

Contrastive Learning for Blind Super-Resolution via A Distortion-Specific Network

Xinya Wang, Jiayi Ma, *Senior Member, IEEE*, and Junjun Jiang, *Senior Member, IEEE*

Abstract—Previous deep learning-based super-resolution (SR) methods rely on the assumption that the degradation process is predefined (e.g., bicubic downsampling). Thus, their performance would suffer from deterioration if the real degradation is not consistent with the assumption. To deal with real-world scenarios, existing blind SR methods are committed to estimating both the degradation and the super-resolved image with an extra loss or iterative scheme. However, degradation estimation that requires more computation would result in limited SR performance due to the accumulated estimation errors. In this paper, we propose a contrastive regularization built upon contrastive learning to exploit both the information of blurry images and clear images as negative and positive samples, respectively. Contrastive regularization ensures that the restored image is pulled closer to the clear image and pushed far away from the blurry image in the representation space. Furthermore, instead of estimating the degradation, we extract global statistical prior information to capture the character of the distortion. Considering the coupling between the degradation and the low-resolution image, we embed the global prior into the distortion-specific SR network to make our method adaptive to the changes of distortions. We term our distortion-specific network with contrastive regularization as CRDNet. The extensive experiments on synthetic and real-world scenes demonstrate that our lightweight CRDNet surpasses state-of-the-art blind super-resolution approaches.

Index Terms—Blind super-resolution, contrastive learning, deep learning, image super-resolution (SR).

I. INTRODUCTION

SINGLE image super-resolution (SISR) aims to restore a high-resolution (HR) image from a given low-resolution (LR) observation. This task is a prerequisite of many applications, including medical diagnosis and video surveillance, to name a few. As an inverse problem, the SISR task is coupled with the degradation process and is highly ill-posed, which requires further study in low-level computer vision. A popular strategy for solving the SISR problem is to construct con-

Manuscript received April 15, 2022; revised June 29, 2022; accepted July 14, 2022. This work was supported by the National Natural Science Foundation of China (61971165), and the Key Research and Development Program of Hubei Province (2020BAB113). Recommended by Associate Editor Wenguan Wang. (*Corresponding author: Jiayi Ma*.)

Citation: X. Y. Wang, J. Y. Ma, and J. J. Jiang, “Contrastive learning for blind super-resolution via a distortion-specific network,” *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 1, pp. 78–89, Jan. 2023.

X. Y. Wang and J. Y. Ma are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: wangxinya@whu.edu.cn; jyma2010@gmail.com).

J. J. Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: jiangjunjun@hit.edu.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JAS.2022.105914

volution neural networks with the assumption that the degradation process is predefined and fixed, e.g., bicubic downsampling [1]–[11]. In this way, a great number of HR-LR image pairs are obtained by using predefined degradation to drive the sophisticated networks. Thus, evaluated on the same degradation, these methods have achieved increasing objective SR performance, i.e., peak signal-to-noise ratio (PSNR), and structure similarity (SSIM). However, in real-world scenarios, the degradation process is usually unknown and diverse among images due to imaging and transmission. The mismatch between the simplistic degradation assumption in existing SISR algorithms and the intrinsic degradations of real inputs brings about the difficulty for these data-driven SISR methods in real applications. Consequently, their performance suffers from severe deterioration even if the degradation is slightly different. Therefore, recently, researchers turn to pay more attention to blind SR, where the true degradation process is unknown.

For blind SR task, most existing methods generate LR images from HR inputs via the following mathematically model:

$$\mathbf{y} = (\mathbf{x} \otimes \mathbf{k}) \downarrow_s + \mathbf{n} \quad (1)$$

where \mathbf{y} is the LR image, \mathbf{x} is the corresponding HR image, \otimes represents the blur operation with the kernel \mathbf{k} , \downarrow_s denotes the downsampling process with a scale factor s , and \mathbf{n} denotes the noise level. In this case, several learning-based methods cast into a two-step solution, i.e., estimating the blur kernel from the given LR input and then recovering the SR result based on the estimated kernel, or iteratively optimize the kernel and the SR estimation [12]–[14]. Nevertheless, it is still a challenge to customize an effective and efficient method for practical use. The challenges mainly come from three aspects. Firstly, the iterative optimization process will inevitably increase the computational load. Secondly, the mismatch of the blur kernel would further contribute to the SR reconstruction error. For example, if the real blur kernel is spatially variant in real scenarios, the performance of these two-step methods would degrade badly. Thirdly, the generated images regularized by the pixel loss would be over-smoothed.

To address the above three challenges, instead of estimating the kernel, we consider the global statistical information of the distorted LR image as prior information and impose the contrastive regularization on the reconstructed results to solve the blind SR problem, which is implemented by a distortion-specific SR network with contrastive regularization, termed as CRDNet. Inspired by [15], due to the presence of distortions, scene statistics of locally normalized luminance coefficients

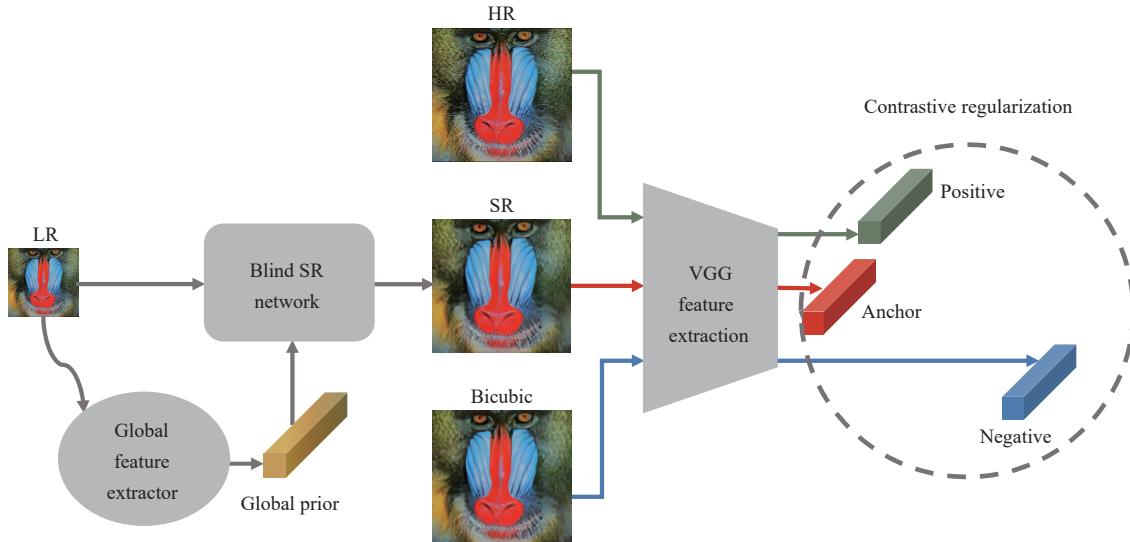


Fig. 1. The whole framework of our proposed method.

can be used to quantify potential losses of “naturalness” in the image. In addition, equipped with global characteristic statistical properties, mean subtracted contrast normalized (MSCN) coefficients can greatly reduce dependencies between neighborhoods, which is complementary to the features captured in the neural network. Based on this observation, refraining from an explicit characterization of degradations, we introduce the MSCN coefficients into the network to extract the global prior, aiming to capture the distortion-oriented global statistical properties. Specifically, we embed the global prior into both shallow and deep features via a tailored statistical feature extractor. In this way, our blind SR network would be adaptive to the changes of distortions. Note that, our method does not require the iterative process of generating the SR results with degradation kernels, which avoids the accumulation of estimation errors. In addition, there is no need to provide ground-truth kernels for supervised learning in our proposed method. To realize the best trade-off between performance and parameters, we also design a compact blind SR network by adopting the up-and-down strategy to make dense convolution calculation in the low-dimensional space. The information loss from the reduction of resolution can be compensated by a long skip connection. Compared to these two-step methods involving kernel estimation, our method is simple and time-saving.

