Better "CMOS" Produces Clearer Images: Learning Space-Variant Blur Estimation for Blind Image Super-Resolution

Xuhai Chen[1] Jiangning Zhang[2*] Chao Xu[1] Yabiao Wang[2] Chengjie Wang[2] Yong Liu[1†]
[1] APRIL Lab, Zhejiang University [2]Youtu Lab, Tencent
{22232044, 21832066}@zju.edu.cn, yongliu@iipc.zju.edu.cn
{vtzhang, caseywang, jasoncjwang}@tencent.com

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.00165

Abstract

Most of the existing blind image Super-Resolution (SR) methods assume that the blur kernels are space-invariant. However, the blur involved in real applications are usually space-variant due to object motion, out-of-focus, etc.*, resulting in severe performance drop of the advanced SR methods. To address this problem, we firstly introduce two new datasets with out-of-focus blur,* i.e.*, NYUv2-BSR and Cityscapes-BSR, to support further researches of blind SR with space-variant blur. Based on the datasets, we design a novel* ***C****ross-****MO****dal fu****S****ion network (CMOS) that estimate both blur and semantics simultaneously, which leads to improved SR results. It involves a feature Grouping Interactive Attention (GIA) module to make the two modalities interact more effectively and avoid inconsistency. GIA can also be used for the interaction of other features because of the universality of its structure. Qualitative and quantitative experiments compared with state-of-the-art methods on above datasets and real-world images demonstrate the superiority of our method,* e.g.*, obtaining PSNR/SSIM by +1.91↑/+0.0048↑ on NYUv2-BSR than MANet*[1].

1. Introduction

Blind image SR, with the aim of reconstructing High-Resolution (HR) images from Low-Resolution (LR) images with unknown degradations, has attracted great attention due to its significance for practical use [2,5,6,12,15,22–24, 29]. Two degradation models, bicubic downsampling [35] and traditional degradation [26,32], are usually used to generate LR images from HR images. The latter can be modeled by:

$$\boldsymbol{y} = (\boldsymbol{x} \bigotimes \boldsymbol{k}) \downarrow_s + \boldsymbol{n}. \tag{1}$$

It assumes the LR image $\boldsymbol{y}$ is obtained by first convolving the HR image $\boldsymbol{x}$ with a blur kernel $\boldsymbol{k}$, followed by a down-sampling operation with scale factor s and an addition of noise $\boldsymbol{n}$. On top of that, some works [38, 48] propose more complex and realistic degradation models, which also assume that blur is space-invariant. However, in real-world applications, blur usually changes spatially due to factors such as out-of-focus and object motion, so that the mismatches will greatly degrade the performance of existing SR methods. Fig. 1 gives an example when the LR image suffers from space-variant blur. Since both KernelGAN [1] and DCLS [28] estimate only one blur kernel for an image, there are a lot of mismatches. In the first row of Fig. 1, where the kernel estimated by the two methods are sharper than the real one of the patch, SR results are over smoothing and high frequency textures are significantly blurred. In the second row, where the kernels estimated are smoother than the correct one, SR results show ringing artifacts caused by over-enhancing high-frequency edges. This phenomenon illustrates that mismatch of blur will significantly affect SR results, leading to unnatural outputs. In this paper, we focus on the space-variant blur estimation to ensure that the estimated kernel is correct for each pixel in the images.

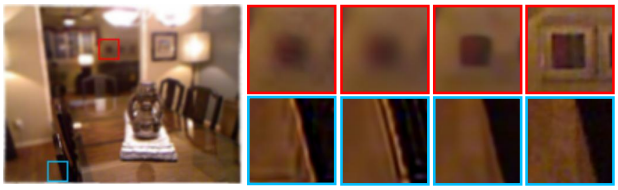


Figure 1. SR results of KernelGAN [1], DCLS [28] and the proposed CMOS on a space-variant blurred LR image. For KernelGAN and DCLS, patches are blurry in the first row and have artifacts in the second row, while CMOS performs well in both cases.

A few recent works [15, 23, 43] have taken space-variant blur into account. Among them, MANet [23] is the most representative model, which assumes that blur is space-invariant within a small patch. Based on this, MANet uses a moderate receptive field to keep the locality of degradations. However, there are still two critical issues. 1) Because there

*Equal contribution.
†Corresponding author.
[1]https://github.com/ByChelsea/CMOS.git

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.00165

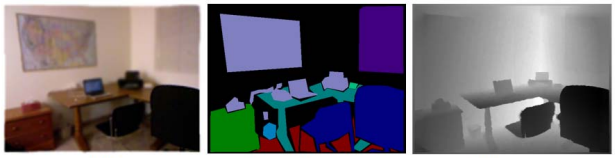


Figure 2. A condition in which blur and semantic information are inconsistent. This image comes from our dataset NYUv2-BSR.

is no available dataset containing space-variant blur in SR field, MANet is trained on space-invariant images, resulting in blur deviation of the training and testing phase. 2) Even limiting the size of the receptive field, the estimation results are still poor at the boundaries of different kernels, leading to mean value prediction of space-variant blur.

To address the aforementioned challenges, we first introduce a new degradation method and propose two corresponding datasets, *i.e.*, NYUv2-BSR and Cityscapes-BSR to support relevant researches of space-variant blur in the SR domain. As a preliminary exploration, out-of-focus blur is studied as an example in this paper and it is generated according to the depth of the objects using the method proposed in [19]. Besides, we also add some space-invariant blur into the datasets so that the models trained on them can cope with both spatially variant and invariant situations.

