ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20241024002089
Original scientific paper

DHANet: Dual-Stage Hybrid Attention Network for Blind Image Super-Resolution

Ningjiang MA

Abstract: In recent years, blind image super-resolution (SR) methods have demonstrated promising performance but remain limited by inaccurate blur kernel evaluation and difficulties in global feature extraction. This paper introduces DHANet, a Dual-Stage Hybrid Attention Network, combining CNN and Transformer-based modules for blind image SR. DHANet includes a blur kernel predictor, a hybrid attention dual-path module (HADM) for enhanced feature extraction, and a feature refinement module (FRM) for reconstructing refined high-resolution images. Experiments on benchmark datasets demonstrate superior performance in terms of quality and efficiency. Specifically, our method improves the average PSNR metric on four benchmark datasets from 34.98 to 35.29 at SR×2 compared to the second-best comparison method, and shows varying degrees of improvement in PSNR and SSIM metrics on different datasets at SR×3 and SR×4.

Keywords: blind image super-resolution; channel attention residual block; hybrid attention mechanism; transformer

1 INTRODUCTION

Single image super-resolution (SISR) is a subjective low-level visual problem that restores high-resolution (HR) images by reconstructing more details from observed low-resolution (LR) images. This technology can reduce hardware equipment costs and has huge practical application prospects in military, medical and other fields [1]. Similar to other restoration tasks of restoring high-quality images from low-quality images [2, 3], SISR is an ill posed problem where one LR image can restore multiple HR images. In order to address this highly challenging issue, researchers have done a lot of meaningful work.

The existing SISR methods can be divided into three categories [4] based on the way HR images are generated: interpolation-based methods [5], reconstruction-based methods [6, 7], and learning-based methods [8-25]. Interpolation-based methods estimate unknown pixels in the HR space using neighborhood information. Reconstruction-based approaches often leverage mathematical frameworks like maximum a posteriori probability or Bayesian analysis. Learning-based methods, including sparse representation [8-11], neighborhood embedding [12, 13], and deep neural networks (DNN) [14-25], reconstruct HR images by modeling the mapping between LR and HR image pairs. In recent years, DNN has achieved significant results in SISR due to its powerful feature learning ability.

LR images are usually considered to be obtained by convolving HR images with blur kernels *k*, then down sampling them and adding additive noise. The SISR degradation model can be is defined as follows.

$$I_{LR} = \left(I_{HR} \odot k\right)\downarrow_S + n \tag{1}$$

where I_{HR} is the original HR image, and I_{LR} is the degraded LR image, k is the blur kernel applied to the original HR image, $\odot$ represents the two-dimensional (2D) convolution operation, $\downarrow S$ presents the down sampling factor and n represents the Gaussian white noise.

In recent years, advanced SISR methods based on DNN have achieved excellent performance. Most methods assume that the blur kernel *k* in Eq. (1) is known and fixed, but in real-world scenarios, the blur kernel is often unknown and complex. Due to the domain gap between the assumed degradation and the actual degradation [4], the performance of these methods will significantly decrease, leading to the problem of excessively smooth or sharp edges in the reconstructed image. Considering the complex degradation factors of real degraded images, some researchers focus on the more challenging task of blind image super-resolution (BISR), where the blur kernel *k* is unknown and needs to be evaluated from observed LR images. The evaluated blur kernel is used for the SR reconstruction task of images, which can increase the generalization of the network model. This is crucial in fields with complex degradation and high demands for image quality, such as medical imaging and surveillance. In these areas, BISR effectively addresses such challenges, offering practical solutions for real-world scenarios applications.

At present, most BISR methods [27, 28] based on convolutional neural networks (CNN) generally consist of two consecutive steps: 1) estimating the blur kernels based on the observed LR images; 2) restoring SR images based on the estimated kernels. While these methods have made significant progress, they still face two limitations. First, most existing researches [26, 29] focus on building deeper and more complex neural networks, which brings heavy computational cost and memory consumption. In [26], the authors made it easier to estimate the blur kernels by sending LR and SR images to the estimator. The iterative principle can gradually make the generated SR images closer to the ground-truth images, but it will consume more computational costs and make the inference time longer [30]. Due to the limited local receptive field, CNN pays more attention to local description and can process local detail information well, but it is difficult to simultaneously obtain long-distance global information. As an advanced solution of CNN, the Transformer architecture has shown excellent performance in the fields of natural language and computer vision tasks. In [31], an efficient Transformer model can achieve excellent results in image restoration tasks by capturing long-distance features. However, the Transformer structure does not perform as well as CNN in learning local features. Balancing the learning of local and global features remains another challenge in the field of blind super-resolution.

Based on the above analysis, we propose an end-to-end dual-stage hybrid attention network (DHANet) to progressively obtain reconstructed HR images. DHANet consists of two stages: low-resolution feature reconstruction stage and high-resolution feature refinement stage, named LR-Stage and HR-Stage. In LR-stage, a dual-branch sub-network is constructed to reconstruct the features in LR space, in order to obtain a coarse SR image. The upper branch aims to extract features of different depths from LR images, while the lower branch is used to achieve deblurring of shallow features and further refine deblurring features. To extract richer features, A hybrid attention dual-path module (HADM) is constructed based on the combination of CNN and transformer attention mechanisms to achieve the extraction of global and local features. In HR-Stage, a feature refinement module (FRM) is designed to explore the available information in HR space to reconstruct a SR image with more detail information. The main contributions of this paper are as follows.

1) A DHANet consisting of two stages, namely LR feature reconstruction stage (LR-Stage) and HR feature refinement stage (HR-Stage), is proposed to progressively achieve the reconstruction of HR images.

2) In LR-Stage, a blur kernel predictor is constructed to learn a fuzzy kernel from the input LR image, and a HADM based on a channel attention residual block (CARB) and a transformer block (TB) is constructed to reconstruct the features in LR space.

3) In HR-stage, a FRM is constructed to explore the HR features in HR space to make up for the missing information in LR space, in order to reconstruct a HR image with rich detail information.

4) Extensive experiments on several different datasets demonstrate that the proposed method achieves competitive results in both quantitative and qualitative evaluations and has high inference efficiency.

2 RELATED WORK

In this section, we mainly introduce some works related to the proposed method, including blind SR reconstruction and application of Transformer in computer vision.