In order to alleviate the over-smooth result caused by the pixel loss, we impose the contrastive regularization on the SR images by the triple loss. As illustrated in Fig. 1, we regard the restored SR prediction generated by the blind SR network as the anchor and its corresponding HR image (i.e., ground truth), and the bicubic upsampling result of the input LR image as the positive, and negative, respectively. We expect that the contrastive learning would pull the prediction closer to the HR image and push the prediction farther away from the bicubic upsampling result in the feature domain. Therefore, the contrastive regularization constrains the target images into the closed upper and lower bounds by contrastive learning, which benefits blind SR prediction sharper and

clearer. Furthermore, since it can be removed for testing, the contrastive regularization improves the performance for blind image SR without increasing extra computation/parameters during inference.

In summary, our contribution in this paper includes the following three aspects. First, we propose a novel CRDNet to effectively produce high-quality super-resolved images by contrastive regularization and compact distortion-specific SR network. Our CRDNet realizes the best parameter-performance trade-off, compared to the state-of-the-art methods. Second, we propose a novel contrastive regularization without extra computation to generate more satisfactory results. Third, the extracted global prior could capture the nature of degradation to make the SR network sensitive to the distortion, which further improves the performance of the network.

II. RELATED WORK

A. Super-Resolution for Bicubic Downsampling

With the progress of deep learning technology, learning-based algorithms have dominated the SR field because of the outstanding performance. Usually, learning-based SR methods train on plenty of paired HR and LR images that are difficult to acquire. Therefore, previous works predefined the degradation progress as bicubic interpolation to synthesize LR images from corresponding HR images. In this way, Dong *et al.* [1] pioneered three convolutional layers to solve the SR problem successfully, which achieved better performance than traditional SR methods. After that, many strategies have been proposed to design sophisticated networks for the image SR task, such as residual learning [2]–[4], attention mechanism [8], [16], and generative adversarial network [17], [18]. Besides, some works were devoted to exploring efficient models to reduce the computational cost and time. Hui *et al.* [7] proposed a novel information distillation network for real-time reconstruction. Lan *et al.* [19], [20] integrated multiscale correlation learning and nonlocal operations to enhance the representational capability with a reasonable number of parameters.

Although the aforementioned methods have obtained remarkable quantitative or qualitative performance under the bicubic downsampling assumption, it is still difficult for SR methods to be applied in real scenarios. Since the real degradation processes are usually unknown and more complicated than the bicubic interpolation, there exists a domain gap between the real world and synthesized data. Therefore, when SR methods designed for bicubic downsampling are applied to real data, they inevitably give less pleasing results [8], [21].

B. Super-Resolution for Multiple Degradations

To cope with various degradations, some methods recently work on multiple degradations to solve the non-blind SR problem. Zhang *et al.* [22] first concatenated degradation map with the LR image as prior inputs to generate the SR result correlated to the real blur kernel and noise level. Inspired by zero-shot learning, Shocher *et al.* [23] constructed a small image-specific CNN that is optimized in test phase to exploit the internal recurrence information of a neural image. To reduce the number of iterations, a meta-learning strategy is utilized in MZSR [24] for making the network adaptive to a specific degradation within a few steps. Recently, Zhang *et al.* [25] interpreted the SR problem as maximum a posteriori estimation (MAP) optimization and alternately solved a data subproblem and a prior sub-problem by an unfolding SR network. To enhance off-the-shelf deep SR network, Hussein *et al.* [26] utilized a closed-form filter to correct an LR input in line with the one produced by bicubic degradation. Zhou *et al.* [27] designed a general framework to exploit scale-related features among the multiple tasks. Although important advances have been achieved by these SR methods, as pointed out in [12], the SR results of the aforementioned methods are sensitive to the provided blur kernel. When the input kernel deviates from the predefined distributions, none of these multiple degradation methods could live up to our expectations.

C. Blind Super-Resolution

Since the degradation is regarded as a significant input for the non-blind SR methods, early, previous solutions for blind SR naturally combined a kernel-estimation algorithm with a non-blind SR method. As a result, the kernel prediction task is critical for blind SR. Michaeli and Irani [28] pioneered to estimate the blur kernel by exploring the internal patch recurrence. In KernelGAN [13], the blur kernel is explicitly extracted from a generative network and the discriminator is used to verify whether the distribution of generated LR image is consistent with the original input. By introducing the flow prior for kernel learning, Liang *et al.* [29] design an invertible mapping to generate reasonable kernel initialization. Although combining the advanced kernel estimation method with the non-blind SR method could obtain better SR results, the SR performance highly relies on degradation estimation. Thereby, the errors in kernel estimation could inevitably cause unexpected artifacts to the SR predictions. To overcome the defect, Gu *et al.* [12] developed an iterative kernel correction (IKC) method to correct the estimated kernel by observing generated SR results. Luo *et al.* [14] unfolded the alternating optimization (DAN) for predicting the degradation the SR image

simultaneously. In [30], a degradation representation has been learned for blind SR in an unsupervised way. However, degradation estimation could cause extra computational load for the SR task and the accumulative estimation error would deteriorate the reconstruction performance. Instead, our proposed method refrains the kernel estimation and resorts to extracting the global prior to capture the changes of degradation.

III. PROBLEM FORMULATION

According to (1), given the LR image \mathbf{y} , there are two variables that need to be determined. Thereby, existing end-to-end blind SR methods recover clear images by minimizing image reconstruction loss, kernel estimation loss with the regularization term simultaneously, which can be formulated as

$$\arg \min_{\theta_1, \theta_2} \|\mathbf{x} - H(\mathbf{y}; \theta_1)\| + \|\mathbf{k} - M(\mathbf{y}; \theta_2)\| + \lambda \phi(H(\mathbf{y}; \theta_1)) \quad (2)$$

where \mathbf{x} is the corresponding HR image, \mathbf{k} is the ground-truth kernel, $H(\cdot; \theta_1)$ is the SR network with parameter θ_1 , $M(\cdot; \theta_2)$ is the kernel prediction network with parameter θ_2 , and $\phi(\cdot)$ is regularization term with a penalty parameter λ . Consequently, previous methods adopt the two-step or alternative strategies to optimize (2) and the introduction of the blur kernel would inevitably cause the heavier computational complexity.

In our method, refraining from an explicit blur kernel, we customize global prior information to capture the characterization of distortion, which can be expressed as

$$\arg \min_{\theta_1} \|\mathbf{x} - H(\mathbf{y}, \mathbf{z}; \theta_1)\| + \lambda \phi(H(\mathbf{y}, \mathbf{z}; \theta_1)) \quad (3)$$

in which \mathbf{z} is the global prior captured by our global feature extractor. Besides, differently from the previous regularization, we propose a contrastive regularization to improve the quality of the reconstructed images.

IV. OUR PROPOSED METHOD

Our blind SR network consists of a global feature extractor and a distortion-specific SR network, as shown in Fig. 1. First, the LR input is sent into the global feature extractor to obtain a global prior. Differently from the estimated kernel iteratively generated with SR result, the global prior is expected to capture the distortion-oriented properties without being supervised by the ground truth kernel [12], [14]. Then, this global prior is incorporated into the distortion-specific SR network to produce the super-resolved result, which makes our method adaptive to the changes of distortions.