Furthermore, to improve the performance at the boundaries of different blur regions, we present a novel model named *Cross-MOdal fuSion network* (CMOS). Our intuition is that the sharp semantic edges are usually aligned with out-of-focus blur boundaries and it can help to distinguish different blur amounts. This raises a critical concern that how to effectively introduce semantics into the process. Specifically, we firstly predict blur and semantics simultaneously instead of using the semantics as an extra input, which not only avoids using extra information during test phase, but also enables non-blind SR methods to recover finer textures with the two modalities. Secondly, to enhance accuracy at the blur boundaries, we conduct interaction between the semantic and blur features for complementary information learning inspired by multi-task learning [36, 42]. However, in some cases these two modalities are inconsistent. As shown in Fig. 2, the wall and the picture on it are completely different in the semantic map, with clear boundaries. But the depth of them are almost the same, so the blur amounts depending on depth are also very similar. In this case, not only can the two modalities fail to use common features, but they can also negatively influence each other. Besides, since we add some space-invariant blurred images with uniform blur maps in the datasets, it will also greatly increase the inconsistency.

Motivated by these observations, we propose a feature Grouping Interactive Attention (GIA) module to help the interaction of the two modalities. GIA has two parallel streams: one operating along the spatial dimension and the other along the channel dimension. Both streams employ group interactions to process the input features and make adjustments. Moreover, GIA has an upsampling layer based on the flow field [21] to support inputs of different resolutions. Its universal structure allows it to be used for more than just interactions between the two modalities.

The main contributions of this work are as follows:

- To support researches on space-variant blur in the field of SR, we introduce a new degradation model of out-of-focus blur and propose two new datasets, *i.e.*, NYUv2-BSR and Cityscapes-BSR.
- We design a novel model called CMOS for estimating space-variant blur, which leverages extra semantic information to improve the accuracy of blur prediction. The proposed GIA module is used to make the two modalities interact effectively. Note that GIA is universal and can be used between any two features.
- Combined with existing non-blind SR models, CMOS can estimate both space-variant and space-invariant blur and achieve SOTA SR performance in both cases.

2. Related Work

2.1. Degradation Model

SR methods give rise to poor performance if the assumed degradation deviates from those in reality. Many works [4, 45, 49] use the traditional model (Eq. 1) to generate their training data. Compared to bicubic downsampling [40, 50], although traditional model has taken more factors into account, it is still too simple to simulate real degradation. Consequently, Real-ESRGAN [38] proposes a flexible high-order degradation model by applying traditional model repeatedly, while BSRGAN [48] adjusts the degradation order of the traditional model and use randomly shuffled blur, downsampling and noise. Liang *et al.* [23] go a step further to simulate space-variant blur by dividing images into patches and applying different kernels. Unfortunately, it cannot well simulate the real situations. As a result, to support relevant researches, we introduce space-variant out-of-focus blur into SR, and propose two corresponding datasets, *i.e.*, NYUv2-BSR and Cityscapes-BSR.

2.2. Kernel Estimation

One of the mainstream methods of blind SR is to estimate degradation first and then use it as prior information for non-blind SR. KernelGAN [1] proposes to learn a kernel from the internal distribution of image patches, while IKC [6] uses an iterative correction scheme to learn the PCA features of kernels. Luo *et al.* [28] transfer blur estimation into LR space and learn kernel weights instead of kernel itself. However, these methods only estimate a unique kernel,

Figure 3. Original RGB images, the generated out-of-focus images and blur maps. The changes from dark to light in blur maps indicate that the corresponding out-of-focus image changes from clear to blur. The first three columns are images from NYUv2-BSR, and the last three columns are images from Cityscapes-BSR.

thus the performance will be significantly reduced on space-variant situations. Accordingly, KOALAnet [15] proposes to learn specific kernels for each pixel, and MANet [23] designs a network with moderate receptive field to adapt to the locality of degradation. However, they still have limitations, such as the moderate receptive field might limit the capacity of the model. By contrast, with the help of semantic information, our CMOS can predict space-variant blur effectively and accurately.

2.3. Non-blind SR

Non-blind SR aims to restore images with known degradations. Early non-blind SR methods [13, 14, 18, 25] are based on bicubic downsampling, which struggle to generalize to images with more complex degradations. To address this problem, SRMD [49] first proposes to stretch the blur and noise to the size of LR images, and take the concatenated images and degradation maps as input to restore the HR counterparts. Following SRMD, SFTMD [6] uses SFT layer [39] to combine the stretching degradation maps instead of simply concatenation, while UDVD [44] employs per-pixel dynamic convolution to more effectively deal with variational degradations across images. Besides, zero-shot methods [11, 32, 34] have also been investigated in non-blind SR with multiple degradations. What is noteworthy is that our CMOS can be easily combined with most non-blind SR methods to achieve excellent blind SR performance.