2.1 Blind Image SR Reconstruction

Initially, Dong et al. [14] were the first to use a three-layer convolutional network for image super-resolution reconstruction. Later, in order to improve the performance of SR reconstruction, Zhang et al. [17] proposed a channel attention mechanism, which enables the network to focus on more useful channels and enhances the model's discriminative learning ability. Most SR methods directly learn the mapping relationship between LR and HR images, without considering the degradation factors of the images. In addition, the training set used was obtained by assuming that LR images are degraded by Gaussian blur, which makes the generalization of SR methods relatively weak and cannot achieve good results when used for real degraded images. In order to improve the performance of the models in practical applications, some researchers have begun to focus on blind SR reconstruction, which considers the problem of blur degradation to enhance the generalization of the reconstructed model.

Gu et al. [32] proposed an iterative optimization scheme, namely Iterative Kernel Correction (IKC), which achieves the reconstruction of HR images by jointly optimizing the degradation estimation and image restoration processes. The DASR proposed by Wang et al. [29] benefited from the research progress of contrastive learning and improves the generalization ability of the model by accurately learning the degradation information in the LR image feature space. Luo et al. [26] proposed an end-to-end network that alternates the processes of image restoration and blur kernel estimation, achieving fine-grained reconstruction of HR images while implementing blur kernel estimation. This training method greatly reduces the inference time of the model. Liang et al. [33] used moderate convolutional receptive fields and inter-channel dependencies to estimate spatially varying blur kernels. Later, Luo et al. [34] reformulated the degradation formula to estimate blur kernels closer to the ground-truth kernels and successfully deblurred the images. Despite numerous studies on blind SR tasks, most methods still suffer from detail loss and blurring in the reconstructed HR images.

2.2 Transformer for Computer Vision

The impressive Transformer in natural language tasks has aroused great interest in the field of computer vision and has been successfully used in various computer vision tasks [35-37] such as image recognition, object detection, and image restoration. ViT [36] was the first to directly apply Transformer to image recognition tasks, dividing images into blocks and using the linear embedding sequence of these blocks as input to the Transformer structure. Chen et al. [37] explored the use of a pre-trained Transformer model called IPT in computer vision tasks. Yang et al. [35] were among the first to introduce the transformer architecture into image generation and proposed a texture Transformer consisting of four closely related modules for image SR reconstruction task. This approach achieved a more powerful feature representation by stacking multiple texture Transformers. Liang et al. [25] proposed an image restoration model called SwinIR based on Swin Transformer [38], which has fewer parameters and delivers impressive image SR effects. Zamir et al. [31] developed an efficient Transformer model for restoration tasks by designing the building blocks with multi-head attention and feed-forward network to capture long-range pixel interactions while accommodating large images. In our work, we successfully integrated two powerful feature extractors, the Transformer block and CNN, to achieve the extraction of local and global features.

3 THE PROPOSED METHOD

3.1 The Overall Structure of DHANet

In this section, we proposed a DHANet for blind image super-resolution, as shown in Fig. 1, which is composes of two stages: LR feature reconstruction stage (LR-stage) and HR feature refinement stage (HR-stage).

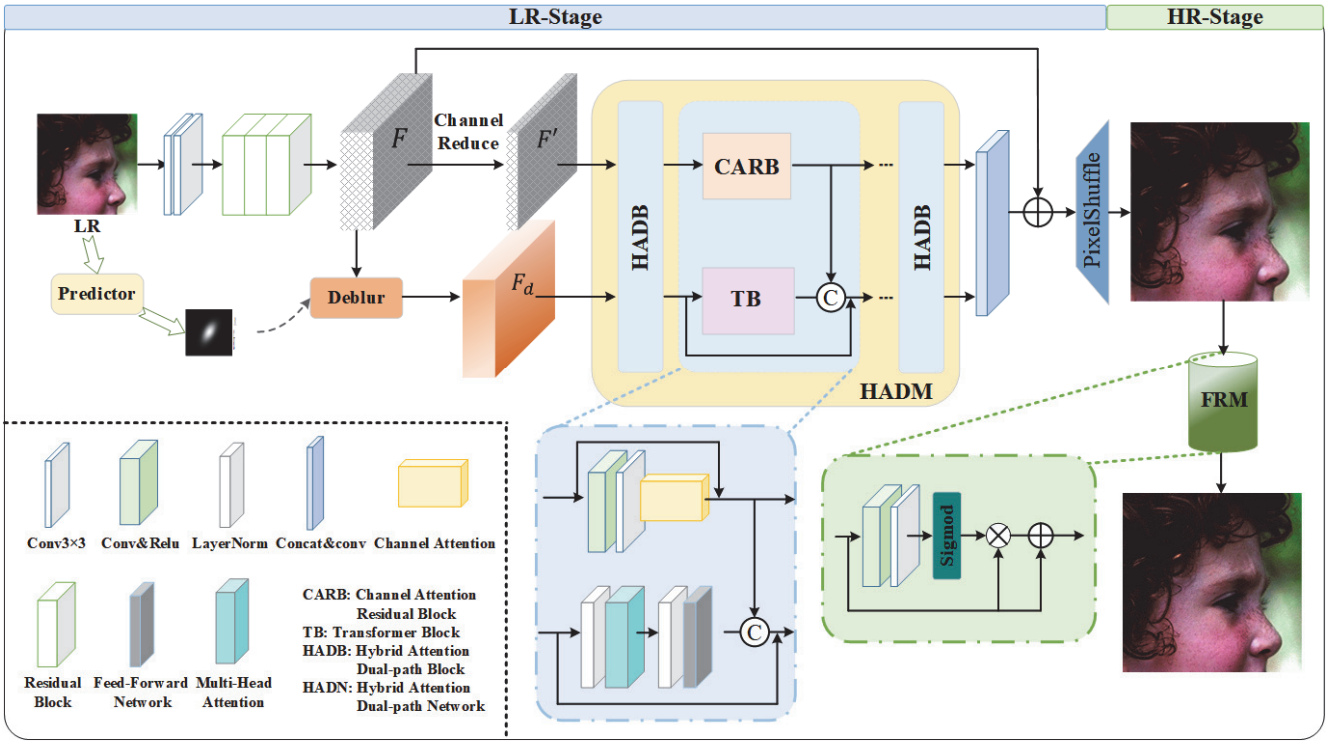

Figure 1 Overall architecture of DHANet

In the LR-stage, a dual-branch reconstruction network is constructed to achieve deblurring of image features and reconstruction of detail features, in order to obtain the reconstructed features in the low-resolution space. In the upper branch, two convolution operations with a kernel size of 3 × 3 are first used to extract shallow features from the input LR image I_{LR}. Then, three cascaded residual blocks are used for further feature extraction and enhancement to obtain the feature maps F. Each residual block consists of two convolutional layers and one channel attention layer to emphasize useful features and suppress useless features. The above operations can be represented by the following equation.