A. Global Prior

Reviewing the visual science literature, the regularity of natural scene statistics has been well established in the spatial domain as well as the wavelet domain [31], [32]. As claimed in [15], the presence of distortions would influence the statistical properties of natural images and quantifying naturalness will make it possible to capture the distortion properties. Therefore, we use the proposed mean subtracted contrast normalized (MSCN) coefficients in [15] to capture features correlated well with distortions.

To calculate MSCN coefficients, we first obtain locally normalized luminance via local mean subtraction and divisive

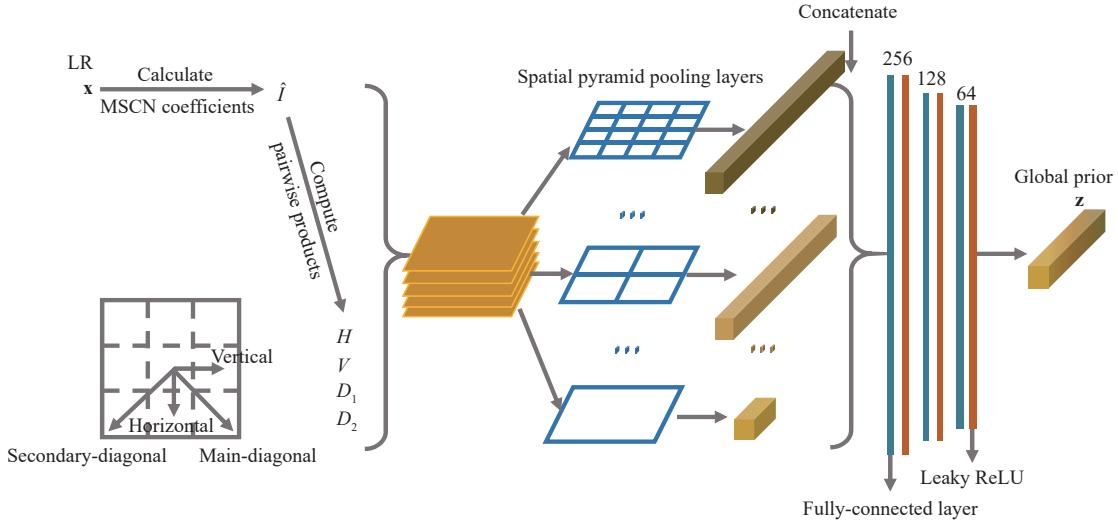


Fig. 2. The detailed operations in global feature extractor.

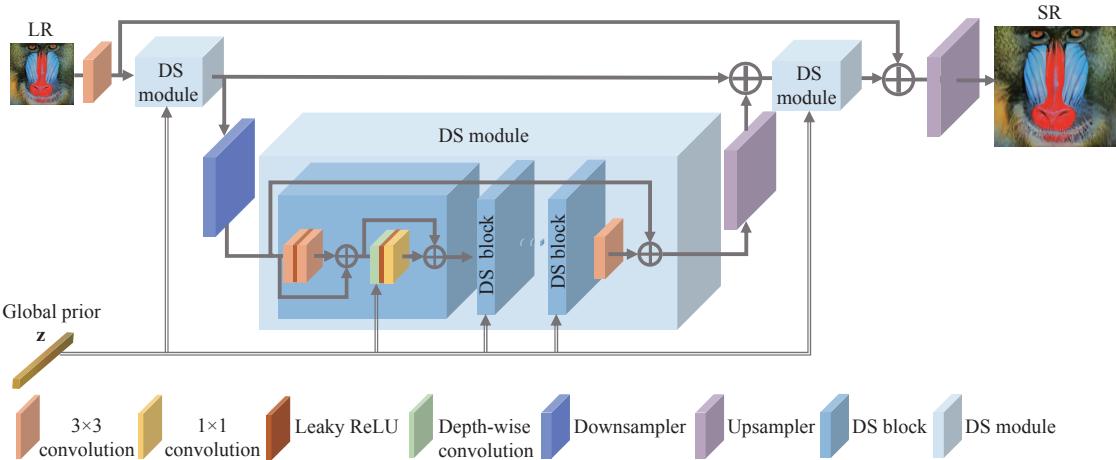


Fig. 3. The whole structure of the distortion-specific SR network at the scale factor 4.

normalization, which has a decorrelating effect. Such an operation may be applied to a given intensity image $I(i, j)$ to produce

$$\hat{I}(i, j) = \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + C} \quad (4)$$

where $\mu(i, j)$ is the local mean, $\sigma(i, j)$ is the local variance and $C = 1$. While MSCN coefficients are more homogenous for pristine images, the signs of adjacent coefficients also present a regular structure, which is disturbed by the distortion. Referring to [15], we also model this structure by the empirical distributions of pairwise products of neighboring MSCN coefficients along four directions: horizontal (H), vertical (V), main-diagonal (D_1) and secondary-diagonal (D_2), as illustrated in Fig. 2. Specifically,

$$\begin{cases} H(i, j) = \hat{I}(i, j)\hat{I}(i, j+1) \\ V(i, j) = \hat{I}(i, j)\hat{I}(i+1, j) \\ D_1(i, j) = \hat{I}(i, j)\hat{I}(i+1, j+1) \\ D_2(i, j) = \hat{I}(i, j)\hat{I}(i+1, j+1). \end{cases} \quad (5)$$

However, instead of modeling distorted image statistics by

Gaussian distribution, we concatenate these features (\hat{I}, H, V, D_1, D_2) and feed them into the neural network to extract more compact vectors to be well correlated with distortions. As illustrated in Fig. 2, these features are sent into five spatial pyramid pooling layers (SPP), following by three fully-connected layers to extract global statistical properties, which can be expressed as:

$$\mathbf{z} = E([\hat{I}, H, V, D_1, D_2]) \quad (6)$$

in which E represents the operation of these layers. The global prior is expected to capture the character of the degradation to make the SR network distortion-specific. This global feature extractor is trained together with the blind SR network. As a result, the generated global feature \mathbf{z} would be well adapted for the blind SR task.

B. Distortion-Specific SR Network

Incorporated with prior information extracted from the MSCN coefficients, a distortion-specific SR network is designed to super-resolve the low-resolution input, as illustrated in Fig. 3. In the feature extraction phase, the distortion-specific module (DS module) is deployed as the building part

and the up-and-down sampling strategy is involved in dealing with the complex degradation. Our CRDNet network consists of three DS modules, with each module comprising several DS blocks. In general, most of the existing SR methods tend to extract features at the original resolution to avoid loss of information, whereas, considering the complicated degradation, we engage the up-and-down sampling strategy to produce clearer results. After the first DS module, we downsample the features with the scale of 2, reducing the resolution in half, which can be formulated as

$$F_2 = \mathcal{D}(\mathcal{H}_{DS}(F_1)) \quad (7)$$

where F_1 denotes the input feature of the first DS module with the operation \mathcal{H}_{DS} . \mathcal{D} represents the downsampling operation that could reduce the influence of noise and make the network obtain a larger receptive field. Thus, the second DS module takes F_2 as the input, performing in a lower resolution. Then, we upsample the features to the original resolution to generate the input feature F_3 of the last DS module

$$F_3 = \mathcal{U}(\mathcal{H}_{DS}(F_2)) \quad (8)$$

where \mathcal{U} represents the upsampling operation. Subsequently, the third DS module conducts at the original resolution. The residual learning is used for alleviating the information loss to produce the feature (F_R) before reconstruction, which can be expressed as

$$F_R = \mathcal{H}_{DS}(F_3) + F_1. \quad (9)$$

The up-and-down strategy is implemented by the shuffle pixel layer [33] and its inverse version.

Within each DS module, we embed the global prior feature \mathbf{z} into each DS block. Since the MSCN maps exhibit a largely homogeneous appearance with a few low-energy residual object boundaries [15], we expect that prior feature \mathbf{z} could capture statistical property changes made by the presence of distortion. Therefore, integrated with the prior feature, the SR network could be aware of the distortion, adaptively handling various degradation.