3. The Proposed Datasets

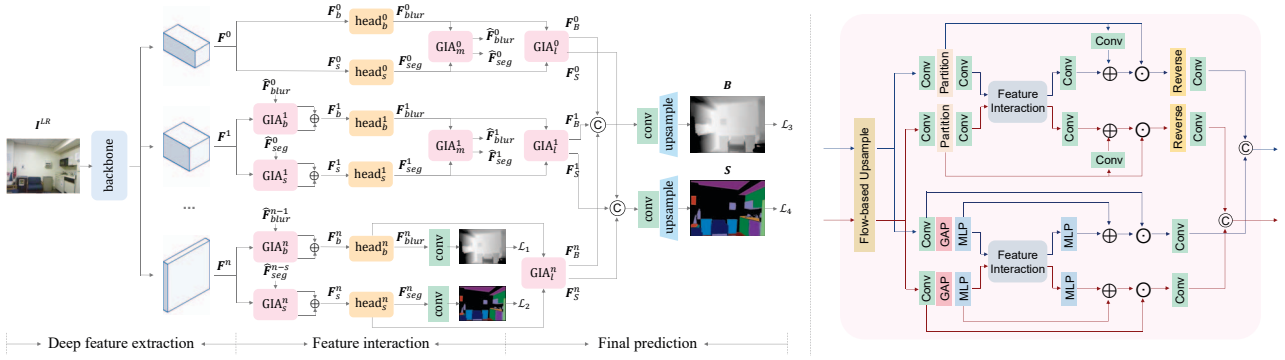
To support researches on space-variant blur, we propose two novel datasets, NYUv2-BSR and Cityscapes-BSR, where BSR stands for Blind image SR. To the best of our knowledge, we are the first to introduce out-of-focus, one of the most common space-variant blur in real world, into blind image SR. Out-of-focus is caused by differences in depth. Every point that is not in the plane of focus corresponds to a Circle Of Confusion (COC) in image plane. The blur can be simulated by isotropic Gaussian kernels with standard deviation σ related to the diameters of COCs [17], which can be calculated using thin lens model [30]. We employ the method proposed in [19] to blur the images and the ground truth blur map is constructed by σ of each pixel.

Dataset	NYUv2-BSR			Cityscapes-BSR		
	VA	IVA	Total	VA	IVA	Total
Train	636	159	795	2380	595	2975
Val	-	-	-	400	100	500
Test	524	130	654	1220	305	1525

Table 1. Details of NYUv2-BSR and Cityscapes-BSR. VA and IVA represents the number of images with space-variant out-of-focus blur and space-invariant blur respectively.

As mentioned above, we need depth-color image pairs to generate images with out-of-focus blur. Thus, we select NYUv2 [33] and Cityscapes [3] as original datasets. NYUv2 is an indoor dataset. It contains 1449 pairs of RGB and depth images, in which 795 pairs are used for training and the rest 654 for testing. Cityscapes is an outdoor dataset and the fine-annotated part consists of training, validation and test sets containing 2975, 500, and 1525 images, respectively. Since the depth maps in Cityscapes contain invalid measurements, which are not conducive to the generation of out-of-focus images, we use CREStereo [20], a deep learning-based stereo matching method, to generate disparity maps and calculate the corresponding depth maps based on the camera parameters. Fig. 3 shows the original RGB images of NYUv2 and Cityscapes, as well as the generated out-of-focus images and corresponding blur maps.

In terms of parameters of the isotropic Gaussian kernels, the kernel width range is set to $[0.0, 5.0]$ and $[0.0, 15.0]$ for NYUv2 and Cityscapes, respectively. The kernel size is fixed to 21×21 and 61×61, and the downsampling scale



(a) Overview of the proposed CMOS (b) Sturcture of GIA module

Figure 4. **Architecture of CMOS and GIA.** (a) Given an LR image, CMOS outputs the estimated blur map $\boldsymbol{B}$ and semantic map $\boldsymbol{S}$ simultaneously in the HR space. (b) GIA has two parallel streams to effectively interact features in both spatial and channel dimensions. It includes a flow-based upsample module to support inputs with different resolutions. If the input resolutions are the same, a feature alignment will also be performed through the learned flow field.

factor is set to 4. Besides, 1/4 of the images are blurred by space-invariant kernels, so that the models trained on the datasets are not limited by the space-variant situations. Tab. 1 shows the details. In addition, to ensure the adequacy and fairness of experiments, we created five test groups for each dataset, in which each group had a different 1/4 of the images blurred by space-invariant kernels.

4. Method

As stated before, sharp semantic edges can increase the accuracy of space-variant blur estimation near the boundaries. Motivated by this, we propose a Cross-MOdal fuSion network (CMOS) to predict both blur and semantic maps simultaneously by mutual supervision of them.

4.1. Overview

Inspired by [36], CMOS is a multi-scale network, which consists of three main stages, as shown in Fig. 4. In the first stage, a fully convolutional encoder capable of generating multi-scale features is used to extract deep features $\{\boldsymbol{F}^0, \boldsymbol{F}^1, \cdots, \boldsymbol{F}^n\}$. In the next stage, for each scale i, we apply two task-specific heads, head_b^i and head_s^i, to predict initial blur and semantic features $\boldsymbol{F}_{blur}^i$ and $\boldsymbol{F}_{seg}^i$. Then, we use a proposed GIA_m^i module to achieve effective information interaction between the two modalities to obtain more accurate features $\hat{\boldsymbol{F}}_{blur}^i$ and $\hat{\boldsymbol{F}}_{seg}^i$ in a mutually supervised manner, formulated as:

$$\boldsymbol{F}_{blur}^i = \text{head}_b^i(\boldsymbol{F}_b^i), \tag{2}$$

$$\boldsymbol{F}_{seg}^i = \text{head}_s^i(\boldsymbol{F}_s^i), \tag{3}$$

$$\hat{\boldsymbol{F}}_{blur}^i, \hat{\boldsymbol{F}}_{seg}^i = \text{GIA}_m^i(\boldsymbol{F}_{blur}^i, \boldsymbol{F}_{seg}^i), \tag{4}$$

where $\boldsymbol{F}_b^i$ and $\boldsymbol{F}_s^i$ denotes the input of the task-specific heads. To make better use of the multi-scale information, we use GIA_b^i and GIA_s^i to fuse the adjacent low-scale features, so the input of the heads can be written as:

$$\boldsymbol{F}_b^0 = \boldsymbol{F}_s^0 = \boldsymbol{F}^0, \tag{5}$$

$$\boldsymbol{F}_b^i = \text{Sum}(\text{GIA}_b^i(\boldsymbol{F}^i, \hat{\boldsymbol{F}}_{blur}^{i-1})), \tag{6}$$

$$\boldsymbol{F}_s^i = \text{Sum}(\text{GIA}_s^i(\boldsymbol{F}^i, \hat{\boldsymbol{F}}_{seg}^{i-1})), \tag{7}$$

where $\text{Sum}(\cdot)$ represents for adding outputs of the modules. At the highest resolution n, the task-specific features are fed into two convolution layers to generate auxiliary blur and semantic maps for additional supervision, which is beneficial to further improve the accuracy of the final prediction.

The last stage consists $n+1$ GIA_l^i modules to get the final features $\boldsymbol{F}_B^i$ and $\boldsymbol{F}_S^i$ of each scale as:

$$\boldsymbol{F}_B^i, \boldsymbol{F}_S^i = \text{GIA}_l^i(\boldsymbol{F}_{blur}^i, \boldsymbol{F}_{seg}^i). \tag{8}$$

These features are then concated and convolved to obtain the prediction of blur and semantic maps. In this way, we can build a shorter way for each scale to the supervision and further facilitate the interaction between blur and semantics. Besides, since blur is done in the HR space, we upsample the outputs using bi-linear interpolation by scale factor s.

4.2. Grouping Interactive Attention Module

GIA is designed to help blur and semantics interact more effectively and avoid inconsistency. Besides, it can also be used for other features because of the universal structure. GIA involves two parallel streams operating on spatial and channel dimensions, and it can handle inputs of different resolutions by using a flow-based upsample module [21].

Spatial Grouping Feature Interaction. The input features may be similar on most patches, but different on some. As shown in Fig. 2, the picture hanging on the wall brings difference between the blur and semantic maps. As a result,

we propose to adjust the spatial weight map in the general spatial attention [7, 46, 47] mechanism to take advantage of similar information and avoid inconsistencies.

In the top half of Fig. 4 (b), each input is first passed through a convolution layer and divided into windows denoted by $\boldsymbol{F}_w^j$. These windows are then further processed by another convolution layer before being fed into the feature interaction module (last part of Sec. 4.2). The spatial adjusting weight map $\boldsymbol{M}_a^j \in \mathbb{R}^{1\times H\times W}$ can be obtained by a 1×1 convolution layer after the interaction. Additionally, each input has its own spatial weight map $\boldsymbol{M}_o^j \in \mathbb{R}^{1\times H\times W}$ extracted from the windows by another 1×1 convolution layer directly. Thus, the outputs $\boldsymbol{F}^j$ corresponding to the two inputs can be expressed as:

$$\boldsymbol{F}^j = \mathrm{Mul}(\boldsymbol{F}_w^j, \mathrm{Add}(\boldsymbol{M}_o^j, \alpha \boldsymbol{M}_a^j)), j = 1, 2, \quad (9)$$

where α is a learnable parameter. Finally, windows are restored as features and final output is obtained by smoothing out possible seams with a layer of 3×3 convolution.

Channel Grouping Feature Interaction. As spatial feature interaction concentrates on local details, we further introduce channel grouping feature interaction to calibrate global information inspired by [8]. Firstly, we transfer the input $\boldsymbol{F}_{in}^j$ to channel-wised attention vector $\boldsymbol{A}_o^j \in \mathbb{R}^C$ by applying global average pooling and an MLP layer. Then, the vectors are fed into a feature interaction module, and two adjusting attention vectors $\boldsymbol{A}_a^j \in \mathbb{R}^C$ integrating the two features are obtained through another MLP layer. Similar to the spatial one, the final outputs can be obtained by:

$$\boldsymbol{F^j} = \mathrm{Mul}(\boldsymbol{F}_{in}^j, \mathrm{Add}(\boldsymbol{A}_o^j, \beta \boldsymbol{A}_a^j)), j = 1, 2 \quad (10)$$

where β is a learnable parameter. Since global information is important for both blur [31] and semantic estimation [27], feature interaction of channel dimension is also essential.

Feature Group Interaction. This module is designed to interact spatial or channel features in groups. For spatial interaction, the input size is $C \times H \times W$. We regard the features of each pixel as a group, and the size of the grouped features is $N \times D$, where $N = HW$, $D = C$. For channel interaction, the input size is C. It will be divided into N groups with length D, where $C = ND$. In this way, both spatial and channel inputs can be represented as $\boldsymbol{G}_i \in \mathbb{R}^{N\times D}, i = 1, 2$ after grouping, where i represents two different inputs. Then, we use inner product for feature interaction and get the interactive feature $\boldsymbol{F}_{fuse} \in \mathbb{R}^{N\times N}$,

$$\boldsymbol{F}_{fuse} = \boldsymbol{G}_1 {\boldsymbol{G}_2}^T. \quad (11)$$

After that, for spatial interaction, one of the output can be obtained by reshaping $\boldsymbol{F}_{fuse}$ to $H \times W \times N$, and the other can be obtained by reshaping $\boldsymbol{F}_{fuse}$ to $N \times H \times W$. For channel interaction, the two final outputs are the same and can be both obtained by simply flatten $\boldsymbol{F}_{fuse}$.