$$F = Res\left(...Res\left(Conv_{3\times3}\left(Conv_{3\times3}\left(I_{LR}\right)\right)\right)\right) \tag{2}$$

where $Conv_{3\times3}(\bullet)$ denotes a convolution operation with a kernel size of 3 × 3, and $Res(\bullet)$ denotes the operation of a residual block. Next, to reduce the computational complexity of the network, a 1 × 1 convolution operation is performed to compress the channels of feature maps F while achieving feature integration. Here, the number of channels of feature maps F is reduced to one fourth of its original size. Finally, the compressed feature maps F' are fed into the upper branch of the HADM, which consists of multiple cascaded HADBs, to achieve deep feature extraction and enhancement. The specific operation is as follows.

$$\begin{aligned} DF_{up} &= HADM_{up}\left(Conv_{1\times1}\left(F\right)\right) \\ &= HADB_{up}\left(\cdots\left(HADB_{up}\left(Conv_{1\times1}\left(F\right)\right)\right)\right) \end{aligned} \tag{3}$$

where $Conv_{1\times1}(\bullet)$ denotes a convolution operation with a kernel size of 1 × 1. $HADM_{up}(\bullet)$ and $HADB_{up}(\bullet)$ denote the upper branch operation in HADM and HADB. DF_{up} denotes the output feature maps of the upper branch in the LR-stage.

In the lower branch of the LR stage, first, a blur kernel predictor is constructed to blindly predict the blur kernel of the input image, which is applied to perform the deblurring operation on shallow features. Then, the deblurring features F_d are fed into the upper branch of the HADM, which consists of multiple cascaded HADBs, to achieve the learning of long-distance features. The specific operations are represented as follows.

$$\begin{aligned} DF_{lo} &= HADM_{lo}\left(D_{blur}\left(Pre\left(I_{LR}\right),F\right)\right) \\ &== HADB_{lo}\left(\cdots\left(HADB_{lo}\left(D_{blur}\left(Pre\left(I_{LR}\right),F\right)\right)\right)\right) \end{aligned} \tag{4}$$

where $Pre(\bullet)$ denotes the blur kernel predictor, $D_{blur}(\bullet)$ denotes the deburring operation, $HADM_{lo}(\bullet)$ and $HADB_{lo}(\bullet)$ denote the lower branch operation in HADM and HADB, DF_{lo} denotes the output feature maps of the lower branch in the LR-stage. The output features of the upper and lower branches are finally concatenated and integrated through a 3 × 3 convolutional layer. The integrated features and shallow features are merged through a residual link to obtain reconstructed features in LR space. The PixelShuffle operation rearranges elements of a low-resolution feature map into a higher-resolution output, increasing spatial resolution without adding extra parameters. Therefore, we apply PixelShuffle to the

reconstructed features here to generate a coarse HR image $\hat{I}_{CHR}$. The above operations can be represented as follows.

$$\hat{I}_{CHR} = PS\left(\left(Conv_{3\times3}\left(Concat\left(DF_{up}, DF_{lo}\right)\right)\right) + F\right) \quad (5)$$

where $Concat(\bullet)$ denotes the concatenation operation, and $PS(\bullet)$ denotes the PixelShuffle operation.

In the HR-stage, the coarse HR image from the LR-stage is sent to an FRM to refine the features in HR space, in order to output the final reconstructed HR image $\hat{I}_{HR}$. Below, we will provide a detailed introduction to the two stages of HATS-Net. Below, we will provide a detailed introduction to the components in the network, such as blur kernel predictor, HADM, and FRM.

3.2 Blur Kernel Predictor

In the degradation model (1), the bias of the blur kernel k will significantly reduce the performance of the SR model, resulting in overly smooth or sharp results. Most current methods assume that the blur kernel is known, fixed, or even discarded to simplify the problem. To improve the performance of the model, we constructed a blur kernel predictor that uses several convolutional layers to gradually increase the number of feature channels while progressively reducing their spatial size. This dynamic, parameterized approach enables blind estimation of the blur kernel for LR image degradation, as shown in Fig. 2.

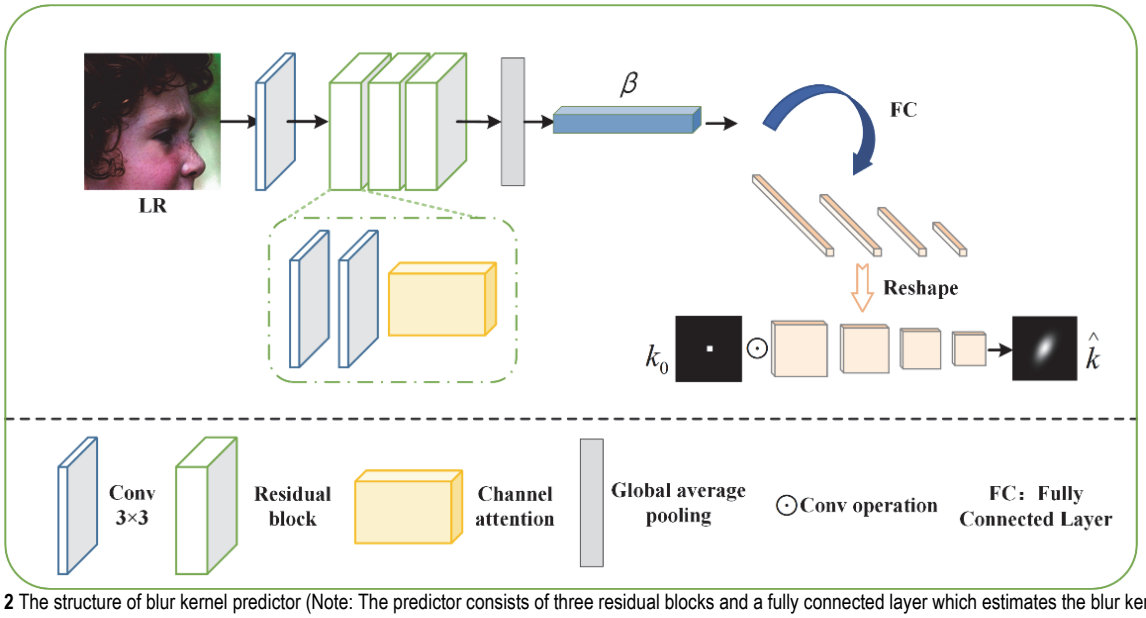

Figure 2 The structure of blur kernel predictor (Note: The predictor consists of three residual blocks and a fully connected layer which estimates the blur kernel by generating filters at four scales.)

