




Multi-Stage Degradation and Content Embedding Fusion for Blind Super-Resolution

Haiyang Zhang¹ , Mengyu Jiang²  and Liang Liu³ 

School of Computer Science, Beijing University of Posts and Telecommunications

Abstract

To achieve promising results on blind image super-resolution (SR), some Unsupervised Degradation Prediction (UDP) methods narrow the domain gap between the degradation embedding space and the SR feature space by fusing the degradation embedding with the additional content embedding before multi-stage SR. However, fusing these two embeddings before multi-stage SR is inflexible, due to the variation of the domain gap at each SR stage. To address this issue, we propose the Multi-Stage Degradation and Content Embedding Fusion (MDCF), which adaptively fuses the degradation embedding with the content embedding at each SR stage rather than before multi-stage SR. Based on the MDCF, we introduce a novel UDP method, called MDCFnet, which contains an additional Dual-Path Local and Global encoder (DPLG) to extract the degradation embedding and the content embedding separately. Specially, DPLG diversifies receptive fields to enrich the degradation embedding and combines local and global features to optimize the content embedding. Extensive experiments on real images and several benchmarks demonstrate that the proposed MDCFnet can outperform the existing UDP methods and achieve competitive performance on PSNR and SSIM even compared with the state-of-the-art SKP methods.

CCS Concepts

• Computing methodologies → Reconstruction;

1. Introduction

As a fundamental low-level vision problem, single image super-resolution (SISR) aims at restoring the high-resolution (HR) image from the low-resolution (LR) input with a pre-defined degradation process (e.g., bicubic). However, as revealed in [GLZD19], SISR methods are impractical due to the mismatch between the pre-defined degradation and the real one. Therefore, more attention is paid to SR with unknown degradations, also known as blind image super-resolution (BSR). Typically, it is achieved by two sequential steps: (1) estimate the degradation from the LR input and (2) fuse the degradation into the SR feature for image restoration. According to the form of degradation estimation, existing BSR methods can be divided into two groups: Supervised Kernel Prediction (SKP) and Unsupervised Degradation Prediction (UDP).

Most existing SKP methods utilize the classic degradation model [BKSI19, GLZD19] to represent the degradation. In general, these methods explicitly estimate blur kernels from LR images and leverage the kernel stretching strategy to fuse degradation information into non-blind SR networks. The main challenges they encounter are the ambiguity produced by downsampling (e.g., bicubic) [LHY*22] and the low robustness to incorrect kernel estimation [GLZD19]. To address these issues, [GLZD19, HLW*20, FWX*22] design alternating optimization algorithms, and [BKSI19, LZG*21, TJW*21, LHY*22] optimize the form of

kernel representations. Nevertheless, SKP methods are still unrealistic due to the unavailable real-world blur kernels.

Different from SKP methods, UDP methods learning abstract degradation embeddings are more suitable for real applications without supervision from the ground-truth kernel label. As the most representative method, DASR [WWD*21] utilizes contrastive learning to learn degradation embeddings. The main idea of DASR is to distinguish various degradations in the embedding space rather than explicit estimation in the pixel space. However, the inconsistency between the degradation embedding and the SR feature is still challenging. To address this issue, CDSR [ZLL*22] observes that content information can reduce the inconsistency. Specially, CDSR fuses the degradation embedding with the content embedding before SR and proposes a Domain Query Attention based module (DQA) to adaptively reduce the inconsistency. Like most BSR methods, CDSR reconstructs the SR features in a multi-stage manner. Nevertheless, CDSR don't consider that the SR features varying at each SR stage cause the variation of the inconsistency at each SR stage. Thus, the degradation embedding fused before multi-stage SR is inflexible and limits the results.

