CUSTOMIZED TREATMENT PER PIXEL FOR BLIND IMAGE SUPER-RESOLUTION

*Guanqun Liu, Xiaoshuai Hao**

Vision Computing Lab, Samsung Research China - Beijing (SRC-B), Chaoyang District, Beijing

ABSTRACT

Blind image super-resolution task aims at restoring high-resolution images from their low-resolution counterparts by reversing the unknown degradation. Existing methods have achieved promising results when handling degradation with isotropic or anisotropic gaussian blur, whereas suffer from performance drop when addressing degradation with motion blur. Compared with gaussian blur, motion blur is more diverse, *i.e.*, each pixel moves a peculiar distance in distinct orientation, thereby each pixel requires individual treatment. To tackle this degradation with motion blur issue, we propose a novel blind image super-resolution method named deform**A**ble **rec**eptive **S**uper **R**esolution (ArcSR), which provides deformable receptive field and unique parameters for each pixel. Specifically, we propose Deformable Mutual convolution (DMconv) and Kernel Guided convolution (KGconv) for blur kernel estimation and super-resolution, respectively. The DMconv explores the correlation within channels of image features and achieves deformable receptive field by redesigning deformable convolution to generate kernels for each pixel. Meanwhile, the KGconv views the estimated kernel as an attention matrix for convolutional parameters and gives each pixel disparate convolutional parameters to rewrite the missing high-frequency information. Comprehensive experiments demonstrate the superiority of our method.

***Index Terms*—** Blind Super-resolution, Motion Blur, Deformable Receptive Field, Customized Parameters

1. INTRODUCTION

Given a low-resolution image (LR), blind image super-resolution (SR) methods aim to restore the high-resolution counterpart (HR) by reversing the unknown degradation, which commonly consists of blurriness, noise, and downsampling. However, since the degradation is not specified in this situation and one LR image has plenty HR counterparts, the reverse mapping of degradation is not unique. Therefore, how to solve this ill-posed issue draws a great deal of attention from academia and industry.

Existing methods [1, 2, 3, 4] have achieved remarkable results on this field. They generally estimate [2, 4] or represent [3, 5] the degradation and super-resolve the LR image with it. These methods generate promising HR images when handling gaussian blur, whereas they suffer from severe performance drops when dealing with motion blur as they ignore the motion blur. Compared with gaussian blur, motion blur is more flexible, *i.e.*, each pixel moves a peculiar distance in distinct orientation and requires unique processing. Hence, a method that can handle degradation compound with motion blur needs to be further explored.

In this paper, we propose deform**A**ble **rec**eptive **S**uper **R**esolution (ArcSR). As the motion blur is spatially variant and every pixel around one pixel has an unequal influence on it, we argue that the network should adopt an unfixed receptive field and dissimilar convolutional parameters for each pixel, and introducing multi-head self-attention mechanism will bring too much computational cost. To this end, we propose Deformable Mutual convolution (DMconv) for kernel estimation and Kernel Guided convolution (KGconv) for super-resolution. Specifically, our DMconv redesigns the deformable convolution to simultaneously achieve unfixed receptive field and explore the correlation within channels of image features which is proved to be useful for spatially variant blur kernel estimation [4]. Furthermore, to remove the degradation and rewrite the missing high-frequency information, our KGconv gives each pixel unique convolutional parameters by viewing each estimated blur kernel as an attention matrix for convolutional parameters. We conduct comprehensive comparison experiments and prove that ArcSR is remarkably superior to state-of-the-art methods on existing benchmarks. Moreover, we conduct ablation experiments and reveal the function of our DMconv and KGconv. Our main contributions are summarized as follows:

- To tackle the unknown degradation with motion blur problem in blind SR task, we propose a novel blind image super-resolution method named deform**A**ble **rec**eptive **S**uper **R**esolution (ArcSR), which provides deformable receptive field and unique parameters for each pixel.
- Specifically, we propose Deformable Mutual convolution (DMconv) and Kernel Guided convolution (KGconv) for blur kernel estimation and super-resolution, respectively.
- ArcSR is remarkably superior to SOTAs on existing methods, revealing the effectiveness of our approach.

*Corresponding author.