Firstly, feature extraction was performed on the LR image using 3 × 3 convolutional layers and three residual blocks, where each residual block consists of two convolutional kernels and one channel attention. Then, a global averaging pooling was employed to compress the features to obtain a vector β containing channel information.

$$\beta = Gp\left(Res\left(...Res\left(Conv_{3\times3}\left(I_{LR}\right)\right)\right)\right) \quad (6)$$

Next, a dynamic depth linear kernel estimation method was designed to obtain the estimated blur kernel. Specifically, the fully connected (FC) operation is used to map the vector β into four filters of sizes 11 × 11, 7 × 7, 3 × 3, and 1 × 1, in order to search for different filter parameters.

$$f_i = FC_i\left(\beta\right), i = 1,2,3,4 \quad (7)$$

where $Gp(\bullet)$ denotes the global averaging pooling operation. $FC_i(\bullet)$ denotes the i-th $(i = 1,2,3,4)$ FC operation, and its output is the i-th filter f_i.

Finally, we initialized a blur kernel k_0 with a size of 21 × 21, with only one value of 255 at the center point and all other values being 0. Four filters are sequentially convolved with k_0 to obtain a compact kernel with a step size of 1, which is the estimated blur kernel. This process can be mathematically represented as follows.

$$\hat{k} = k_0 * f_1 * f_2 * f_3 * f_4 \quad (8)$$

where $*$ denotes the convolution operation.

3.3 Hybrid Attention Dual-Path Module (HADM)

To learn more diverse features, we construct an HADM composed of multiple HADBs, where the dual-branch cascades of HADBs form the dual branches of the HADM. The upper branch of HADB consists of a CARB for learning local spatial features, while the lower branch contains a TB for learning long-distance features. CARB consists of 3 × 3 convolutional layers, a ReLU activation layer, and a channel attention, mainly focusing on learning local features and enhancing features by learning the correlation of features in the channel dimension. The upper

branch operation of the *t*-th HADB can be represented as follows.

$$DF_{up}^{t} = CARB\left(DF_{up}^{t-1}\right) = CA\left(Conv_{3\times3}\left(\mathrm{Re}lu\left(Conv_{3\times3}\left(DF_{up}^{t-1}\right)\right)\right)\right)+DF_{up}^{t-1}, \quad DF_{up}^{0}=F' \tag{9}$$

where $CARB(\bullet)$ denotes the operation of CARB, $Relu(\bullet)$ denotes the ReLU function, and $CA(\bullet)$ denotes the channel attention operation. DF_{up}^{t} denotes the output of the upper branch in the *t*-th HADB, which are used as supplementary features for the lower branch.

In the lower branch of HADB, firstly, a TB consisting of two normalization layers, a multi-head attention module, and a feedforward network is used to learn the global contextual information by calculating the self-attention on the channel dimension. The, the feature maps DF_{up}^{t} from the upper branch are concatenated with the feature maps output by TB to form the output DF_{lo}^{t} of the lower branch. To reduce the computational burden caused by feature map stacking, the concatenated feature maps are compressed in the channel dimension through a 1 × 1 convolutional layer. In addition, a residual link is used to stabilize network training and reuse shallow features. The upper branch operation of the *t*-th HADB can be represented as follows.

$$DF_{lo}^{t} = Conv_{1\times1}\left(Concat\left(DF_{up}^{t}, TB\left(DF_{lo}^{t-1}\right)\right)\right)+DF_{lo}^{t-1} \tag{10}$$

where $TB(\bullet)$ denotes the operation of TB, and DF_{lo}^{t} denotes the output of the lower branch.

3.4 Fine Reconstruction Module (FRM)

In HR-stage, an FRM is constructed to refine features in HR space to reconstruct HR images with rich details. Firstly, the coarse HR image is fed into 3 × 3 convolution operations and a ReLu function to extract the features in HR space. Then, a Sigmoid function is used to obtain a spatial weight map that weights rough HR images to enhance image features. Finally, a residual link is constructed between the coarse HR image and the enhanced features to output the final reconstructed HR image $\hat{I}_{HR}$. The above operations are expressed as follows.

$$\hat{I}_{HR} = \hat{I}_{CHR} \otimes Sig\left(Conv_{3\times3}\left(\mathrm{Re}lu\left(Conv_{3\times3}\left(\hat{I}_{CHR}\right)\right)\right)\right)+\hat{I}_{CHR} \tag{11}$$

where $\otimes$ denotes the element-wise multiplication operation. $Sig(\bullet)$ denotes the Sigmod function.

3.5 Loss Function

To better train the network, a loss function containing two loss terms, namely blur kernel loss and reconstruction loss, is defined as follows.

$$L = \left\|k-\hat{k}\right\|_1 + \left\|I_{GT}-\hat{I}_{HR}\right\|_1 \tag{12}$$

where k and $\hat{k}$ denote the ground-truth and the predicted blur kernels, respectively. I_{GT} donates the ground-truth (GT) image. $\|\cdot\|_1$ donates the L_1 norm.

Table 1 Quantitative results of different comparison methods on four benchmark datasets. (The best one is highlighted in bold and the second-best is highlighted in underline.)