In this work, we polish a UDP method, called MDCFnet, through two aspects: (1) To adapt to the variation of the domain gap between the degradation embedding space and the SR feature space, the Multi-Stage Degradation and Content Embedding Fusion (MDCF)

is proposed, which fuses the degradation embedding with the content embedding individually at each SR stage, as shown in Figure 1. Moreover, MDCF adaptively calculates the weights of the degradation embedding and the content embedding by a self-attention-like mechanism [XZS*20, XCY*21]. (2) In order to generate the degradation embedding and the content embedding separately and effectively before SR, a Dual-Path Local and Global encoder (DPLG) is proposed, which contain a Multiple Receptive Fields Depthwise Separable Convolution (MConv) and a Local and Global Combination (LGC) module, as shown in Figure 1. Specifically, in order to enrich the degradation embedding, the MConv is applied in the pixel-wise branch to diversify its receptive field without increasing it. In order to optimize the content embedding, the LGC combines the local and global features in the patch-wise branch.

The main contributions of this paper are as follows:

- We propose a novel UDP method, called MDCFnet, based on the Multi-Stage Degradation and Content Embedding Fusion (MDCF), which adaptively fuses the degradation embedding with the content embedding at each SR stage rather than before multi-stage SR.
- We present a Dual-Path Local and Global encoder (DPLG) to extract degradation and content information separately and effectively. Specially, DPLG diversifies receptive fields to enrich the degradation embedding and combines local and global features to optimize the content embedding.
- Extensive experiments on real images and several benchmarks demonstrate that the proposed MDCFnet can outperform the existing UDP methods and achieve competitive performance on PSNR and SSIM even compared with the state-of-the-art SKP methods.

2. Related Work

2.1. Non-blind Super-Resolution

Since the arising of SRCNN [DLHT15] which learns the mapping from LR to HR image with DNNs, plenty of DNN-based elaborate architecture designs [SCH*16, KLL16a, KLL16b, HSU18, ZTK*18, ZLL*18, LCS*21] and training strategies [LTH*17, BYT18, WYW*18] are proposed due to their remarkable performance. Nevertheless, these methods assume that degradation is known and fixed (e.g., bicubic). When the assumed degradation deviates from the real one, SR reconstruction inevitably leads to inferior results. To be more flexible, some methods [ZZZ18, ZZZ19, XTT*20, SCI18] handle multiple degradations by giving corresponding priorities (i.e., blur kernels). However, these priorities are unavailable in real world.

2.2. Blind Super-Resolution

Supervised Kernel Prediction. A feasible way to deal with unknown degradations is that explicitly estimate blur kernels from LR images and fuse them into SR feature. S2K [TJW*21] estimates blur kernels in frequency domain and demonstrates that feature representation in frequency domain is more conducive for blur kernel estimation than in spatial domain. IKC [GLZD19] proposes an iterative kernel correction method to estimate accurate

blur kernels. DAN [HLW*20] adopts an alternating optimization algorithm to estimate the blur kernel and restore the SR image in a single network. MANet [LSZ*21] uses a moderate receptive field and exploits channel interdependence to estimate kernels. KXNet [FWX*22] integrates the learning process with the inherent physical mechanism to generate blur kernels with clear physical patterns. DCLS [LHY*22] introduces dynamic deep linear kernel to provide more equivalent choices of possible optimal solutions for kernel and fuses blur kernel into LR image in the feature domain to obtain clean feature. Nevertheless, these SKP methods are sensitive to kernel estimation errors and can't deal with real-world degradations deviating from the training degradation distribution.

Unsupervised Degradation Prediction. Instead of requiring the supervision from the ground-truth kernel label, UDP methods learn abstract degradation embeddings to distinguish various degradations in the embedding space rather than explicit estimation in the pixel space. DASR [WWD*21] is the first to leverage contrastive learning [CH21, HFW*20, CFGH20, CKNH20, DSRB14, HCL06] to learn degradation embeddings based on the assumption that the degradation is the same in each image and varies for different images. CDSR [ZLL*22] demonstrates that content information can serve as a cue to narrow the domain gap between the degradation embedding space and the SR feature space. Although existing UDP methods develop a more suitable manner for real-world applications, the strategy of fusing the degradation embedding with the content embedding is still inflexible.

3. Method

We now formally introduce the MDCFnet, which consists of two newly-established modules: the Multi-Stage Degradation and Content Embedding Fusion (MDCF) and the Dual-Path Local and Global encoder (DPLG). Besides, the SR network is based on the DQA module [ZLL*22] and the RRDB module [WYW*18]. As shown in Figure 1.