Specifically, taking the n -th DS block for example, we first deploy a simplified residual block for the input feature f_{in}^n

$$f_{mid}^n = \mathcal{H}_{res}(f_{in}^n) \quad (10)$$

in which \mathcal{H}_{res} expresses the simplified residual block to produce the middle feature in the DS block. Considering the global nature of the prior features extracted from the MSCN coefficients, we integrate the prior features into the network in the form of a convolution kernel. According to different pictures corresponding to different priors, the depth-wise convolution followed by the 1×1 convolution is used to make the features image-specific and local residual is also added back to ease the learning difficulty, which is

$$f_{out}^n = \mathcal{H}_{1 \times 1}(\mathcal{H}_{d-w}(f_{mid}^n)) + f_{mid}^n \quad (11)$$

where \mathcal{H}_{d-w} and $\mathcal{H}_{1 \times 1}$ represent the depth-wise convolution and the 1×1 convolution, respectively.

C. Contrastive Regularization

Most existing blind SR algorithms based on CNN are dedicated to minimizing the pixel-level reconstruction loss, such

as $L1$ and $L2$. Consequently, the reconstruction results tend to be over-smooth, and thus, we try to impose an extra regularization on the SR results to alleviate the issue.

Inspired by contrastive learning, our method introduces contrastive regularization to help improve the quality of the reconstructed image. Generally, contrastive learning attempts to acquire a representation that distinguishes the target sample from other samples. Specifically, contrastive learning expects that the target sample in the representation space can be as close as possible to the positive sample and far away from the negative sample. Therefore, for the blind SR problem, there are two aspects that need to be considered: how to construct positive and negative samples and how to define the metric space. There is no doubt that the HR image should be the positive sample that the reconstruction result needs to be close to. In theory, other different images can be treated as negative samples. However, in order to temper the over-smoothness of the reconstructed image instead of making the reconstruction result distinguishable, we regard the result of bicubic upsampling as the negative sample. Since the image SR task early attempts to achieve the inverse process of bicubic downsampling, the result of bicubic upsampling can be regarded as the hardest sample to some extent. In our method, owing to the superiority of the perceptual loss defined in the VGG feature representation, we project the samples to the latent space through the VGG network and constrain the distance of these samples in the feature space, which is investigated to regularize the network learn better feature space in [34] and [35]. Thus, the objective of our blind SR network can be represented as

$$\arg \min_{\theta_1} \|\mathbf{x} - \mathbf{H}(\mathbf{y}, \mathbf{z}; \theta_1)\| + \lambda \phi(\mathbf{G}(\mathbf{x}), \mathbf{G}(\mathbf{H}(\mathbf{y}, \mathbf{z}; \theta_1)), \mathbf{G}(\mathbf{b})) \quad (12)$$

where the first term is the reconstruction loss for constraining the recovered image and its corresponding HR image in the pixel level. We employ $L1$ loss, since it could obtain better performance compared to $L2$ loss [5]. The second term is the contrastive regularization among HR image \mathbf{y} , restored image $\mathbf{H}(\mathbf{y}, \mathbf{z}; \theta_1)$, and the bicubic upsampling result \mathbf{b} under the same latent feature space, which works on pulling the restored image to its ground-truth image and pushing the restored away to the bicubic upsampling result. λ is a hyperparameter to balance the reconstruction term and contrastive regularization. From the perspective of classification, considering the target sample, the positive and negative examples belong to the same category, so we employ the triple loss here. It is required that for the same class of positive and negative examples, the target sample should be at least away from the negative sample than the positive sample above a threshold. Therefore, the overall loss function can be formulated in detail as

$$\begin{aligned} & \arg \min_{\theta_1} \|\mathbf{x} - \hat{\mathbf{x}}\|_1 + \lambda \max(d(\mathbf{G}(\mathbf{x}), \mathbf{G}(\hat{\mathbf{x}})) \\ & \quad - d(\mathbf{G}(\mathbf{x}), \mathbf{G}(\mathbf{b})), \tau, 0) \end{aligned} \quad (13)$$

where $\hat{\mathbf{x}}$ represents the restored image in brief, that is, $\hat{\mathbf{x}} = \mathbf{H}(\mathbf{y}, \mathbf{z}; \theta_1)$ and $d(\cdot, \cdot)$ is the $L2$ distance between two samples. Adopting the bicubic upsampling result as the negative to constrain the solution space, the prediction of our method could be sharper and clearer as well as perceptual superior.

TABLE I
QUANTITATIVE COMPARISON WITH SOTA SR METHODS ON NOISE-FREE DEGRADATIONS WITH ISOTROPIC GAUSSIAN KERNELS. THE BEST RESULTS ARE INDICATED IN BOLD

Method	Scale	Set5		Set14		BSD100		Urban100		Manga109	
		0.6	1.8	0.6	1.8	0.6	1.8	0.6	1.8	0.6	1.8
Kernel width		0.6	1.8	0.6	1.8	0.6	1.8	0.6	1.8	0.6	1.8
Bicubic		32.66/0.9167	28.25/0.8183	29.53/0.8471	26.10/0.7153	28.88/0.8171	26.06/0.6795	26.17/0.8152	23.22/0.6724	29.62/0.9183	25.00/0.8030
RCAN		35.91/0.9505	28.50/0.8269	32.31/0.9018	26.32/0.7264	31.15/0.8780	26.25/0.6906	29.80/0.9039	23.44/0.6857	34.68/0.9657	25.30/0.8146
IKC	×2	37.41/0.9579	34.66/0.9285	33.42/0.9136	30.67/0.8556	31.92/0.8918	29.69/0.8238	30.91/0.9143	27.03/0.8290	37.75/0.9750	31.76/0.9336
DAN		37.82/0.9584	35.78/0.9380	33.33/0.9129	31.82/0.8745	32.06/0.8947	30.52/0.8527	31.14/0.9157	29.04/0.8756	38.09/0.9757	35.28/0.9580
DASR		37.43/0.9558	35.49/0.9377	32.95/0.9059	31.61/0.8717	31.79/0.8875	30.58/0.8507	30.72/0.9098	29.02/0.8748	37.72/0.9751	34.84/0.9347
CRDNet		38.13/0.9590	36.04/0.9392	33.64/0.9149	32.07/0.8766	32.27/0.8976	30.82/0.8586	31.35/0.9186	29.27/0.8784	38.14/0.9762	35.30/0.9582
Kernel width		1.2	3.6	1.2	3.6	1.2	3.6	1.2	3.6	1.2	3.6
Bicubic		27.69/0.7904	24.44/0.6774	25.59/0.6820	23.24/0.5811	25.58/0.6461	23.80/0.5570	22.72/0.6341	20.83/0.5305	24.27/0.7661	21.63/0.6668
RCAN		30.26/0.8636	24.66/0.6883	27.47/0.7512	23.41/0.5896	26.89/0.7067	23.93/0.5640	24.71/0.7399	20.98/0.5400	27.49/0.8640	21.83/0.6767
IKC	×4	31.75/0.8870	30.26/0.8585	28.37/0.7709	26.63/0.7100	27.42/0.7240	26.41/0.6854	25.62/0.7676	24.07/0.7024	29.41/0.8921	26.61/0.8265
DAN		31.97/0.8920	30.94/0.8663	28.44/0.7714	27.68/0.7378	27.52/0.7289	26.95/0.6956	25.63/0.7678	24.98/0.7320	30.75/0.9101	29.27/0.8824
CRDNet		32.22/0.8936	31.62/0.8874	28.65/0.7800	28.12/0.7680	27.65/0.7348	27.43/0.7310	26.20/0.7869	25.08/0.7398	30.82/0.9118	29.31/0.8829

D. Implementation Details

In our CRDNet network, the convolutional layers for feature extraction are equipped with 64 filters, except for the last layer, since the output channel number of the last layer is three. The kernel size of the depth-wise convolution is 8×8 . To keep the size of the feature map unchanged, the zero-padding strategy is used for these convolutional layers. For the upsampler employed in our network, we use the shuffle pixel layer proposed in [33] and the downsample is the inverse version of the shuffle pixel layer. For the loss function, we empirically set $\lambda = 0.3$. Our CRDNet method is implemented by PyTorch and with one NVIDIA TITAN RTX GPU. For the training phase, we choose a mini-batch size of 32 with random rotation augmentation. The LR patches are of 48×48 sizes. The model is trained by Adam optimizer with exponential decay rates β_1 and β_2 equal to 0.9 and 0.999, respectively. The initial learning rate is set to 1×10^{-4} and decays by 10 times every 150 epoches while the total epoch is 600.