Scale	Methods	Set5 PSNR/SSIM	Set14 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM	AVE PSNR/SSIM	Percent improvement in PSNR
×2	Bicubic	29.42/0.8666	27.13/0.7797	24.23/0.7430	25.60/0.8498	26.59/0.8097	32.7%
	RCAN [25]	29.81/0.8797	27.54/0.8804	24.69/0.7727	-	27.35/0.8443	29%
	IKC [32]	34.31/0.9287	31.69/0.8789	29.22/0.8752	36.93/0.9667	33.03/0.9123	6.8%
	DASR [29]	37.22/0.9515	32.93/0.9029	30.63 /0.9079	-	33.59/0.9208	5.1%
	KXNet [39]	37.58/**0.9552**	33.36/0.9091	31.48/**0.9192**	37.48/0.9737	34.98/0.9393	0.9%
	Ours	**37.67**/**0.9552**	**33.41**/**0.9103**	**31.53**/0.9188	**38.55**/**0.9750**	**35.29**/**0.9398**	-
×3	Bicubic	26.19/0.7716	24.58/0.6671	22.07/0.6216	22.98/0.7576	23.96/0.7045	31.4%
	RCAN [25]	26.37/0.7840	24.73/0.6800	22.18/0.6366	-	24.43/0.7002	28.9%
	IKC [32]	32.90/0.8997	29.41/0.8106	26.85/0.8087	32.43/0.9316	30.40/0.8627	3.6%
	DASR [29]	33.78/0.9200	29.94/0.8266	27.28/0.8307	-	30.33/0.8591	3.8%
	KXNet [39]	34.22/**0.9238**	30.27/0.8340	**28.00**/0.8457	33.34/**0.9422**	31.45/0.8864	0.1%
	Ours	**34.31**/**0.9238**	**30.29**/**0.8342**	27.97/**0.8491**	**33.41**/0.9416	**31.49**/**0.8869**	-
×4	Bicubic	24.43/0.7045	23.25/0.6036	20.96/0.5544	21.50/0.6933	22.54/0.6390	30.1%
	RCAN [25]	24.52/0.7148	23.30/0.6109	20.96/0.5608	-	22.93/0.6288	27.9%
	IKC [32]	29.83/0.8375	26.88/0.7301	24.42/0.7112	28.91/0.8782	27.51/0.7893	6.6%
	DASR [29]	31.68/0.8854	28.26/0.7668	25.49/0.7621	-	28.48/0.8048	3.0%
	KXNet [39]	31.94/**0.8912**	28.67/**0.7782**	**26.18**/0.7873	30.50/**0.9089**	29.32/**0.8414**	0.1%
	Ours	**32.04**/0.8899	**28.69**/0.7765	26.11/**0.7878**	**30.51**/**0.9089**	**29.33**/0.8407	-

4 EXPERIMENT RESULTS AND ANALYSIS

4.1 Dataset and Experimental Setup

Based on existing work [32] we collected a total of 3450 HR images, including 800 images from the DIV2K dataset [40] and 2650 images from the Flickr2K dataset [41], which were used as our training data. To verify the generalization of the model, we generate LR images using two different blur kernels based on equation (1). One is the isotropic Gaussian blur kernel, and another is the anisotropic Gaussian blur kernel. Two different degradations were applied to two types of datasets: four commonly used benchmark datasets including Set5 [42], Set14 [43], Urban100 [44] and Manga109 [45], as well as DIV2KRK dataset [27].

Isotropic Gaussian kernel. Blind SR experiments on the isotropic Gaussian blur kernels were carried out according to the parameter settings [32]. Specifically, the kernel size is set to 21. In the training, we sampled the kernel widths uniformly from the ranges of [0.2, 2.0], [0.2, 3.0] and [0.2, 4.0], and the corresponding SR scale factors were set to 2, 3 and 4, respectively. For the test set, the blur kernel is set to Gaussian8, which samples uniformly the 8 cores in the range of scale factors 2, 3 and 4, respectively.