3.1. Multi-Stage Degradation and Content Embedding Fusion

The Multi-Stage Degradation and Content Embedding Fusion (MDCF) is proposed to adaptively fuse the degradation embedding $E_d \in \mathbb{R}^{1 \times L}$ with the content embedding $E_c \in \mathbb{R}^{1 \times L}$ at each SR stage. The MDCF can be divided into the Multi-Stage Fusion (MSF) strategy and the Adaptive Embedding Fusion (AEF) module.

Multi-Stage Fusion Strategy. In order to handle the domain gap between the degradation embedding space and the SR feature space varying at each SR stage, the MSF is proposed to fuse the degradation embedding with the content embedding individually at each SR stage. Specifically, as shown in Figure 1, each SR stage contains three sequential operations: (1) adaptively fuse E_d with E_c to generate the fused degradation embedding E_f by the AEF; (2) fuse E_f into the SR feature by DQA [ZLL*22]; (3) further restore image by RRDB [WYW*18]. It should be mentioned that the previous work, CDSR [ZLL*22], fuses the degradation embedding E_d with the content embedding E_c before multi-stage SR and then feeds the fused degradation embedding E_f to all SR stages. Due to each SR stage in CDSR only containing operation (2) and (3), it can't adapt

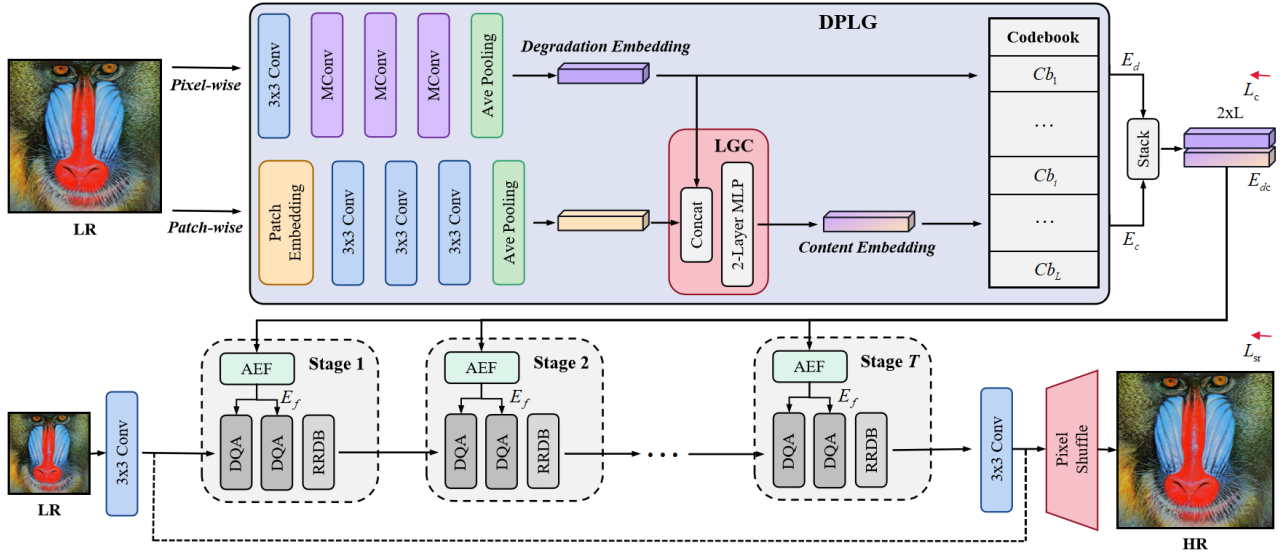


Figure 1: The overall architecture of MDCFnet. Top: Dual-Path Local and Global encoder (DPLG). DPLG includes degradation embedding generation and content embedding generation. Bottom: SR network composed of proposed Multi-Stage Degradation and Content Embedding Fusion (MDCF). The output degradation embedding E_d and content embedding E_c are adaptively fused at each SR stage.