V. EXPERIMENTS AND RESULTS

To fully investigate the proposed method, we conduct extensive experiments on both synthetic and real images. For synthetic images, we evaluate quantitative and qualitative results under different settings and perform an ablation study to analyze the proposed method. For real images, we provide a qualitative comparison to show the advantage of our method.

A. Experiments on Noise-Free Degradations With Isotropic Gaussian Kernels

We first perform the experiments on noise-free degradations with isotropic Gaussian kernels following [12]. Following the settings in [12], we collect 800 images in DIV2K [36] and 2650 training images in Flickr2K [37] for the training dataset, and evaluate the result on four benchmark datasets including Set5 [38], Set14 [39], BSD100 [40], Urban100 [41] and Manga109 [42]. We synthesized LR images according to (1) for training and testing. In this setting, the size of the Gaussian kernel is set as 21×21 . During training, the kernel

width is uniformly sampled in $[0.2, 2.0]$ and $[0.2, 4.0]$ for scale factors 2 and 4 respectively. For testing, the HR images are first processed by the selected blur kernels at two widths and then directly downsampled to form synthetic test images. We compare our CRDNet with several state-of-the-art SR methods, including Bicubic, RCAN [16], IKC [12], DAN [14] and DASR [30]. RCAN is a state-of-the-art (SOTA) SISR method specially designed for bicubic degradation. IKC [12] is the representative two-step SR method that only considers degradations with isotropic Gaussian kernels. DAN [14] and DASR [22] are the SOTA blind SR method considering both isotropic/anisotropic Gaussian kernels and noises.

Quantitative results in terms of PSNR and SSIM are exhibited in Table I while visualization results are shown in Fig. 4. It can be noticed from Table I that when the test degradations are different from the pre-defined one, the RCAN method does not achieve considerable performance even if it performs well in bicubic-downsampling settings. Considering the estimation of SR image and degradation with the iterative scheme, IKC and DAN have achieved better performance. However, according to the testing time provided in Table II, they are time-consuming due to the iterations. Although the degradation is not accurately estimated in DASR, the involved degradation representation greatly increases the complexity of the network, which can be seen in Table II. Based on the experimental results, equipped with global prior and contrastive learning, the proposed CRDNet outperforms the others in terms of PSNR and SSIM values. In addition, our method is quite stable at different kernel widths, which is demonstrated in Fig. 5. Although the RCAN method achieves better results when the kernel width is extremely small, the performance of our proposed CRDNet shows a more stable trend. This is mainly because we embed the global prior in both low-level and high-level features to make the network well adaptive for different degradations. In the meanwhile, contrastive learning with the bicubic estimation could promote our method to be perceptually superior.

Visualization reconstructed results generated by different

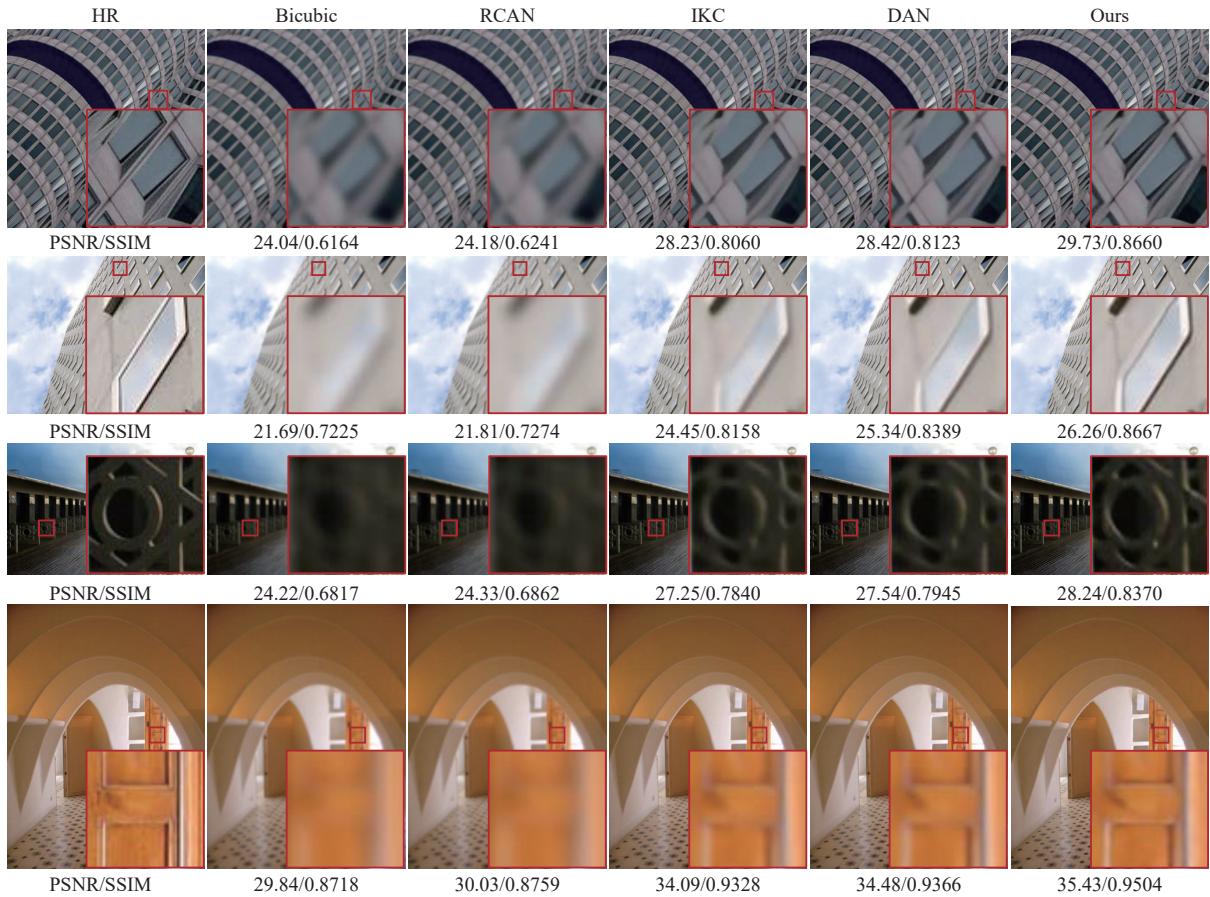


Fig. 4. Visual comparison of noise-free models achieved on four images from Urban100 for $\times 4$ SR with kernel width 3.6: img068, img010, img032, img080.

