR. It demonstrates the incapability of existing modulation methods in real-world scenarios. Though RealSR and ESRGAN achieve better performance on NIQE on the AIM19 test set, the reconstruction quality of these two methods is unsatisfactory, as shown in Fig. 8. From the visual comparison in Fig. 8, we can find that our MM-RealSR can achieve better restoration quality with vivid details, demonstrating that our approach can estimate the degradation scores to guide the restoration process in real-world scenarios.

4.4   Real-World Modulation Performance

Comparison with CUGAN. In this part, We compare our method with the most recent modulation-based SR method CUGAN [3] on real-world data. We visualize the modulation process of our MM-RealSR and CUGAN in Fig. 9. The best visual result of each method is labeled with a red box We can find that the CUGAN trained with simple degradations can not deal with real-world scenarios. On the contrary, our MM-RealSR has better reconstruction quality with a larger modulation range in real-world applications.

    Further Discussions. A direct interactive modulation method for RWSR is to train several models with different degradation levels. The modulation can then be achieved by performing inference on those models and selecting the most satisfactory result. To compare our method with this case, we uniformly divide the degradations in training Real-ESRGAN [35] into four segments, corresponding to four models. In the evaluation, we select the best result (lowest LPIPS)

Table 2. The quantitative comparison between our MM-RealSR and the segmented modulation approach. The quantitative results are evaluated on RealSRSet test set [4].

	Segment-1	Segment-2	Segment-3	Segment-4	Best-LPIPS	MM-RealSR
LPIPS ↓	0.3957	0.3910	0.3950	0.4125	0.3905	**0.3662**
DISTS ↓	0.1764	0.1820	0.1910	0.1942	0.1760	**0.1632**

Bicubic ×4   Segment-1   Segment-2   Segment-3   Segment-4   MM-RealSR
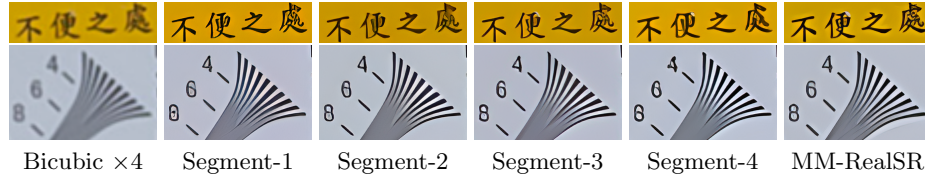
Fig. 10. Visual comparison between our MM-RealSR and the segmented modulation approach. The input degradation scores of our MM-RealSR are estimated by UDEM.

among the four results for each sample. We experiment on the RealSRSet [4] and show the results in Tab. 2 and Fig. 10. Obviously, the performance of such a simple approach is inferior to our method both quantitatively and qualitatively. Besides, this simple method only has four points for modulation.

5   Conclusion

In this paper, we present the first attempt to study the interactive modulation in real-world super-resolution through a novel metric learning based strategy. Specifically, we utilize metric learning to map the unquantifiable degradation level in real-world scenarios to a metric space and train it in an unsupervised manner. Equipped with the metric space, we can get a relative ranking score of the unquantifiable real-world degradation to guide the restoration and modulation processes. We also propose an anchor point strategy to constrain the distribution in the metric space. To adapt to the complex degradations in real-world scenarios, we also carefully design controllable dimensions in which each degradation factor has a more general degradation meaning. Extensive experiments demonstrate that our proposed MM-RealSR achieves excellent modulation and restoration performance in real-world super-resolution.

Limitations. Our MM-RealSR adopts the complex and high-order degradation synthesis [43,35] to simulate the real-world degradations. However, it is still hard to directly apply our method to train on real samples. There are also some failure cases (*e.g.*, local jitter in Fig. 6, and our UDEM regards some art effects such as bokeh as degradation). Our work makes the first attempt towards interactive modulation in RWSR, and there is still a long way with challenges.