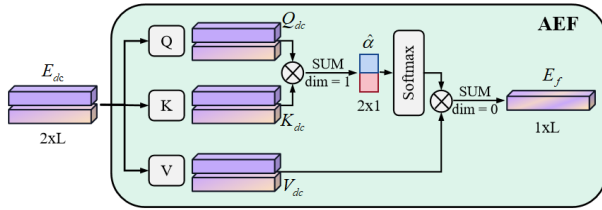


Figure 2: The architecture of Adaptive Embedding Fusion (AEF).

to the variation of the domain gap between the degradation embedding space and the SR feature space.

Adaptive Embedding Fusion Module. The AEF is designed to adaptively calculate the weights of degradation and content embeddings. As shown in Figure 2, like self-attention mechanism, the AEF first generates a query $Q_{dc} = [Q_d, Q_c] \in \mathbb{R}^{2 \times L}$, a key $K_{dc} = [K_d, K_c] \in \mathbb{R}^{2 \times L}$ and a value $V_{dc} = [V_d, V_c] \in \mathbb{R}^{2 \times L}$ from the stacked degradation and content embedding $E_{dc} = [E_d, E_c] \in \mathbb{R}^{2 \times L}$, which can be formulated as follows:

$$\begin{aligned} Q_{dc} &= f_q(E_{dc}), \\ K_{dc} &= f_k(E_{dc}), \\ V_{dc} &= f_v(E_{dc}), \end{aligned} \quad (1)$$

where $f_q(\cdot)$, $f_k(\cdot)$ and $f_v(\cdot)$ represent a simple single-layer MLP (Multi-Layer Perceptron) to project the input embedding to another embedding space. An important step for AEF to work is to adaptively calculate the weights of degradation and content embeddings. The main idea of AEF is to learn a mapping from the embedding to its importance. The feed-forward networks (i.e., $f_q(\cdot)$ and $f_k(\cdot)$) can be considered as the learnable parameters, and the similarity

between the query and the key can be regarded as the importance. Therefore, AEF implements the weight $\hat{\alpha} \in \mathbb{R}^{2 \times 1}$ as:

$$\hat{\alpha} = \left[\frac{\langle Q_d, K_d \rangle}{\sqrt{L}}, \frac{\langle Q_c, K_c \rangle}{\sqrt{L}} \right], \quad (2)$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product and L represents the length of embedding. The weight $\hat{\alpha}$ is then normalized to $\alpha \in \mathbb{R}^{2 \times 1}$ by softmax as:

$$\alpha = \text{softmax}(\hat{\alpha}, \text{dim} = 0). \quad (3)$$

Finally, V_d and V_c are combined to form an optimal fused degradation embedding $E_f \in \mathbb{R}^{1 \times C}$ via the weight α :

$$E_f = \text{SUM}(V_{dc} \otimes \alpha, \text{dim} = 0), \quad (4)$$

where \otimes denotes the element-wise multiplication with broadcast.

3.2. Analysis on Encoder

We first investigate the key point on: what kind of encoder is really needed for the generation of degradation and content embeddings? As shown in Table 1, we conduct experiments based on the encoder LPE [ZLL*22] to further investigate the effect of the patch-wise branch and the pixel-wise branch.

In Table 1, LPE_L , which only extracts local degradation features by the pixel-wise branch, achieves the highest degradation classification accuracy [ZLL*22]. Although there exists the inconsistency between the degradation embedding space and the SR feature space, it can be solved by adding content information. Thus, the best way to generate degradation embeddings is to learn local features. Due to the extraction of content information, LPE and LPE_P achieve a higher SR performance than LPE_L by reducing the inconsistency. Moreover, LPE achieves a higher degradation classification accuracy and SR performance than LPE_P by extracting

Table 1: Degradation classification accuracy of embedding (Acc.) [ZLL*22] and PSNR of $\times 2$ SR. We conduct experiments on three different encoders. LPE (used in CDSR): containing both pixel-wise and patch-wise branches. LPE_L : only containing pixel-wise branch. LPE_P : only containing patch-wise branch.

Encoder	Acc.	Set5	Set14	B100	Urban100
LPE	55.70%	37.48	33.23	31.92	31.12
LPE_L	64.20%	37.36	33.12	31.79	31.00
LPE_P	17.80%	37.42	33.17	31.85	31.06

additional local degradation features. Therefore, the combination of local and global features is more suitable for the generation of content embeddings than global features alone, due to a better capability of degradation classification.