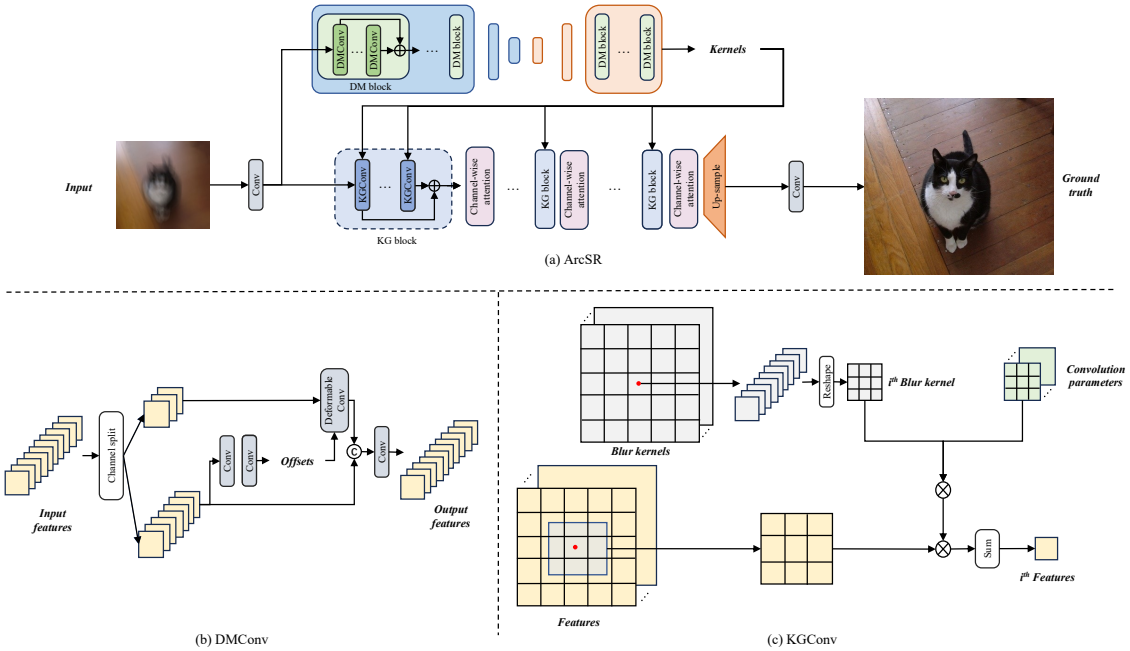

Fig. 1. **Overview of the proposed ArcSR.** (a) The structure of our ArcSR. We first estimate the blur kernel through a U-net network and then super-resolve the LR image. (b) The structure of our DMconv, which explores the correlation between the channels of input features through deformable receptive field. (c) The structure of our KGconv, which rewrites the missing textures by customizing unique convolutional parameters for each pixel.

2. RELATED WORK

Recently, researchers adopted bicubic downsampling as the degradation and proposed many excellent methods [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. SRCNN [14] is the first method that introduces deep learning into the field. ESPCN [12] proposes an upsampling method named pixel-shuffle. RDN [15] enlarges the neural network by combining Resnet and Densenet. Nowadays, to apply the SR method in practice, researchers explore real-world degradation. IKC [8] iteratively estimates the blur kernel and correspondingly super-resolves the LR image. DASR [5] adopts unsupervised methods to accelerate the blind SR process. MANET [4] explores the mutual information with the image feature for spatially variant degradation. CMOS [2] extends MANET with the guidance of semantic information.

However, these methods ignore the motion blur and generate HR images with severe artifacts when addressing the degradation compound of motion blur. Compared with these methods, our method manages to handle motion blur by customizing treatment per pixel.

3. METHODOLOGY

3.1. Problem Formulation

Blind SR methods aim to super-resolve LR images with unknown degradation. Mathematically, the degradation can be formulated as:

$$I^{LR} = (I^{HR} * k)_{\downarrow s} + n \tag{1}$$

where I^{LR}, I^{HR}, $*$, k, s, and n represent LR image, HR image, convolutional operation, blur kernel, downsampling scale, and additional noise. In this paper, we focus on the scenario in which the motion blur is involved in the degradation.

3.2. Our Method

3.2.1. Overview

In this paper, we propose deform**A**ble **rec**eptive **S**uper-**R**esolution (ArcSR) and depict the overall structure in Fig. 1. As shown in Fig. 1(a), the ArcSR first estimates the blur kernel for each pixel and then super-resolves the LR image.