4.3. Loss Function

We use the mean absolute error (MAE) for blur estimation and the cross-entropy (CE) loss for semantic segmentation. As shown in Fig. 4 (a), the auxiliary loss $\mathcal{L}_1$ and loss $\mathcal{L}_3$ are both MAE, while the auxiliary loss $\mathcal{L}_2$ and loss $\mathcal{L}_4$ are both CE, specifically:

$$\mathcal{L}_1 = \mathcal{L}_3 = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \|\boldsymbol{B}_{i,j} - \hat{\boldsymbol{B}}_{i,j}\|_1 \quad (12)$$

$$\mathcal{L}_2 = \mathcal{L}_4 = -\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{C} \boldsymbol{S}_{i,j}^c \log(\hat{\boldsymbol{S}}_{i,j}^c) \quad (13)$$

where $\hat{\boldsymbol{B}}_{i,j}$ and $\boldsymbol{B}_{i,j}$ denote the estimated blur map and the corresponding ground-truth at position (i, j). Similarly, $\hat{\boldsymbol{S}}_{ij}^c$ and $\boldsymbol{S}_{ij}^c$ represent the estimated semantic map and the ground-truth at position (i, j) of the c-th category. C is the number of object categories, and H, W are the height and width of the maps. We do not adopt a particular loss weighing strategy, but simply sum the losses together,

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \mathcal{L}_4 \quad (14)$$

5. Experiments

5.1. Experimental Setup

Settings of CMOS. We select HRNet [37] as our backbone and change the stride of the first two convolutions to 1. This translates to 4 scales of the input LR images (1, 1/4, 1/8, 1/16). The task-specific heads are implemented as two basic residual blocks [9]. As for semantic segmentation, we use the official 40 classes for NYUv2-BSR and 19 classes for Cityscapes-BSR. All our experiments are conducted with the pre-trained ImageNet weights.

Settings of Non-Blind SR. For non-blind SR, we use RRDB-SFT proposed in [23]. To feed both blur and semantics into it, we use a GIA module. Finally, we finetune RRDB-SFT on blur and semantic maps estimated by CMOS. The loss between SR and HR images is also MAE.

Implementation Details. The image sizes are selected as 640×480 for both NYUv2-BSR and Cityscapes-BSR. We augment the training data by scaling with a randomly selected ratio in $\{1, 1.2, 1.5\}$ and the blur values are divided by the ratio. We also flip the training samples with a possibility of 0.5. Adam optimizer [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is used to train the model for 700 epochs, with a batch size of 8. The learning rate is initialized as 0.0001 and a cosine learning rate schedule with 10 warm-up epochs is adopted. Implemented with PyTorch, it takes about 28 hours to train CMOS on an RTX 3090 GPU.

Evaluation Metrics. For blur estimation, we use PSNR and SSIM [41]. For semantic segmentation, we use mIoU. For the final SR images generated by RRDB-SFT with the

Method	Group1	Group2	Group3	Group4	Group5	Avg.
KernelGAN [1]	23.10/0.7430	23.13/0.7439	23.18/0.7449	23.16/0.7449	23.18/0.7461	23.15/0.7446
KOALAnet [15]	27.69/0.8773	27.73/0.8768	27.60/0.8734	27.73/0.8754	27.74/0.8760	27.70/0.8758
DCLS [28]	27.89/0.8799	27.94/0.8798	27.82/0.8760	27.91/0.8768	27.89/0.8781	27.89/0.8781
DAN [10]	27.90/0.8809	27.98/0.8808	27.83/0.8771	27.91/0.8775	27.88/0.8791	27.90/0.8791
MANet [23]	30.16/0.9117	30.20/0.9111	30.07/0.9095	30.07/0.9099	30.10/0.9107	30.12/0.9106
CMOS(ours)	32.09/0.9168	32.08/0.9159	31.99/0.9145	31.96/0.9147	32.01/0.9153	32.03/0.9154
Upper Bound	33.80/0.9309	33.78/0.9303	33.69/0.9290	33.73/0.9301	33.74/0.9298	33.75/0.9300

Table 2. Average PSNR/SSIM of different methods for spatially variant blind SR on NYUv2-BSR. Avg. represents the average results on the 5 test groups. The best and second best results are highlighted in red and blue colors, respectively.

Method	Group1	Group2	Group3	Group4	Group5	Avg.
KernelGAN [1]	28.96/0.8461	29.02/0.8475	28.88/0.8464	28.96/0.8468	28.99/0.8477	28.96/0.8469
KOALAnet [15]	32.40/0.9173	32.45/0.9177	32.29/0.9149	32.38/0.9166	32.40/0.9166	32.38/0.9166
DCLS [28]	32.41/0.9174	32.46/0.9176	32.28/0.9151	32.44/0.9168	32.38/0.9166	32.39/0.9167
DAN [10]	32.33/0.9162	32.38/0.9165	32.21/0.9140	32.36/0.9156	32.30/0.9155	32.32/0.9156
MANet [23]	34.24/0.9293	34.29/0.9294	34.16/0.9273	34.27/0.9288	34.27/0.9285	34.25/0.9287
CMOS(ours)	35.58/0.9388	35.61/0.9389	35.50/0.9373	35.60/0.9385	35.60/0.9381	35.58/0.9383

Table 3. Average PSNR/SSIM of different methods for spatially variant blind SR on Cityscapes-BSR. Avg. represents the average results on the 5 test groups. Note that, there is no official ground truth semantic maps for the test sets of Cityscapes [3], so the upper bound is not available here. The best and second best results are highlighted in red and blue colors, respectively.