3.3. Dual-Path Local and Global Encoder

As shown in Figure 1, we introduce the Dual-Path Local and Global encoder (DPLG), which contains a pixel-wise branch and a patch-wise branch to produce degradation embeddings and content embeddings, respectively. Moreover, a codebook [ZLL*22] is employed to constrict the basis of embedding space.

Degradation embedding generation. As shown in Section 3.2, the best way to generate degradation embeddings is to learn local features. As claimed in [LSZ*21], a moderate receptive field (22×22) on the LR image input keeps the locality of degradation. To enrich degradation embeddings, a Multiple Receptive Fields Depthwise Separable Convolution (MConv) is proposed to diversify the receptive field without increasing it. As shown in Figure 3, there exist three parallel groups of a pointwise convolution (PWConv) and a depthwise convolution (DWConv) with different kernel sizes (3×3 , 5×5 and 7×7). Then the concatenated features are fused by two sequential pointwise convolutions. Specifically, the depthwise separable strategy is applied to reduce the computation and parameters. The parallel depthwise convolutions diversify the receptive field of pixel-wise branch, which is ranged from 9×9 to 22×22 , as shown in Figure 1.

Content embedding generation. As mentioned in Section 3.2, the best way to generate content embeddings is to learn the combination of local and global features. We propose a Local and Global Combination module (LGC). As shown in Figure 1, the patch-wise branch first extracts the global features through a patch embedding layer and three standard convolution layers. Then, LGC concatenates the global features with the local features extracted by the pixel-wise branch and merges them into a content embedding E_c by a MLP with two layers. The content embedding E_c generated by LGC not only contains global content information but also local degradation information to enhance the capability of degradation classification.

3.4. Unsupervised Degradation Representation Learning

To conduct degradation representation learning in an unsupervised way, followed by DASR [WWD*21], we apply MoCov2

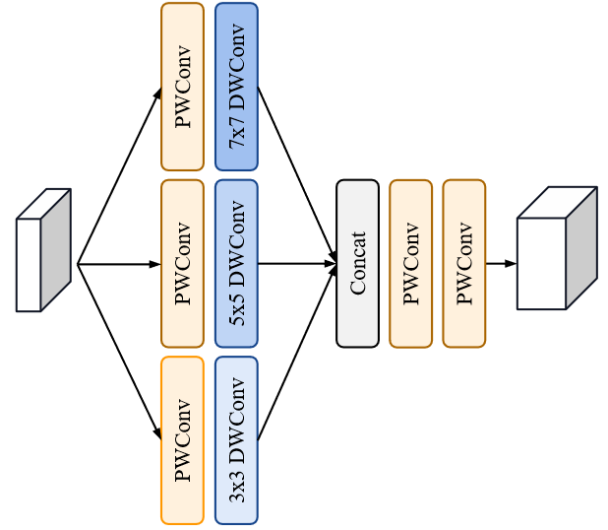


Figure 3: The architecture of Multiple Receptive Fields Depthwise Separable Convolution (MConv).

[CFGH20] to conduct the contrastive learning based on the assumption that the degradation is the same in an image but can vary for different images. In detail, given B HR images, we first randomly crop two patches from each HR image and degrade them with random blur kernels. It is worth noting that the two HR patches cropped from the same image are degraded with the same blur kernel. Then, these $2B$ LR patches are encoded into $\{d_i^1, c_i^1, d_i^2, c_i^2 \in \mathbb{R}^{256}\}$ using DPLG, where d_i^j is the degradation embedding E_d of the j^{th} patch from the i^{th} LR image and c_i^j is the content embedding E_c of the j^{th} patch from the i^{th} LR image. For the i^{th} LR image, we refer to p_i^1 and p_i^2 as query and positive samples, where $p_i^j = d_i^j + c_i^j$. The contrastive loss L_c is defined as:

$$L_c = \sum_{i=1}^B -\log \frac{\exp(E(p_i^1) \cdot E(p_i^2)/\tau)}{\sum_{j=1}^{N_{\text{queue}}} \exp(E(p_i^1) \cdot E(p_{\text{queue}}^j)/\tau)}, \quad (5)$$

where $E(\cdot)$ is the projection head [CKNH20], N_{queue} denotes the number of negative samples in the queue, p_{queue}^j represents the j^{th} negative sample and τ is a temperature hyper-parameter. Furthermore, the total loss function L is defined as $L = L_c + L_{sr}$, where L_{sr} denotes the L_1 distance between SR result and the HR ground-truth.