For kernel estimation, we adopt a Unet structure as the backbone and view the downsampling process as encoder and the upsampling process as decoder. Each level in the encoder and decoder adopts the same structure, which adopts two DM-blocks which consist of multiple Deformable Mutual convolutions (DMconv) and employs skip-connection mechanism to avoid gradient vanishing. The details of DMconv are illustrated in section 3.2.2. Furthermore, to avoid the information missing during the downsampling process, we take the addition of outputs generated by the i^{th} level of the encoder and the $i-1^{th}$ level of the decoder as the input of the i^{th} level

of the decoder. Hence, we estimate the blur kernel matrix $k \in \mathbb{R}^{ks^2 \times H \times W}$, where ks, H, and W represent kernel size, height of the input image, and width of the input image.

After obtaining blur kernel matrix k, our ArcSR super-resolves the LR image with several KGblocks, and each KGblock is equipped with one channel-wise attention block. The KGblock rewrites the missing high-frequency information with the assistance of estimated blur kernels and the channel-wise attention block compels the network to give extra attention to missing textures, which is proved to be useful in low-level tasks (*e.g.*, SR, image restoration [16]). Concretely, the KGblock consists sequentially of a couple of KGconvs and adopts the skip-connection mechanism to aggregate the input and output of the KGblock. The details of KGconv are illustrated in section 3.2.3. Furthermore, the channel-wise attention block enhances features by generating attention matrices with a convolutional layer, an average pooling layer, and a full connection layer and multiplying with given features. After processed by multiple KGblocks and channel-wise attention blocks, the input feature is then upsampled by the pixel-shuffle operation and reconstructed to HR image by a convolutional layer.

3.2.2. Deformable Mutual Convolution

To estimate the blur kernel for each pixel, we propose the DMconv and depict the detail in Fig. 1(b). The DMconv delves the correlation between channels of image features with a deformable receptive field. Firstly, we split the image feature $F \in \mathbb{R}^{C \times H \times W}$ into two parts $F_s \in \mathbb{R}^{\frac{C}{4} \times H \times W}$ and $\hat{F}_s \in \mathbb{R}^{\frac{3C}{4} \times H \times W}$ along the channel dimension. Secondly, we conduct two convolutional layers on $\hat{F}_s$ to obtain the offset. Thirdly, we perform deformable convolution on F_s with the obtained offset. Finally, we concatenate F_s and $\hat{F}_s$ and adopt a convolutional layer to fuse the information within channels. Specifically, to fully explore the correlation, we employ 4 DMconv layers in DMblock, and different DMconv in one DMblock split the channel with different indexes.

3.2.3. Kernel Guided Convolution

To super-resolve the LR image with the assistance of estimated blur kernels, we propose the KGconv and illustrate the structure in Fig. 1(c). We believe that the estimated kernel demonstrates how pixels around the center pixel influence the center pixel and for each pixel, the proposed KGconv customizes convolutional parameters by utilizing the estimated kernel. Therefore, given one pixel, we view the corresponding blur kernel as the attention matrix for convolutional parameters. Specifically, we extract the corresponding blur kernel from the estimated kernel k, reshape it to an attention matrix $\in \mathbb{R}^{ks \times ks}$, and multiply with parameters within the convolutional layer. Thereafter, we perform convolutional operation on input features with the obtained customized parameters.

Method	DMconv	KGconv	REDS		GoPro	
			PSNR	SSIM	PSNR	SSIM
Baseline	✗	✗	23.63	0.64	24.17	*0.70*
ArcSR (w/o DMconv)	✗	✔	*23.76*	*0.65*	*24.28*	*0.71*
ArcSR (w/o KGconv)	✔	✗	*23.74*	*0.64*	*24.23*	*0.71*
ArcSR (full)	✔	✔	**24.04**	**0.66**	**24.33**	**0.72**

Table 1. **Quantitative results of ablation experiments on DMconv and KGconv**. Noise level is 0.

3.2.4. Loss Function

We adopt the L1 distance between the super-resolved HR image I^{SR} and the ground truth HR image I^{HR}. The formulation is formulated as:

$$\mathscr{L} = ||I^{SR}, I^{HR}||_1 \tag{2}$$

4. EXPERIMENTS

4.1. Dataset and Metrics

We adopt REDS [17] and GoPro [18] dataset as training set and testing set, as our method focuses on degradation contains with motion blur and commonly used SR datasets (*e.g.*, set5 [19], set14 [20], BSD100 [21], Urban100 [22], Manga109 [23, 24]) contains no motion blur kernel. Following previous works [2, 3, 4], we adopt Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and visual performance for comparison metrics.