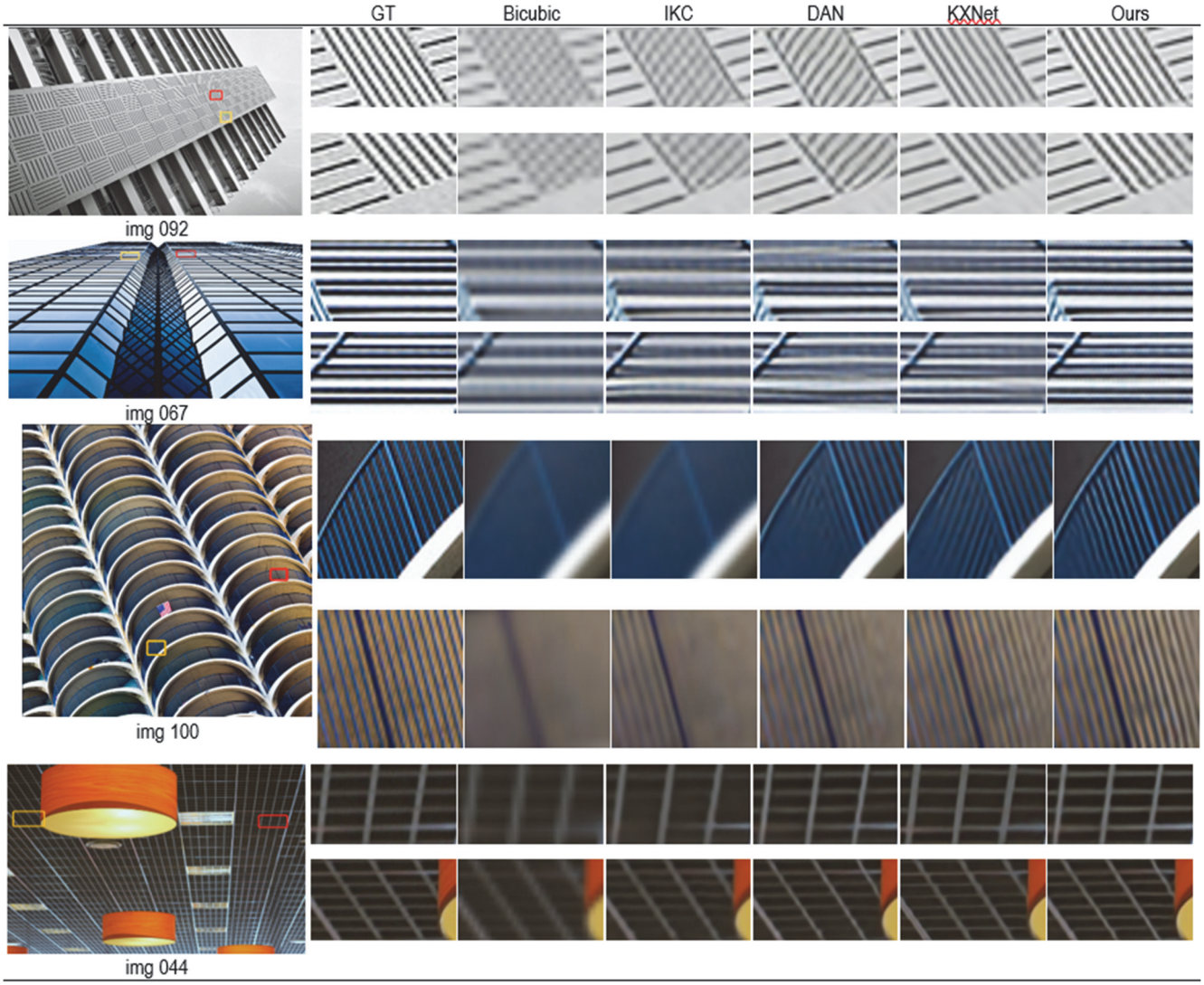

Figure 3 Comparison of SR×2 on img 092 and img 067, SR×2 on img100, and SR×4 on img 044 in the Urban100 dataset (The right side shows zoomed-in insets for each method)

Anisotropic Gaussian kernel. Blind SR experiments on the anisotropic Gaussian kernels were carried out according to the parameter settings [27]. We employed the method similar to generating isotropic Gaussian blur kernels to generate anisotropic Gaussian blur kernels. The kernel widths on the two axes are uniformly distributed in (0.6, 5.0). In addition, the random angle rotation was selected to generate an anisotropic Gaussian kernel, which is used for the DIV2KRK dataset to obtain the test set.

The proposed model was implemented using PyTorch 1.8, and trained on an Ubuntu 20.04 system with an NVIDIA GeForce GTX A6000. For all scale factors, the input LR image block size is 64 × 64, and the batch size is 64. The model was trained for 5×10^5 iterations using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$. The learning rate was initialized to 4×10^{-4} and attenuated by a factor of 0.5 every 2×10^{-4} iterations. To increase the diversity of training data, we used data augmentation operations such as random horizontal flipping and 90-degree rotation on the training set. For all the experiments except the ablation studies, the number of HADM is set to18 In the upper branch of HADMs, each layer in CARBs contained 16

channels and the others of DHANet contained 64 channels. The SR results obtained by all comparison methods were evaluated on the luminance channel (Y channel) in the YCbCr space. two widely-used evaluation metrics, Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), were used for quantitative comparison.

4.2 Comparison with State-of-the-Art Methods

To validate the effectiveness of our model, we conducted extensive experiments on four benchmark datasets and the DIV2KRK dataset. The quantitative results of our method were compared with those of several recent state-of-the-art methods, including non-blind SISR method RCAN [25] and blind SISR method, including IKC [32], DASR [29], DAN [26], KXNet [39], KernelGAN [27] + ZSSR [21], KOALAnet [22], and DCLS [34]. Fig. 3 and Tab. 1 present the subjective and objective results of some SOFA methods and our model on four benchmark datasets, respectively. Tab. 2 gives the subjective and objective results of some SOTA methods and our model on the DIV2KRK dataset.

Fig. 3 shows the SR × 2 results for img 044, img 067, SR × 3 results for img100, and the SR × 4 results for img 092 in the Urban100 dataset. From the figure, we can see that the non-blind SR method IKC loses more texture information compared to blind SR methods including DAN, KXNet, and our method. KXNet and our method reconstruct more texture structures, but our method leverages the combined strengths of CNN and Transformer architectures, enabling more effective capture of both local and global features, as well as better feature interaction. As a result, our method produces sharper edges and fewer artifacts in the reconstructed images compared to KXNet. Tab. 1 provides the quantitative results obtained by the comparison methods on four benchmark datasets degraded by isotropic Gaussian kernels. In the last column of the table, we present the average PSNR and SSIM on four datasets, demonstrating the average performance of each method. As shown in Tab. 1, we can see that blind SR methods are superior to non-blind SR methods (Bicubic and RCAN), and the proposed DHANet has achieved mostly optimal results at different scales compared to other methods. Our results are relatively close to those obtained by the KXNet method, but in objective comparison, our results have sharper edges compared to the KXNet method. Therefore, considering both objective and subjective results, our method outperforms other comparative methods.

The objective values of comparison methods in Tab. 2 are all from the original references provided by the author. Due to the inability to obtain testing models for some comparison methods, the experiments conducted on the DIV2KRK dataset in this paper only provided objective comparisons. From Tab. 2, we can see that the blind SR method is also superior to the non-blind SR methods (Bicubic and RCAN). Compared with other methods, the proposed DHANet achieved the highest PANR values at both scales. Therefore, our method is also effective for the degradation of anisotropic Gaussian kernels.

Table 2 Quantitative results of different comparison methods on the DIV2KRK dataset. (The best one is highlighted in bold and the second-best is highlighted in underline.)

Scale	Methods	DIV2KRK PSNR/SSIM	Percent improvement in PSNR
×2	Bicubic	28.73/0.8040	14.3%
	RCAN[25]	29.20/0.8223	12.5%
	KernelGAN [27]+ZSSR[21]	30.36/0.8669	8.2%
	KOALAnet[22]	31.89/0.8852	3.0%
	DAN [26]	32.56/0.8997	0.89%
	DCLS[34]	<u>32.75</u>/**0.9094**	<u>0.3%</u>
	Ours	**32.85**/<u>0.9062</u>	-
×4	Bicubic	25.33/0.6795	14.5%
	RCAN[25]	25.66/0.6936	13.0%
	KernelGAN[27]+ZSSR[21]	26.81/0.7316	8.2%
	KOALAnet[22]	27.77/0.7637	4.4%
	DAN[26]	27.55/0.7582	5.3%
	DCLS[34]	<u>28.99</u>/<u>0.7946</u>	<u>0.03%</u>
	Ours	**29.00/0.7961**	-

4.3 Ablation Study

1) Number of HADB. We conducted experiments to verify the impact of HADB quantity on network performance, and the experimental results are shown in Tab. 3.

Table 3 Ablation study on the impact of the number N of HADBs on model performance. (The best one is marked in bold.)