4. Experiments

4.1. Datasets and Implementation Details

Following [FWX*22], we collect 800 HR images from DIV2K [AT17] and 2650 HR images from Flickr2K [TAVG*17] as the training dataset. Furthermore, we synthesize the corresponding LR images via two different degradation kernel settings: (1) isotropic Gaussian kernels with noise-free and (2) anisotropic Gaussian kernels with noise. With the two settings, we can fully study the influence of degradations and the performance of the proposed method which is evaluated by PSNR and SSIM on only the luminance channel of the SR results (YCbCr space).

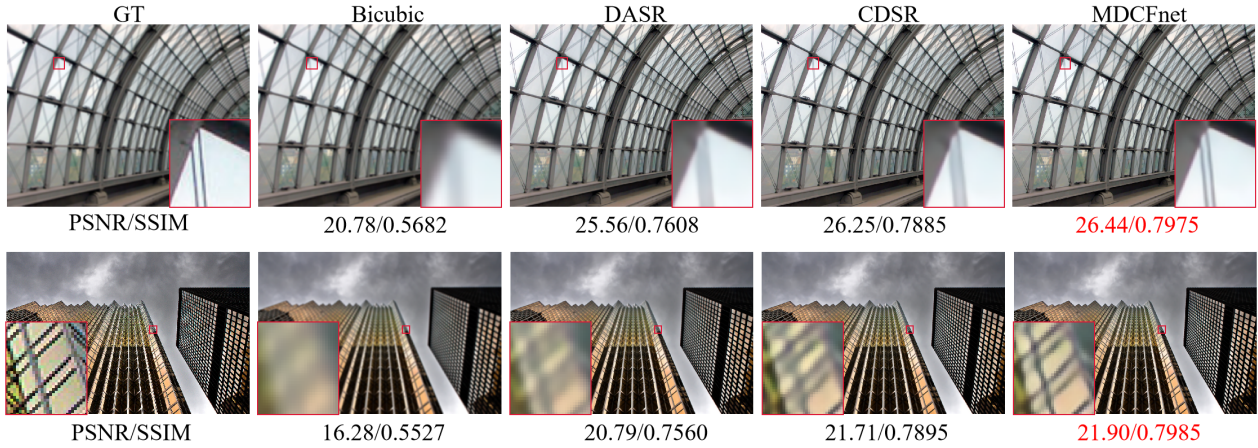


Figure 4: Visual results of img 002 and img 019 in Urban100 [HSA15], for scale factor 4 and kernel width 2.4. Best viewed in red color.

Table 2: Quantitative comparison on different test sets with isotropic Gaussian kernels. The best two results are marked in red and blue colors, respectively.

Method	Scale	Set5		Set14		B100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	x2	28.82	0.8577	26.02	0.7634	25.92	0.7310	23.14	0.7258
CARN		30.99	0.8779	28.10	0.7879	26.78	0.7286	25.27	0.7630
Bicubic + ZSSR		31.08	0.8786	28.35	0.7933	27.92	0.7632	25.25	0.7618
DASR		37.22	0.9513	32.72	0.8979	31.64	0.8829	30.29	0.9022
CDSR		37.48	0.9526	33.23	0.9044	31.92	0.8867	31.12	0.9116
MDCFnet(Ours)		37.55	0.9550	33.41	0.9088	32.00	0.8905	31.37	0.9160
Bicubic	x4	24.57	0.7108	22.79	0.6032	23.29	0.5786	20.35	0.5532
CARN		26.57	0.7420	24.62	0.6226	24.79	0.5963	22.17	0.5865
Bicubic + ZSSR		26.45	0.7279	24.78	0.6268	24.97	0.5989	22.11	0.5805
DASR		31.52	0.8805	28.00	0.7540	27.29	0.7140	25.12	0.7417
CDSR		31.83	0.8850	28.39	0.7667	27.49	0.7216	25.72	0.7641
MDCFnet(Ours)		31.99	0.8874	28.50	0.7686	27.52	0.7246	25.82	0.7693