4.2. Implemenataion Details

As for the structure of our ArcSR, we employ 4 DMconvs in DMblock and 2 DMblocks in each level during the kernel estimation process. Furthermore, we also adopt 3 KGconvs in KGblock and 8 KGblocks during the SR process. For paired data construction, we synthetic LR images for training and testing. Concretely, the REDS dataset provides LR images blurred by motion blur, while the GoPro dataset does not, thereby we downsample the GoPro dataset with bicubic interpolation method and add gaussian noise on images in these two datasets. The level of gaussian noise is set to (0.2, 20). We also adopt a data argument strategy that crops HR image patches with 256 $\times$ 256 and rotation them with 90, 180, and 270 degrees and correspondingly process the LR images. For optimization, we adopt the Adam optimizer and set the learning rate to 1e -4 and β to 0.9. We train ArcSR in 4 $\times$ 10^5 steps and adopt multi-step learning strategy. The learning is decreased by 0.5 at 2 $\times$ 10^4, 4 $\times$ 10^4, 8 $\times$ 10^4, and 2 $\times$ 10^5 step.

4.3. Ablation Studies

To systematically evaluate the effectiveness of each module of our proposed ArcSR, we train the model by removing each component and present the quantitative results on $\times$ 4 SR

Fig. 2. **Qualitative comparison on $\times$ 4 SR on GoPro dataset [18]**. Our method manages to suppress the ghost edge phenomenon and removes most motion blur.

Noise Level	Method	REDS		GoPro	
		PSNR	SSIM	PSNR	SSIM
0	DASR [5]	23.86	0.65	23.27	*0.72*
	MANET [4]	*23.95*	**0.66**	*23.89*	*0.72*
	UDKE [1]	23.86	0.64	23.82	*0.72*
	ReDegNet [3]	22.56	*0.65*	17.80	0.66
	CMOS* [2]	23.94	*0.65*	*23.89*	0.71
	ArcSR (Ours)	**24.04**	**0.66**	**24.33**	**0.73**
10	DASR [5]	23.36	*0.61*	22.70	0.69
	MANET [4]	23.45	**0.63**	*23.39*	**0.72**
	UDKE [1]	*23.47*	*0.61*	**23.47**	*0.71*
	ReDegNet [3]	18.06	0.58	17.32	0.63
	CMOS* [2]	23.46	*0.61*	*23.47*	0.70
	ArcSR (Ours)	**23.48**	**0.63**	**23.47**	**0.72**

Table 2. **Quantitative results of $\times$ 4 SR comparison experiments**. Our method achieves the highest PSNR and SSIM scores. $*$ denotes our re-implementation.

for noise-free degradation on Table 1. In the main ablation study, we design the following ablation models: (1) **Baseline Model**: we replace the DMconv with conventional convolution and replace the KGconv with a sequential operation which concatenates the blur kernel and the image feature and then performs convolutional operation on the obtained matrix; (2) **ArcSR (w/o DMconv)**: we replace the DMconv with conventional convolution; (3) **ArcSR (w/o KGconv)**: we replace the KGconv with the sequential operation; (4) **ArcSR (full)**: our full ArcSR model.

The results demonstrate that both DMconv and KGconv have positive influence on blind SR task. Specifically, from the observation of ArcSR (w/o DMconv, KGconv) and ArcSR (w/o DMconv), compared with simply concatenating estimated kernels and images features together which expands channels of features, KGconv introduces no additional parameters and gains 0.13dB and 0.11dB PSNR score on REDS [17] and Gopro [18] datasets by utilizing estimated kernels to customize convolutional parameters. Meanwhile, from the observation of **ArcSR (w/o DMconv, KGconv)** and **ArcSR (w/o KGconv)**, DMconv gains 0.11dB and 0.06 dB on REDS [17] and Gopro [18] datasets by exploring the correlation within channels of images features with a deformable receptive field. The results of **ArcSR (w/o DMconv)** and **ArcSR (w/o KGconv)** are inferior to the full ArcSR method, verifying the effectiveness of both components.

4.4. Comparison with the State-of-the-Arts

To demonstrate the superiority of our method, we conduct comparison experiments and select state-of-the-art methods DASR [5], MANET [4], UDKE [1], RedegNet [3], and CMOS [2] as our baselines. To be fair, we retrain these methods on GoPro [18] and REDS [17] datasets and adopt the same training strategy as our method except RedegNet, since RedegNet argues that neural network trained on degradation synthesized from pairs collected from natural face images is more robust to real-world data. Following previous methods [4, 5], we conduct experiments on noise-free degradation and noise degradation which the noise level of additional gaussian noise is configured to 10.