Number	$N = 15$	$N = 17$	$N = 18$	$N = 19$	$N = 20$
PSNR	37.57	37.60	**37.67**	**37.67**	37.65

It can be observed that as N increases, the PSNR value of the SR results increases, and the performance of the model is improved. However, when the number of HADBs increased to 19, the PSNR value reached its maximum. When $N = 20$, the PSNR value begins to decrease. This may be due to the increase in the network's parameters, which leads to redundant feature processing and, consequently, a decline in performance. To ensure that the proposed DHANet model achieves an optimal balance between SR performance and efficiency, we set N to 18.

2) Effectiveness of HADB and FRM. To verify the effectiveness of hybrid attention mechanism based on CNN and Transformer in HADB, we have conducted the following ablation studies. In addition, to further verify the effectiveness of FRM in HR-stage, we conducted experiments by removing HR-stage from the network and retaining only LR-stage. The experimental results are shown in Tab. 4.

Table 4 Ablation study of HADB and the FRM in HR-stage. (The best one is marked in bold.)

Model	w/o HAM	w/o FRM	ours
PSNR	37.28	37.60	**37.67**

We first replace the TB in HADB with CNN-based channel attention module. From the table, we can see that the performance of the model with HADB using only a single attention mechanism has significantly decreased. Similarly, the performance of model that only retains LR-stage after removing FRM also shows a significant decline. Therefore, the hybird attention mechanism in HADB can achieve a balance between capturing local details and long-distance dependencies, and improve the performance of the model. FRM can effectively refine the features in the HR space, contributing to the reconstruction of finer SR images.

4.4 Efficiency Analysis

Tab.5 shows the average testing time of IKC, DAN, KXNet, and our method on the Set5 dataset. Our method focuses on balancing performance and efficiency. Although our method has slightly higher parameter count than several other methods, it achieves the shortest average test time on the Set5 dataset. By utilizing efficient and simple operations, we minimize inference time to the greatest extent, offering better computational efficiency, which is of significant importance for practical applications.

Table 5 Efficiency analysis of different methods. (The best one is marked in bold).

Methods	IKC	DAN	DANv2	KXNet	ours
Times / s	3.28	0.75	0.70	0.40	**0.37**
Parameters / M	**4.1**	4.2	4.5	6.5	9
PSNR	34.31	37.33	37.60	37.58	**37.67**

5 CONCLUSION

In this work, we introduce a novel end-to-end hybrid attention two-path network for blind image super-resolution, called DHANet, which consists of two stages: LR-stage and HR-stage. In LR-stage, a dual-branch sub-network consisting of multiple HDAMs is constructed to achieve feature reconstruction in LR space. In addition, to address the issue of blur degradation in network, a blur kernel predictor is designed to evaluate fuzzy kernels from LR images, which are used to achieve deblurring of shallow features. HDAM is constructed based on a hybrid attention mechanism of CNN and transformer, and uses features extracted from LR images and deblurring features as inputs for the upper and lower branches to achieve the extraction of local and global features. In HR stage, an FRM is constructed to explore the effective information of HR in HR space to reconstruct more refined HR images. The proposed method has been validated through a series of experiments, and the experimental results also show that our method has good performance. However, certain limitations remain that warrant further exploration. For instance, the performance is constrained by degradations not covered by the model, and the generalization capability needs to be further enhanced for handling complex real-world degradation scenarios. In the future, we aim to explore applying this network to tasks like video super-resolution, where temporal information can enhance performance. Additionally, further research on real-world degradation is crucial for reducing the cost of high-quality imaging in real-world and advancing other high-level visual tasks.

6 REFERENCES

[1] Wan, W., Wang, Z., Wang, Z., Gu, L., Sun, J., & Wang, Q. (2024). Arbitrary-Scale Image Super-Resolution via Degradation Perception. *IEEE Transactions on Computational Imaging*, *10*, 666-676. https://doi.org/10.1109/TCI.2024.3393712

[2] Liu, X. (2020). Real-time Defogging of Single Image of IoTs-based Surveillance Video Based on MAP. *Tehnički vjesnik*, *27*(4), 1262-1269. https://doi.org/10.17559/TV-20200527085338

[3] Huang, J. & Jin, Z. (2024). Enhanced Image Denoising with Diffusion Probability and Dictionary Learning Adaptation. *Tehnički vjesnik*, *31*(3), 774-783. https://doi.org/10.17559/TV-20230808000859

[4] Zhang, Y., Dong, L., Yang, H., Qing, L., He, X., & Chen, H. (2022). Weakly-supervised contrastive learning-based implicit degradation modeling for blind image super-resolution. *Knowledge-Based Systems*, *249*, 108984 https://doi.org/10.1016/j.knosys.2022.108984

[5] Zhu, S., He, Z., Liu, S., & Zeng, B. (2017). MMSE-directed linear image interpolation based on nonlocal geometric similarity. *IEEE Signal Processing Letters*, *24*(8), 1178-1182. https://doi.org/10.1109/LSP.2017.2711609

[6] Jiang, J., Ma, X., Chen, C., Lu, T., Wang, Z., & Ma, J. (2016). Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means. *IEEE Transactions on Multimedia*, *19*(1), 15-26. https://doi.org/10.1109/TMM.2016.2599145

[7] Chantas, G., Nikolopoulos, S. N., & Kompatsiaris, I. (2020). Heavy-tailed self-similarity modeling for single image super resolution. *IEEE Transactions on Image Processing*, *30*, 838-852. https://doi.org/ 10.1109/TIP.2020.3038521

[8] Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE transactions on image processing*, *19*(11), 2861-2873. https://doi.org/ 10.1109/TIP.2010.2050625

[9] Zeyde, R., Elad, M., & Protter, M. (2012). On single image scale-up using sparse-representations. *Proceedings of Curves and Surfaces: 7th International Conference, Avignon,* 711-730. https://doi.org/10.1007/978-3-642-27413-8_47

[10] Zhu, Z., Guo, F., Yu, H., & Chen, C. (2014). Fast single image super-resolution via self-example learning and sparse representation. *IEEE Transactions on Multimedia*, *16*(8), 2178-2190. https://doi.org/10.1109/TMM.2014.2364976

[11] Zhao, J., Hu, H., & Cao, F. (2017). Image super-resolution via adaptive sparse representation. *Knowledge-Based Systems*, *124*, 23-33. https://doi.org/10.1016/j.knosys.2017.02.029

[12] Timofte, R., De Smet, V., & Van Gool, L. (2015). A+: Adjusted anchored neighborhood regression for fast super-resolution. *Proceedings of Computer Vision--ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore,* Part IV 12, 111-126. https://doi.org/10.1007/978-3-319-16817-3_8