TABLE II
FLOPS AND TESTING TIME COMPARISONS WITH SOTA SR METHODS AT THE SCALE 2

	Bicubic	RCAN	IKC	DAN	DASR	CRDNet (Ours)
Flops (10^9)	\	261.46	80.59	256.45	42.44	33.04
Time (ms)	2.8	162	339	89	48	32

SR algorithms are displayed in Fig. 4 for comparison. We highlight some small areas in red boxes. It is obvious that since RCAN method is simply optimized on the bicubic downsampling, this method fails to produce clear results and behaves badly as the Bicubic method. IKC and DAN cannot restore more details and the edges in their estimations are still blurry due to the accumulated estimation error. Compared to other methods, our CRDNet can recover sharper edges with pleasurable perceptual quality.

B. Experiments on General Degradations With Anisotropic Gaussian Kernels and Noises

We also train our network on more general degradations with anisotropic Gaussian kernels and noises. Following [30], anisotropic Gaussian kernels are characterized by a Gaussian probability density function $N(0, \Sigma)$ (with zero mean and varying covariance matrix Σ). The covariance matrix Σ is determined by two random eigenvalues $\lambda_1, \lambda_2 - U(0.2, 4)$ and a random rotation angle $\theta - U(0, \pi)$. The range of noise level is set to $[0, 25]$. For testing, we use the benchmark dataset

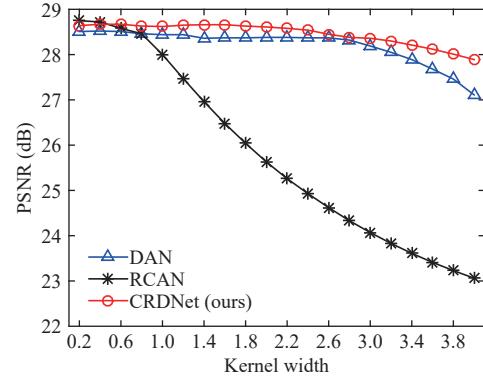


Fig. 5. Comparison of SOTA noise-free models on Set14 dataset for $\times 4$ SR with different kernel widths.

DIV2KRK that is used in [13]. We deploy 7 typical blur kernels and 3 different noise levels for performance evaluation. To handle the noise images using RCAN, we first denoise the LR images by using DnCNN [43] (an SOTA denoising method) under blind settings. We also utilize the method in [44] and the kernelGAN [30] to estimate the noise level and the degradation kernel for the SOTA non-blind SR method USRNet [25] to cope with more general degradations. Since the pre-trained model of DAN is trained on anisotropic Gaussian kernels only, we further fine-tuned this model with noises for a fair comparison. We use the officially released model of DASR for testing in this setting.

TABLE III

QUANTITATIVE COMPARISON WITH SOTA SR METHODS ON NOISY DEGRADATIONS WITH ANISOTROPIC GAUSSIAN KERNELS. THE BEST RESULTS ARE INDICATED IN **BOLD**

Method	Param	Noise	Blur kernel						
			0	27.65/0.7660	28.76/0.7968	26.08/0.7101	26.28/0.7195	26.16/0.7144	25.48/0.6853
DnCNN+RCAN	650K+15.2M	0	27.13/0.7398	27.97/0.7632	25.82/0.6946	25.99/0.7028	25.89/0.6982	25.28/0.6747	25.22/0.6724
		5	26.39/0.7122	27.01/0.7296	25.33/0.6762	25.48/0.6823	25.39/0.6785	24.87/0.6603	24.83/0.6588
		15	27.38/0.7545	28.36/0.7815	25.66/0.7033	26.43/0.7230	26.23/0.7166	25.45/0.6836	25.30/0.6831
kernelGAN+USRNet	193K+17M	0	26.82/0.7355	28.07/0.7723	25.76/0.6950	25.95/0.7036	25.92/0.7022	25.25/0.7650	25.31/0.6753
		5	26.65/0.7216	27.26/0.7405	25.45/0.6805	25.46/0.6831	25.49/0.6821	24.90/0.6623	24.92/0.6614
		15	28.05/0.7816	29.54/0.8197	27.41/0.7598	26.82/0.7416	26.91/0.7442	28.05/0.7723	29.49/0.8047
DAN	1.95M	0	27.87/0.7528	28.50/0.7796	26.60/0.7426	26.76/0.7320	26.66/0.7270	26.47/0.7104	26.22/0.7279
		5	26.84/0.7249	27.47/0.7426	25.59/0.6953	25.67/0.7048	25.60/0.6937	25.26/0.6734	25.23/0.6752
		15	29.39/0.8149	29.59/0.8160	28.92/0.7953	28.94/0.7949	28.93/0.7942	28.72/0.7882	28.71/0.7895
DASR	5.98M	0	28.25/0.7760	28.76/0.7871	27.49/0.7447	27.58/0.7467	27.52/0.7438	27.09/0.7279	27.09/0.7287
		5	27.09/0.7380	27.55/0.7495	26.34/0.7074	26.43/0.7109	26.37/0.7079	25.96/0.6961	25.94/0.6910
		15	29.46/0.8202	29.75/0.8225	29.15/0.7982	29.24/0.8039	29.11/0.7964	28.89/0.7945	28.82/0.8009
CRDNet	2.8M	0	28.30/0.7791	28.87/0.7927	27.60/0.7452	27.75/0.7513	27.56/0.7472	27.13/0.7341	27.12/0.7340
		5	27.11/0.7396	27.61/0.7514	26.38/0.7090	26.55/0.7148	26.45/0.7114	26.06/0.6976	26.05/0.6942
		15							

Table III tabulates the average quantitative performance of all comparing methods on 100 testing images in DIV2KRK dataset. According to the results, the performance of combination models (DnCNN+RCAN and kernelGAN+USRNet) is not competitive due to the accumulated reconstruction errors. As the method is designed for bicubic downsampling, RCAN generates relatively low results on complex degradations. Although USRNet can adapt to the complicated degradations, this method is sensitive to degradation estimation. Therefore, degradation estimation errors generated by the kernel and noise estimation method would be magnified in USRNet, resulting in limited SR performance. Although DAN involves the kernel estimation when restoring clear image and thus achieves better results, it generates worse results on higher noise level. The proposed approach CRDNet consistently outperforms the other methods in terms of both PSNR and SSIM as shown in Table III. Compared to the SOTA blind SR method, our CRDNet achieves the better parameter-performance trade-off.

To visualize the results, we select one setting in Table III to show the super-resolved images reconstructed by five comparing methods. As illustrated in Fig. 6, although the combination of KernelGAN and USRNet can generate slightly sharper edges than DnCNN+RCAN, it cause some artifacts in SR results. The SR image of our method is obviously much cleaner and has more reliable details.

C. Experiments on Real-World Images

We further conduct experiments on real-world scenes to demonstrate the effectiveness of our proposed CRDNet. Since there are no ground-truth HR images, we only visualize some predictions from compared methods. Fig. 7 illustrates the

super-resolved results on real-world inputs. The RCAN is employed as one of the representative non-blind algorithms for comparison and the combination of kernelGAN and USRNet is also included. The method DASR is selected as the SOTA blind SR method for comparison. We can observe that our CRDNet could give satisfactory results with sharper details. Specifically, the recovered letters in the red box are blurry in the result of RCAN, while the combination of kernelGAN and USRNet tends to produce the over-smoothed prediction. In comparison, our CRDNet can produce sharp edges without unsatisfying artifacts due to the contrastive regularization.

D. Ablation Studies

To demonstrate the effectiveness of the proposed CRDNet, we conduct an ablation study on noise-free degradations to analyze different elements, including global prior, up-and-down strategy and contrastive learning. The results are evaluated on the Set14 dataset at the scale 4.