Quantitative Results Quantitative results shown on Table 2 demonstrate that our ArcSR achieves the state-of-the-art PSNR and SSIM scores on degradation with motion blur. Specifically, our method achieves the highest PSNR and SSIM scores on both noise-free and noise degradation. Especially, when handling noise-free degradation, our ArcSR outperforms the state-of-the-art method 0.09dB and 0.44dB on GoPro [18] and REDS [17] datasets.

Visual Results The result of qualitative comparison is shown on Fig. 2, compared with other baselines (*e.g.*, MANET [4], UDKE [1]), our ArcSR removes the most motion blur and manages to suppress the ghost edge phenomenon.

5. CONCLUSION

In this paper, we propose a novel blind image super-resolution method named ArcSR to address the degradation contained motion blur issue. Specifically, we argue blur kernels of each pixel in LR images have discrepancies and each pixel requires individual processing. Therefore, we propose DMconv to explore the correlation within the channels of image features by obtaining deformable receptive field. Furthermore, we propose KGconv to give each pixel unique convolutional parameters by viewing the estimated blur kernel as an attention matrix for convolutional parameters. Sufficient ablation experiments demonstrate functions of our DMconv and KGconv and comprehensive comparison experiments prove the superiority of our ArcSR method.

6. REFERENCES

[1] Hongyi Zheng, Hongwei Yong, and Lei Zhang, "Unfolded deep kernel estimation for blind image super-resolution," in *ECCV*. Springer, 2022, pp. 502–518.

[2] Xuhai Chen, Jiangning Zhang, Chao Xu, Yabiao Wang, Chengjie Wang, and Yong Liu, "Better" cmos" produces clearer images: Learning space-variant blur estimation for blind image super-resolution," in *CVPR*, 2023, pp. 1651–1661.

[3] Xiaoming Li, Chaofeng Chen, Xianhui Lin, Wangmeng Zuo, and Lei Zhang, "From face to natural image: Learning real degradation for blind image super-resolution," in *ECCV*. Springer, 2022, pp. 376–392.

[4] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Mutual affine network for spatially variant kernel estimation in blind image super-resolution," in *ICCV*, 2021, pp. 4096–4105.

[5] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo, "Unsupervised degradation representation learning for blind super-resolution," in *CVPR*, 2021, pp. 10581–10590.

[6] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, Weiping Wang, and Dan Meng, "What matters: Attentive and relational feature aggregation network for video-text retrieval," in *ICME*, 2021, pp. 1–6.

[7] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, and Weiping Wang, "Multi-feature graph attention network for cross-modal video-text retrieval," in *ICMR*, 2021, pp. 135–143.

[8] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong, "Blind super-resolution with iterative kernel correction," in *CVPR*, 2019, pp. 1604–1613.

[9] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li, "Dual alignment unsupervised domain adaptation for video-text retrieval," in *CVPR*, 2023, pp. 18962–18972.

[10] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li, "Mixgen: A new multi-modal data augmentation," in *WACVW*, 2023, pp. 379–389.

[11] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li, "Listen and look: Multi-modal aggregation and co-attention network for video-audio retrieval," in *ICME*, 2022, pp. 1–6.

[12] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016, pp. 1874–1883.

[13] Xiaoshuai Hao and Wanqian Zhang, "Uncertainty-aware alignment network for cross-domain video-text retrieval," in *NIPS*, 2023.

[14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*. Springer, 2014, pp. 184–199.

[15] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *CVPR*, 2018, pp. 2472–2481.

[16] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun, "Simple baselines for image restoration," in *ECCV*. Springer, 2022, pp. 17–33.

[17] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in *CVPR Workshops*, June 2019.

[18] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, July 2017.

[19] Bevilacqua M, Roumy A, and Guillemot C et al., "Low-complexity single-image super-resolution based on non-negative neighbor embedding," in *BMVC*, 2012, pp. 135.1–135.10.

[20] Zeyde R, Elad M, and Protter M., "On single image scale-up using sparse-representations," in *Curves and Surfaces*, 2012, pp. 711–730.

[21] Martin D, Fowlkes C, and Tal D et al., "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001, pp. 416–423.

[22] Huang J, Singh A, and Ahuja N., "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015, pp. 5197–5206.

[23] Matsui Y, Ito K, and Aramaki Y et al., "Sketch-based manga retrieval using manga109 dataset," *MTA*, pp. 21811–21838, 2017.

[24] Aizawa K, Fujimoto A, and Otsubo A et al., "Building a manga dataset "manga109" with annotations for multimedia applications," *IEEE MultiMedia*, pp. 8–18, 2020.