Datasets	PSNR ↑	SSIM ↑
IVA	19.50	0.6840
NYUv2-BSR	**30.12**	**0.9106**

Table 4. Importance of using space-variant blur for training. IVA stands for the NYUv2 dataset with only space-invariant blur.

blur and semantic maps estimated by CMOS, we compare PSNR/SSIM on the Y channel of YCbCr space.

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

We compare CMOS with existing blind SR models: KernelGAN [1], KOALAnet [15], DCLS [28], DAN [10], MANet [23] and the upper bound model (RRDB-SFT given ground-truth blur and semantic maps). We retrained all the comparison methods on NYUv2-BSR and Cityscapes-BSR using their official implementations and settings. KernelGAN is an unsupervised method which trained solely on the LR image at test time. DCLS and DAN are end-to-end methods for space-invariant blur, while KOALAnet and MANet are two-stage methods for space-variant blur. Since we use the non-blind SR model proposed in MANet (*i.e.* RRDB-SFT), we apply same settings to ensure the fairness.

Quantitative comparison. As shown in Tab. 2 and Tab. 3, CMOS leads to the best performance for different test groups in both the two proposed datasets. Notably, methods that estimate only one blur kernel for an image (*i.e.*, KernelGAN, DCLS, and DAN) all suffer from severe performance drop when the real kernels are spatially variant. Although KOALAnet estimates different kernels for different image pixels, it does not include any special handling for space-variant properties and also produces unfavorable results. MANet takes the locality of blur into account, so it performs relatively better. By contrast, the proposed model CMOS effectively utilizes semantic information to help with spatially variant blur estimation and non-blind SR, outperforming MANet by large margins.

Qualitative comparison. We present several representative visual samples in Fig. 5. It can be observed that our CMOS outperforms previous approaches in both removing blur and avoiding artifacts. Other methods may either produce ringing artifacts (especially KernelGAN), or fail to restore texture details, leading the patches still blurry.

5.3. Ablation Study

All the experiments in this section use NYUv2-BSR for training, and the metrics (*i.e.* PSNR, SSIM and mIoU) refer to the mean value across the 5 test sets (Sec. 3).

Importance of Using Space-Variant Blur for Training. According to [23], because of the moderate receptive field, MANet can handle spatially variant cases even if it is trained on spatially invariant blurred images. But we believe that it is necessary to use the images containing space-variant blur for training. To prove it, we trained two MANet

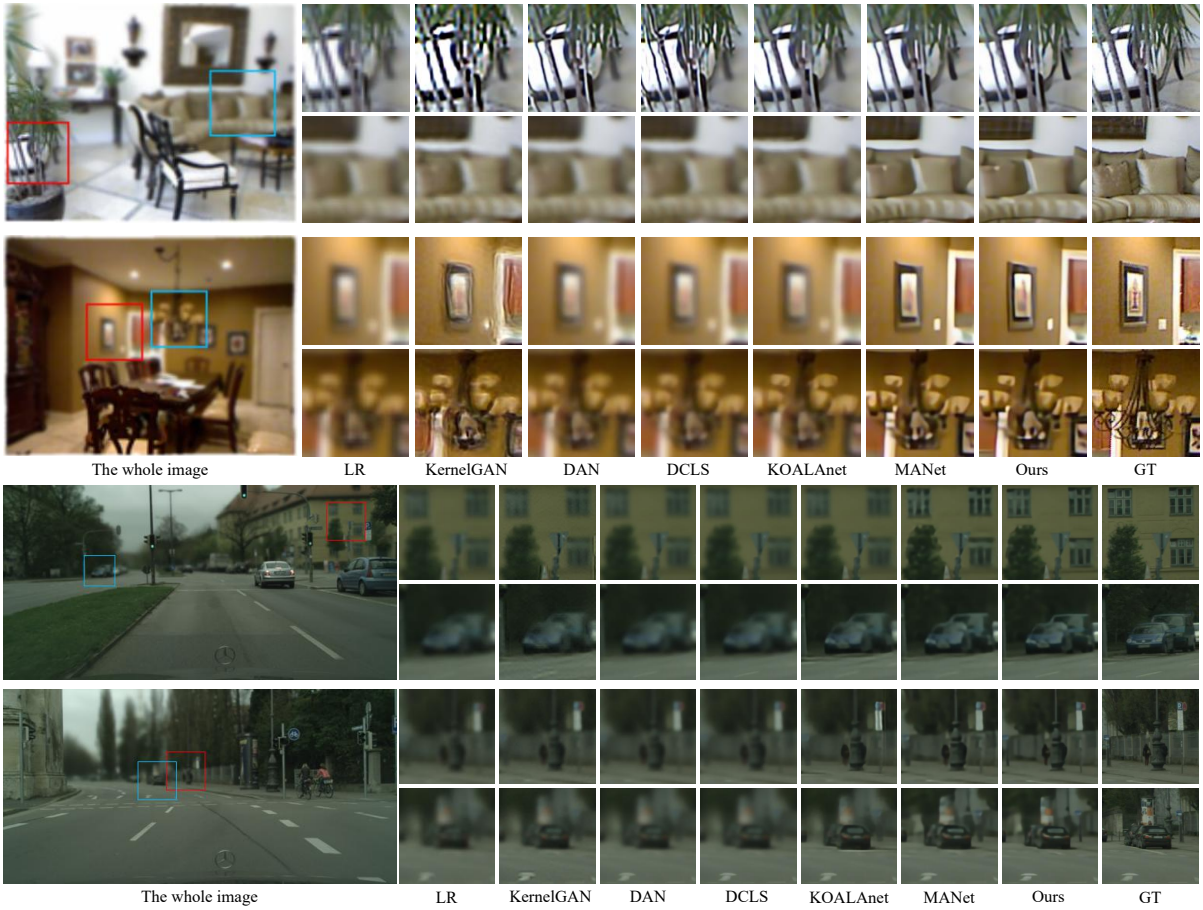


Figure 5. Qualitative comparisons between different SR methods on spatially variant blur (out-of-focus). The first two pictures are from NYUv2-BSR and the last two are from Cityscapes-BSR. (Please zoom in for better view.)