Isotropic Gaussian Kernels with Noise-Free. The blur kernel size is set as 21×21 for all scales. During training, the kernel width is uniformly sampled from the ranges $[0.2, 2.0]$ and $[0.2, 4.0]$ for scale factors 2 and 4 respectively. For evaluation, we use *Gaussian8* [GLZD19] kernel setting to synthesize the testing dataset from four benchmarks: Set5 [BRGM12], Set14 [ZEP10], B100 [MFTM01] and Urban100 [HSA15].

Anisotropic Gaussian Kernels with Noise. The blur kernel size is set as 11×11 and 31×31 for scale factor 2 and 4 respectively. During training, the kernel width at each axis is uniformly distributed in $(0.6, 5)$ and randomly rotated by an angle uniformly distributed in $[-\pi, \pi]$. Moreover, we apply uniform multiplicative noise (up to 25% of each pixel value of the kernel) and normalize it to sum to one. During testing, we evaluate our method by then benchmark DIV2KRC proposed by [BKSI19].

Implementation Details. The size of LR patch is set to 48×48 for all scales ($\times 2$ and $\times 4$). Thus the size of HR patch cropped from HR image is 96 and 192, respectively. The batch size B is set to 32. The

MDCFnet is trained end-to-end. The SR network employs $T = 10$ SR stages. Each stage contains a AEF, two DQA and a RRDB. The length of codebook is set to 1024. The channel number of embedding L is set to 256. As for MoCov2, the τ and N_{queue} in Equation 5 is set to 0.07 and 8192, respectively. The Adam optimizer with the momentum of $\beta_1 = 0.9$, $\beta_2 = 0.999$ is adopted to train out network with the learning rate being initially set to $1e - 4$. The learning rate will decay by half after every 125 epochs by the multi-step decreasing strategy. The training process takes 600 epochs.

4.2. Comparison with State-of-the-arts

Evaluation of isotropic Gaussian kernels with noise-free. Following [GLZD19], we evaluate our method on datasets synthesized by *Gaussian8* kernels. We compare our method with state-of-the-art UDP-based blind SR approaches: DASR [WWD*21] and CDSR [ZLL*22]. We also conduct comparison with CARN [AKS18] and ZSSR [SCI18] (with bicubic kernel).

The quantitative results are shown in Table 2. It is obvious that

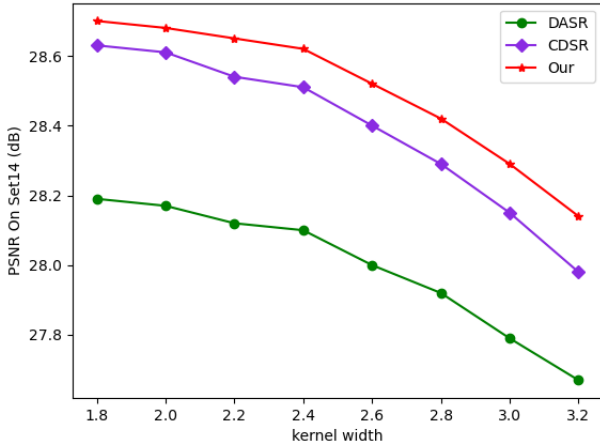


Figure 5: The PSNR curves on Set14 for scale factor 4.

our method leads to the best performance over all datasets. The SISR model suffers severe performance drop when the degradation deviates from the bicubic kernel. DASR learns unsupervised degradation representations by contrastive learning and largely improves the SR performance. Nevertheless, the domain gap between the degradation embedding space and the SR feature space limits the SR results. CDSR jointly extracts degradation and content features to adaptively reduce the inconsistency. Although the SR performance is improved in CDSR, the strategy of fusing degradation and content features is still inflexible. The qualitative results shown in Figure 4 illustrate that MDCFnet contains more useful textures and sharper edges. To show the variable trend with the change of kernel width σ , we also provide the various curves on Set14, as shown in Figure 5.