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 (CVPRW). pp. 126–135 (2017)
2. Bell-Kligler, S., Shocher, A., Irani, M.: Blind super-resolution kernel estimation using an internal-gan. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2019)
3. Cai, H., He, J., Qiao, Y., Dong, C.: Toward interactive modulation for photo-realistic image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 294–303 (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 International Conference on Computer Vision (ICCV). pp. 3086–3095 (2019)
5. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11065–11074 (2019)
6. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(05), 2567–2581 (2022)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 184–199 (2014)
8. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (2015)
9. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: Proceedings of the International Conference on Computer Vision (ICCV). pp. 349–356 (2009)
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 (NeurIPS) **27** (2014)
11. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1735–1742 (2006)
12. He, J., Dong, C., Qiao, Y.: Modulating image restoration with continual levels via adaptive feature modification layers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11056–11064 (2019)
13. He, J., Dong, C., Qiao, Y.: Interactive multi-dimension modulation with dynamic controllable residual learning for image restoration. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 53–68 (2020)
14. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 466–467 (2020)
15. Jiang, J., Zhang, K., Timofte, R.: Towards flexible blind jpeg artifacts removal. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4997–5006 (2021)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 694–711 (2016)

17. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1646–1654 (2016)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
19. Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics **30**(1), 1–50 (2002)
20. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 662–679 (2016)
21. Kulis, B., et al.: Metric learning: A survey. Foundations and trends in machine learning **5**(4), 287–364 (2012)
22. Liu, A., Liu, Y., Gu, J., Qiao, Y., Dong, C.: Blind image super-resolution: A survey and beyond. arXiv preprint arXiv:2107.03055 (2021)
23. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). pp. 1673–1682 (2018)
24. Liu, Y., Wang, S., Zhang, J., Wang, S., Ma, S., Gao, W.: Iterative network for image super-resolution. IEEE Transactions on Multimedia **24**, 2259–2272 (2021)
25. Lu, J., Hu, J., Zhou, J.: Deep metric learning for visual understanding: An overview of recent advances. IEEE Signal Processing Magazine **34**(6), 76–84 (2017)
26. Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., Kandula, P., Suin, M., Rajagoapalan, A., Joon, N.H., et al.: Aim 2019 challenge on real-world image super-resolution: Methods and results. In: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW). pp. 3575–3583 (2019)
27. Maeda, S.: Unpaired image super-resolution using pseudo-supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 291–300 (2020)
28. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2012)
29. Mou, C., Wang, Q., Zhang, J.: Deep generalized unfolding networks for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17399–17410 (2022)
30. Mou, C., Zhang, J., Wu, Z.: Dynamic attentive graph learning for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4328–4337 (2021)
31. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines (1998)
32. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015)
33. 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 (CVPRW). pp. 114–125 (2017)
34. Wang, W., Guo, R., Tian, Y., Yang, W.: Cfsnet: Toward a controllable feature space for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4140–4149 (2019)

35. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the International Conference on Computer Vision Workshops (ICCVW). pp. 1905–1914 (2021)
36. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 606–615 (2018)
37. Wang, X., Yu, K., Dong, C., Tang, X., Loy, C.C.: Deep network interpolation for continuous imagery effect transition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1692–1701 (2019)
38. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW). pp. 0–0 (2018)
39. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 101–117 (2020)
40. You, D., Zhang, J., Xie, J., Chen, B., Ma, S.: Coast: Controllable arbitrary-sampling network for compressive sensing. IEEE Transactions on Image Processing **30**, 6066–6080 (2021)
41. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 701–710 (2018)
42. Zhang, J., Xiong, R., Zhao, C., Ma, S., Zhao, D.: Exploiting image local and nonlocal consistency for mixed gaussian-impulse noise removal. In: Proceedings of the IEEE International Conference on Multimedia and Expo. pp. 592–597 (2012)
43. 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 International Conference on Computer Vision (ICCV). pp. 4791–4800 (2021)
44. 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 (CVPR). pp. 3262–3271 (2018)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)
46. Zhang, Y., Zhang, Y., Zhang, J., Xu, D., Fu, Y., Wang, Y., Ji, X., Dai, Q.: Collaborative representation cascade for single-image super-resolution. IEEE Transactions on Systems, Man, and Cybernetics: Systems **49**, 845–860 (2017)
47. 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)
48. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. In: International Conference on Learning Representations (2019)
49. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2472–2481 (2018)
50. Zhou, R., Susstrunk, S.: Kernel modeling super-resolution on real low-resolution images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2433–2443 (2019)