1) *Global Prior*: We expect the extracted global prior could capture the changes made by the distortion. Thus, integrating the global prior into our model, the blind SR network could be aware of distortion. According to Table IV, the global prior alone brings little gain to the baseline network. To investigate global statistical prior thoroughly, we remove this part in our network and only use the blind SR network, the performance of which is listed in the fourth row of Table IV. Clearly, the performance suffers from severe decreases since the distortion is not involved in the feature extraction. In order to better display the extracted global priors, we visualize the extracted global statistical priors in Fig. 8 with different kernel width settings for isotropic Gaussian kernels. It can be observed that

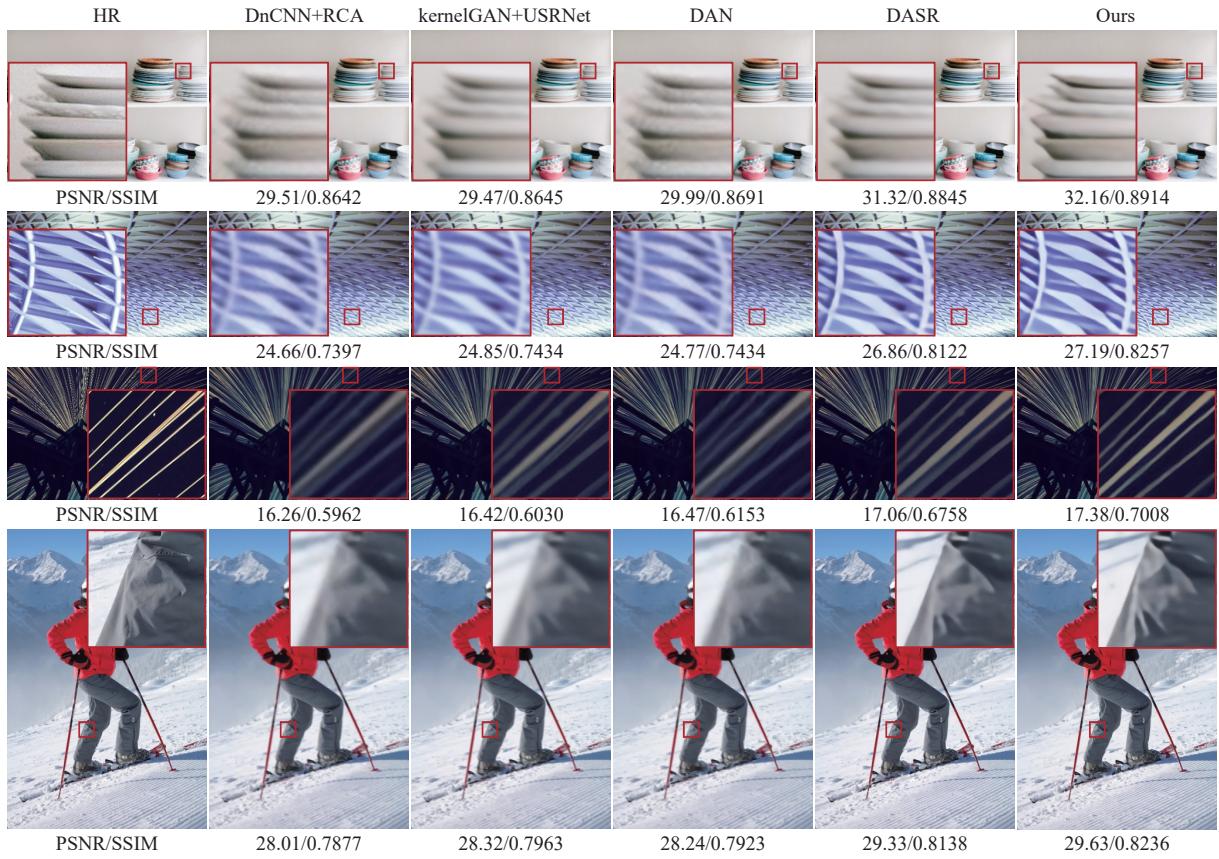


Fig. 6. Visual comparison of $\times 4$ SR models achieved on four images from DIV2KRK for $\theta = 0.125$, $\lambda_1 = 4.0$, $\lambda_2 = 2.0$ with noise level 5: img33, img92, img28, img94.



Fig. 7. Visual results on real-world image for $\times 4$ SR.

the prior information that we extracted has a strong correlation with degradation. For the anisotropic Gaussian kernel demonstrated in Fig. 9, the extracted priors are not always consistent with the ground truth since LR inputs have interfered with noise. In this case, our extracted distortion-oriented prior could promote the SR network to be well adapted to various degradations.

2) *Up-and-Down Strategy*: In our SR network, we deploy the up-and-down sampling modules to deal with the complex degradations, and meanwhile, this strategy would enlarge the receptive field of the network. First, we use this strategy

alone, and we can see that the up-and-down strategy has the greatest contribution to the baseline. Then, we remove the up-and-down sampling modules in our distortion-specific network to confirm their effects. As illustrated in Table IV, our CRDNet benefits from the up-and-down strategy which significantly improves the performance with an increase of 0.12 dB PSNR. Thus, coupling with the global prior, the up-and-down sampling module is an important part of our proposed method to cope with various degradations.

3) *Contrastive Regularization*: Contrastive learning is employed to generate discriminative SR images, in which the

TABLE IV
ABLATION STUDY ON NOISE-FREE DEGRADATIONS EVALUATING ON SET14 DATASET AT SCALE 4

Global prior	Up-and-down	Contr. regular.	PSNR/SSIM
	<i>Baseline</i>		25.17 / 0.6359
✓			25.24 / 0.6366
	✓		25.42 / 0.6513
		✓	25.31 / 0.6459
	✓	✓	25.54 / 0.6643
✓	✓		25.59 / 0.6628
✓		✓	25.52 / 0.6646
✓	✓	✓	25.64 / 0.6652

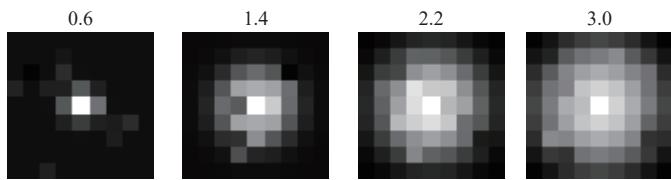


Fig. 8. Visualization of global prior extracted in the proposed network conducted with different kernel widths.

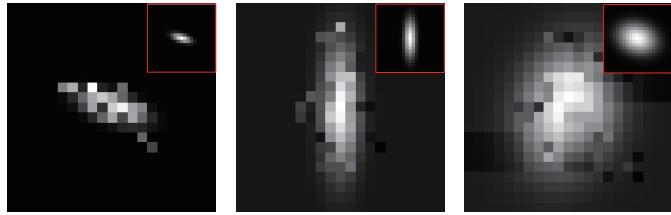


Fig. 9. Visualization of global prior extracted in the proposed network with anisotropic Gaussian kernels.

bicubic interpolation results serve as negative samples. We evaluate the effect of contrastive regularization with or without negative samples. Discarding contrastive regularization represents that only positive ones are applied for training, which is similar to perceptual loss [17]. Compared to other parts, adding contrastive regularization in our model achieves relatively fewer gains in PSNR (+0.05 dB) while higher gains in SSIM (+0.0024). Our proposed CRDNet deploys the designed contrastive regularization with both negative and positive samples during training phrase, which could temper the over-smoothness of the reconstructed image. In Fig. 10, we show that the network trained without the contrastive regularization (w/o CR) generates over-smoothed SR results with some displeased artifacts. In contrast, the SR images reconstructed by the proposed algorithm (our CRDNet) contain satisfactory clean and sharp details.