Method	PSNR ↑	SSIM ↑	mIoU ↑
Ours w/o GIA	23.21	0.8312	32.15
Ours w/ F	23.42	0.8314	33.04
Ours w/ F+C	<u>24.24</u>	<u>0.8336</u>	**36.25**
Ours w/ F+C+S (GIA)	**24.52**	**0.8340**	<u>35.61</u>

Table 5. Effectiveness of GIA. Note that these are the intermediate results, and PSNR/SSIM refer to the blur maps rather than the final SR mages. mIoU evaluates the effect of semantic estimation.

models: one on the proposed NYUv2-BSR dataset, and the other on the space-invariant blurred images generated from the NYUv2 dataset. The comparison results are shown in the Tab. 4. Apparently, training on spatially variant blurred images can increase PSNR and SSIM of SR images dramatically by 10.62 dB ↑ and 0.2266 ↑, respectively. This indicates that maintaining consistency in image blur types during the training and testing phases is crucial.

Effectiveness of GIA Module. We take out the components, *i.e.*, flow-based upsampling (F), channel interaction (C) and spatial interaction (S), of GIA to verify validity. We record the best PSNR and mIoU models individually. As shown in Tab. 5, using only flow-based upsampling improves the results slightly, and when combined with channel interaction, the performance can be significantly enhanced. Furthermore, utilizing all three components, *i.e.*, the complete GIA module, can yield even greater improvements.

Effectiveness of Semantic Information in SR. In order to illustrate that the semantic information is conducive to SR, we ablate it and only input blur maps into RRDB-SFT. It is worth noting that we use the ground truth blur and semantic maps here. As shown in Tab. 6, adding semantic maps improves the PSNR (+0.34 dB ↑) and SSIM (+0.0022 ↑) of the final SR results. We hold the opinion that semantic information may allow the network to take advantage of textural features of related objects it has learned about, and sharp semantic edges may also be helpful in SR.

Effectiveness of Multi-task Learning (MTL). To demonstrate the effectiveness of MTL, firstly, we make separate predictions for blur and semantic maps and compared them

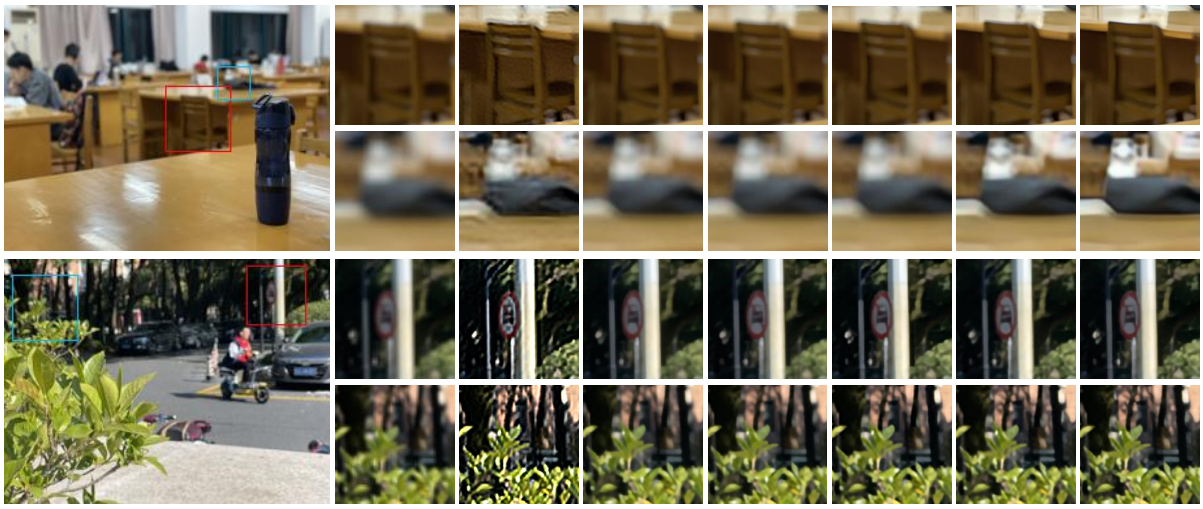


Figure 6. Visual results on real-world images for scale factor 4. The first picture of the indoor scene uses the model trained on NYUv2-BSR, and the second picture of the outdoor scene uses the model trained on Cityscapes-BSR. (Please zoom in for better view.)

Method	PSNR ↑	SSIM ↑
RRDB-SFT w/o semseg	33.41	0.9278
RRDB-SFT w/ semseg	**33.75**	**0.9300**

Table 6. Importance of using semantic information in SR. The ground-truth blur and semantic maps are used in this experiment.

Method	Intermediate Results		SR Results
	PSNR/SSIM ↑	mIoU ↑	PSNR/SSIM ↑
Single Task	**24.58/0.8393**	33.95	30.75/0.9134
CMOS (Ours)	24.52/0.8340	**35.61**	**32.03/0.9154**

Table 7. Effectiveness of MTL. PSNR/SSIM and mIoU of the intermediate results refer to the blur maps and the semantic maps.

with the CMOS results. Secondly, We compare the SR results achieved by solely utilizing the estimated blur maps versus employing both the estimated blur and semantic maps. As shown in Tab. 7, joint estimation improves the results of semantic segmentation (mIoU +1.66↑), albeit with a slight decrease in the performance of blur estimation. But in general, MTL can improve the PSNR/SSIM of the final SR results by +1.28↑/+0.002↑, which proves that semantics is useful to the overall SR process.