[13] Pérez-Pellitero, E., Salvador, J., Ruiz-Hidalgo, J., & Rosenhahn, B. (2016). Antipodally invariant metrics for fast regression-based super-resolution. *IEEE Transactions on Image Processing*, *25*(6), 2456-2468. https://doi.org/10.1109/TIP.2016.2549362

[14] Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, *38*(2), 295-307. https://doi.org/10.1109/TPAMI.2015.2439281

[15] Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646-1654. https://doi.org/10.1109/CVPR.2016.182

[16] Lim, B., Son, S., Kim, H., Nah, S., & Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136-144. https://doi.org/10.1109/CVPRW.2017.151

[17] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. *Proceedings of the European conference on computer vision*, 286-301. https://doi.org/10.1007/978-3-030-01234-2_18

[18] Dai, T., Cai, J., Zhang, Y., Xia, S. T., & Zhang, L. (2019). Second-order attention network for single image super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11065-11074. https://doi.org/10.1109/CVPR.2019.01132

[19] Zhao, X., Zhang, Y., Zhang, T., & Zou, X. (2019). Channel splitting network for single MR image super-resolution. *IEEE transactions on image processing*, *28*(11), 5649-5662. https://doi.org/10.1109/TIP.2019.2921882

[20] Tian, C., Zhuge, R., Wu, Z., Xu, Y., Zuo, W., Chen, C., & Lin, C. W. (2020). Lightweight image super-resolution with enhanced CNN. *Knowledge-Based Systems*, *205*, 106235. https://doi.org/10.1016/j.knosys.2020.106235

[21] Shocher, A., Cohen, N., & Irani, M. (2018). "zero-shot" super-resolution using deep internal learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3118-3126. https://doi.org/10.1109/CVPR.2018.00329

[22] Kim, S. Y., Sim, H., & Kim, M. (2021). Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10611-10620. https://doi.org/10.1109/CVPR46437.2021.01047

[23] Cheng, G., Matsune, A., Du, H., Liu, X., & Zhan, S. (2022). Exploring more diverse network architectures for single image super-resolution. *Knowledge-Based Systems*, *235*, 107648. https://doi.org/10.1016/j.knosys.2021.107648

[24] Zhang, Y., Wei, D., Qin, C., Wang, H., Pfister, H., & Fu, Y. (2021). Context reasoning attention network for image super-resolution. *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 4278-4287. https://doi.org/10.1109/ICCV48922.2021.00424

[25] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. *Proceedings of the IEEE/CVF international conference on computer vision*, 1833-1844. https://doi.org/10.1109/ICCVW54120.2021.00210

[26] Huang, Y., Li, S., Wang, L., & Tan, T. (2020). Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, *33*, 5632-5643. https://dl.acm.org/doi/abs/10.5555/3495724.3496197

[27] Bell-Kligler, S., Shocher, A., & Irani, M. (2019). Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, *32*. https://dl.acm.org/doi/abs/10.5555/3454287.3454313

[28] Liang, J., Zhang, K., Gu, S., Van Gool, L., & Timofte, R. (2021). Flow-based kernel prior with application to blind super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10601-10610. https://doi.org/10.1109/CVPR46437.2021.01046

[29] Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., & Guo, Y. (2021). Unsupervised degradation representation learning for blind super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 0581-10590. https://doi.org/10.1109/CVPR46437.2021.01044

[30] Hui, Z., Li, J., Wang, X., & Gao, X. (2021). Learning the non-differentiable optimization for blind super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2093-2102. https://doi.org/10.1109/CVPR46437.2021.00213

[31] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M. H. (2022). Restormer: Efficient transformer for high-resolution image restoration. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728-5739. https://doi.org/10.1109/CVPR52688.2022.00564

[32] Gu, J., Lu, H., Zuo, W., & Dong, C. (2019). Blind super-resolution with iterative kernel correction. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1604-1613. https://doi.org/10.1109/CVPR.2019.00170

[33] Liang, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Mutual affine network for spatially variant kernel estimation in blind image super-resolution. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4096-4105. https://doi.org/10.1109/ICCV48922.2021.00406

[34] Luo, Z., Huang, H., Yu, L., Li, Y., Fan, H., & Liu, S. (2022). Deep constrained least squares for blind image super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17642-17652. https://doi.org/10.1109/CVPR52688.2022.01712

[35] Yang, F., Yang, H., Fu, J., Lu, H., & Guo, B. (2020). Learning texture transformer network for image super-resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5791-5800. https://doi.org/10.1109/CVPR42600.2020.00583

[36] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929

[37] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., ... & Gao, W. (2021). Pre-trained image processing transformer. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12299-12310. https://doi.org/10.1109/CVPR46437.2021.01212

[38] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[39] Fu, J., Wang, H., Xie, Q., Zhao, Q., Meng, D., & Xu, Z. (2022, October). Kxnet: A model-driven deep neural network for blind super-resolution. *Proceedings of the European Conference on Computer Vision*, 235-253. https://doi.org/10.1007/978-3-031-19800-7_14

[40] Agustsson, E. & Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126-135. https://doi.org/10.1109/CVPRW.2017.150

[41] Timofte, R., Agustsson, E., Van Gool, L., Yang, M. H., & Zhang, L. (2017). Ntire 2017 challenge on single image super-resolution: Methods and results. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 114-125. https://doi.org/10.1109/CVPRW.2017.149

[42] Bevilacqua, M., Roumy, A., Guillemot, C., & Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding. https://doi.org/10.5244/C.26.135

[43] Zeyde, R., Elad, M., & Protter, M. (2012). On single image scale-up using sparse-representations. *Proceedings of the Curves and Surfaces: 7th International Conference*, *7*, 711-730. https://doi.org/10.1007/978-3-642-27413-8_47

[44] Huang, J. B., Singh, A., & Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5197-5206. https://doi.org/10.1109/CVPR.2015.7299156

[45] Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., & Aizawa, K. (2017). Sketch-based manga retrieval using manga109 dataset. *Multimedia tools and applications*, *76*, 21811-21838. https://doi.org/10.1007/s11042-016-4020-z

Contact information:

Ningjiang MA
(Corresponding author)
Tiangong University,
School of Computer Science and Technology,
Tianjin 300387, China
E-mail: ningjiangma@163.com