VI. DISCUSSIONS

We further explore more negative samples on contrastive regularization, which is known to benefit contrastive learning. However, the negative sample mining is still an open and challenging problem. In our method, we first use the complex degradation to downsample the high-resolution images, and

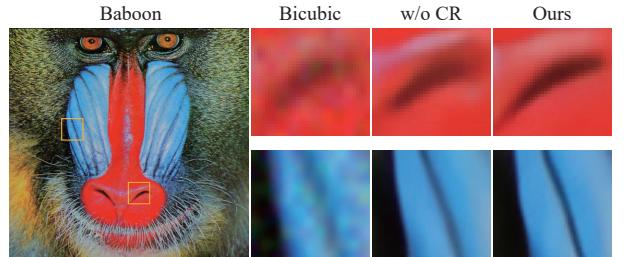


Fig. 10. Contribution of contrastive regularization (CR) in the proposed network conducted on image “Baboon” in Set14.

then upsample the degraded images by different methods to generate negative samples. Intuitively, the negative example generated by upsampling the low-resolution input is harder than the high-resolution image with different degradations. Thereby, we focus on exploiting different even mixed super-resolution techniques to generate more negative examples. In detail, we exploit other interpolation methods (Bilinear, Nearest Neighbor) and existing blind SR methods (RCAN, IKC, DAN, ESRGAN) to generate more negative samples, and randomly select a certain number of negative samples for the contrastive learning. Unfortunately, we consider at most 5 negative samples, because of the limited GPU memory size. Table V demonstrates the PSNR and SSIM results produced by different numbers of negative testing on the Set5 dataset.

TABLE V
THE PSNR AND SSIM RESULTS OF DIFFERENT NUMBERS OF NEGATIVES TESTING ON THE SET5 DATASET

Negative number	PSNR	SSIM
1	32.22	0.8936
3	32.31	0.8947
5	32.48	0.8964

As shown in Table V, adding more negative samples into contrastive regularization achieves better performance. We conjecture that for negative samples, the more negative samples, the farther away from the worse pattern in the blurry images. Thus, our method with 5 negatives achieves the best performance. However, it takes a longer training time when increasing the number of negative samples. For example, our CRDNet with 5 negatives takes about 100 hours in total (i.e., $\times 2$) for training, compared to a total of 14 hours at the rate of 1 : 7. Therefore, we still choose a single negative sample for our compact network.

VII. CONCLUSIONS

In this paper, we proposed a novel CRDNet for blind SR, which consists of contrastive regularization and distortion-specific network with global prior. Contrastive regularization is built upon contrastive learning to ensure that the restored image is pulled closer to the HR image and pushed far away from the blurry image in representation space. Instead of explicitly estimating the distortion, we extract the global prior from the MSCN coefficients to capture the character of the degradation. Embedding the global prior into the SR network

makes our method well adapted to complex degradations. The compact distortion-specific network based on the up-and-down benefits from removing noise and expanding the receptive field to improve the network's representation capability. We have comprehensively evaluated the performance of CRDNet on synthetic and real-world datasets, which demonstrates the superiority of our proposed method.

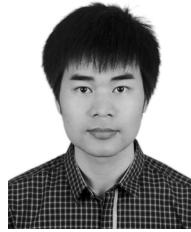
REFERENCES

- [1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2015.
- [2] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.
- [3] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image superresolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1637–1645.
- [4] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3147–3155.
- [5] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 136–144.
- [6] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
- [7] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image superresolution via information distillation network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 723–731.
- [8] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11065–11074.
- [9] N. Hao, H. Liao, Y. Qiu, and J. Yang, "Face super-resolution reconstruction and recognition using non-local similarity dictionary learning based algorithm," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 2, pp. 213–224, 2016.
- [10] L. Sun, Z. Liu, X. Sun, L. Liu, R. Lan, and X. Luo, "Lightweight image super-resolution via weighted multi-scale residual network," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 7, pp. 1271–1280, 2021.
- [11] L. Geng, Z. Ji, Y. Yuan, and Y. Yin, "Fractional-order sparse representation for image denoising," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 2, pp. 555–563, 2017.
- [12] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1604–1613.
- [13] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," arXiv preprint arXiv: 1909.06581, 2019.
- [14] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," arXiv preprint arXiv: 2010.02631, 2020.
- [15] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [16] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image superresolution using very deep residual channel attention networks," in *Proc. Europ. Conf. Comput. Vis.*, 2018, pp. 286–301.
- [17] C. Ledig, L. Theis, F. Huszár, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
- [18] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Europ. Conf. Comput. Vis. Workshop*, 2018, pp. 1–16.
- [19] R. Lan, L. Sun, Z. Liu, H. Lu, Z. Su, C. Pang, and X. Luo, "Cascading and enhanced residual networks for accurate single-image super-resolution," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 115–125, 2021.
- [20] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, 2021.
- [21] T. Köhler, M. Bätz, F. Naderi, A. Kaup, A. Maier, and C. Riess, "Toward bridging the simulated-to-real GAP: Benchmarking super-resolution on real data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2944–2959, 2019.
- [22] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3262–3271.
- [23] A. Shocher, N. Cohen, and M. Irani, "Zero-shot" super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3118–3126.
- [24] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3516–3525.
- [25] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3217–3226.
- [26] S. A. Hussein, T. Tirer, and R. Giryes, "Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1428–1437.
- [27] Y. Zhou, X. Du, M. Wang, S. Huo, Y. Zhang, and S.-Y. Kung, "Crossscale residual network: A general framework for image super-resolution, denoising, and deblocking," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5855–5867, 2022.
- [28] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 945–952.
- [29] J. Liang, K. Zhang, S. Gu, L. Van Gool, and R. Timofte, "Flow-based kernel prior with application to blind super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10601–10610.
- [30] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10581–10590.
- [31] D. L. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems*, vol. 5, no. 4, p. 517, 1994.
- [32] A. Srivastava, A. B. Lee, E. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *J. Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17–33, 2003.
- [33] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video superresolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
- [34] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7303–7313.
- [35] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2582–2593.
- [36] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 126–135.
- [37] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 114–125.
- [38] M. Bevilacqua, A. Roumy, C. Guillemot, and M. line Alberi Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. British Mach. Vis. Conf.*, 2012, pp. 1–10.
- [39] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves and Surfaces*, 2010, pp. 711–730.

- [40] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, pp. 416–423.
- [41] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5197–5206.
- [42] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based MANGA retrieval using MANGA109 dataset," *Multimed. Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, 2017.
- [43] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, 2017.
- [44] G. Chen, F. Zhu, and P. Ann Heng, "An efficient statistical method for image noise level estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 477–485.

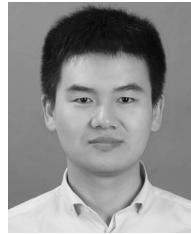


Xinya Wang received the B.S. degree in communication engineering from the Electronic Information School, Wuhan University in 2018. She is currently a Ph.D. candidate in signal and information processing at the Multi-Spectral Vision Processing Laboratory of Wuhan University. Her current research interests include computer vision and image processing.



Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

He has been identified in the 2019–2021 Highly Cited Researcher lists from the Web of Science Group. He has authored or co-authored more than 200 refereed journal and conference papers, including *IEEE TPAMI/TIP, IJCV, CVPR, ICCV, ECCV*. He is an Area Editor of *Information Fusion*, and an Editorial Board Member of *Neurocomputing*.



Junjun Jiang (Senior Member, IEEE) received the B.S. degree in mathematics from the Department of Mathematics, Huaqiao University in 2009, and the Ph.D. degree in computing science from the School of Computer, Wuhan University, in 2014. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include image processing and computer vision.

He won the Finalist of the World's FIRST 10K Best Paper Award at ICME 2017, and the Best Student Paper Runner-up Award at MMM 2015. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and 2015 ACM Wuhan Doctoral Dissertation Award.