Importance of the Auxiliary Supervision. We ablate the auxiliary supervision in CMOS to see if it is necessary for our framework. As shown in Tab. 8, without the auxiliary supervision in the multi-scale structure, although there is a slightly increase in SSIM, PSNR and mIoU droped by 0.38 ↓ and 0.24% ↓, respectively. Therefore, auxiliary supervision can improve the performance of CMOS on the whole.

Methods	PSNR ↑	SSIM ↑	mIoU ↑
CMOS w/o AS	24.14	**0.8347**	35.37
CMOS w/ AS	**24.52**	0.8340	**35.61**

Table 8. Importance of the auxiliary supervision (AS) in CMOS.

5.4. Experiments on Real-Wrold SR

As there is no ground-truth for real images, we only compare visual results of different methods. As shown in Fig. 6, similar to the results on our datasets, KernelGAN still generate ringing artifacts, especially in the ourdoor scene. DAN, DCLS and KOALAnet all produce blurry results, while MANet performs slightly better. In comparision, CMOS can produce realistic and natural textures, and the results are the clearest.

6. Conclusion

In this paper, we introduce out-of-focus blur to SR and propose two new datasets: NYUv2-BSR and Cityscapes-BSR. Besides, we further propose a novel model CMOS to estimate the blur and semantic maps simultaneously. By incorporating semantics, we can restore finer SR results. GIA modules is used to achieve effective feature interaction in both spatial and channel dimensions. Extensive experiments on proposed datasets and real-world images demonstrate that our model can achieve SOTA performance in blind SR when integrated with existing non-blind models.

Acknowledgments: This work is supported by the National Natural Science Foundation of China (61836015).

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 6

[2] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 1

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3, 6

[4] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997. 2

[5] Zhenxuan Fang, Weisheng Dong, Xin Li, Jinjian Wu, Leida Li, and Guangming Shi. Uncertainty learning in kernel estimation for multi-stage blind image super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 144–161. Springer, 2022. 1

[6] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019. 1, 2, 3

[7] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368, 2022. 5

[8] Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhao Zhang, Chengjie Wang, Guannan Jiang, Wei Zhang, Ran Yi, Lizhuang Ma, et al. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4361–4370, 2022. 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[10] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33:5632–5643, 2020. 6

[11] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1428–1437, 2020. 3

[12] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 466–467, 2020. 1

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 3

[15] Soo Ye Kim, Hyeonjun Sim, and Munchurl Kim. Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10611–10620, 2021. 1, 3, 6

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Martin Kraus and Magnus Strengert. Depth-of-field rendering by pyramidal image processing. In *Computer graphics forum*, volume 26, pages 645–654. Wiley Online Library, 2007. 3

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3

[19] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12222–12230, 2019. 2, 3

[20] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 3

[21] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *European Conference on Computer Vision*, pages 775–793. Springer, 2020. 2, 4

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 1

[23] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4096–4105, 2021. 1, 2, 3, 5, 6

[24] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. Flow-based kernel prior with application to

blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10601–10610, 2021. 1

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 3

[26] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013. 1

[27] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015. 5

[28] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17642–17652, 2022. 1, 2, 6

[29] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020. 1

[30] Michael Potmesil and Indranil Chakravarty. A lens and aperture camera model for synthetic image generation. *ACM SIGGRAPH Computer Graphics*, 15(3):297–305, 1981. 3

[31] Siddhant Sahu, Manoj Kumar Lenka, and Pankaj Kumar Sa. Blind deblurring using deep learning: A survey. *arXiv preprint arXiv:1907.10128*, 2019. 5

[32] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018. 1, 3

[33] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 3

[34] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020. 3

[35] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 1

[36] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *European Conference on Computer Vision*, pages 527–543. Springer, 2020. 2, 4

[37] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 5

[38] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data supplementary material. 1, 2

[39] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 3

[40] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[42] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 2

[43] Ruikang Xu, Zeyu Xiao, Jie Huang, Yueyi Zhang, and Zhiwei Xiong. Edpn: Enhanced deep pyramid network for blurry image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 414–423, 2021. 1

[44] Yu-Syuan Xu, Shou-Yao Roy Tseng, Yu Tseng, Hsien-Kai Kuo, and Yi-Min Tsai. Unified dynamic convolutional network for super-resolution with variational degradations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12496–12505, 2020. 3

[45] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 2

[46] Jiangning Zhang, Xiangtai Li, Yabiao Wang, Chengjie Wang, Yibo Yang, Yong Liu, and Dacheng Tao. Eatformer: improving vision transformer inspired by evolutionary algorithm. *arXiv preprint arXiv:2206.09325*, 2022. 5

[47] Jiangning Zhang, Chao Xu, Jian Li, Wenzhou Chen, Yabiao Wang, Ying Tai, Shuo Chen, Chengjie Wang, Feiyue Huang, and Yong Liu. Analogous to evolutionary algorithm: Designing a unified sequence model. *Advances in Neural Information Processing Systems*, 34:26674–26688, 2021. 5

[48] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 2

[49] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3262–3271, 2018. 2, 3

[50] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention. In *European Conference on Computer Vision*, pages 56–72. Springer, 2020. 2