Evaluation of anisotropic Gaussian kernels with noise. Degradation with anisotropic Gaussian kernels and noise is more general and challenging. Similar to isotropic Gaussian kernels, we firstly compare our method with SOTA UDP-based blind SR approaches: DASR [WWD*21] and CDSR [ZLL*22]. We also compare MDCFnet with some SOTA SISR methods such as EDSR [LSK*17], RCAN [KLL16a] and DBPN [HSU18]. In addition, we combine a kernel estimation method KernelGAN [BKSI19] with a non-blind SR method ZSSR [SCI18] to solve blind SR. Furthermore, we compare our method with SOTA SKP-based blind SR methods: IKC [GLZD19], DANv1 [HLW*20], DANv2 [WWD*21], KOALAnet [KSK21] and DCLS [LHY*22]. For most methods, we use their official implementations and pre-trained models.

Table 3 shows the quantitative results on DIV2K [BKSI19]. It can be seen that the proposed MDCFnet outperforms the existing UDP models and achieves the competitive SR performance compared with the SOTA SKP models. ZSSR performs better based on the blur kernel estimated by KernelGAN. IKC iteratively corrects blur kernels. However, the two incompatible sub-networks limit the SR performance. To address this issue, DAN proposes an end-to-end SR model to improve the results. DCLS proposes a more effective form of kernel representation. Although these SKP methods achieve remarkable SR performance, the real world blur kernels

Table 3: Quantitative comparison on DIV2K. SK denotes whether correspond method belongs to SKP. The best two results are marked in red and blue colors, respectively.

Method	SK	DIV2K			
		$\times 2$		$\times 4$	
		PSNR	SSIM	PSNR	SSIM
IKC	✓	-	-	27.70	0.7668
DANv1	✓	32.56	0.8997	27.55	0.7582
DANv2	✓	32.58	0.9048	28.74	0.7893
KOALAnet	✓	31.89	0.8852	27.77	0.7637
DCLS	✓	32.75	0.9094	28.99	0.7946
Bicubic	✗	28.73	0.8040	25.33	0.6795
Bicubic + ZSSR	✗	29.10	0.8215	25.61	0.6911
EDSR	✗	29.17	0.8216	25.64	0.6928
RCAN	✗	29.20	0.8223	25.66	0.6936
DBPN	✗	29.13	0.8190	25.58	0.6910
DBPN + Correction	✗	30.38	0.8717	26.79	0.7426
KernelGAN + ZSSR	✗	30.36	0.8669	26.81	0.7316
AdaTarget	✗	-	-	28.42	0.7854
DASR	✗	32.24	0.8960	28.41	0.7813
CDSR	✗	32.68	0.9039	28.85	0.7901
MDCFnet(Ours)	✗	32.77	0.9119	28.98	0.7927

Table 4: Ablation study in the proposed main components on DIV2K for $\times 2$ SR and degradation classification accuracy (Acc.).

Model	MConv	LGC	MSF	AEF	Acc.		DIV2K	
					E_d	E_c	PSNR	SSIM
CDSR	✗	✗	✗	✗	54.00%	17.80%	29.36	0.8379
Model1	✓	✗	✗	✗	65.40%	18.20%	29.48	0.8402
Model2	✓	✓	✗	✗	62.80%	33.20%	29.53	0.8413
Model3	✓	✓	✓	✗	64.00%	34.00%	29.63	0.8436
MDCFnet	✓	✓	✓	✓	63.60%	33.80%	29.71	0.8451

are still unavailable. To address this issue, some UDP methods, such as DASR and CDSR, are proposed to train the degradation prediction network without supervision from the ground-truth kernel label. However, all of those methods are still inferior to our MDCFnet, which can even achieve the competitive results compared with the SOTA SKP methods.

4.3. Ablation Study

We conduct ablation studies on vital components of our method: Multiple Receptive Fields Depthwise Separable Convolution (MConv), Local and Global Combination module (LGC), Multi-Stage Fusion (MSF) and Adaptive Embedding Fusion (AEF). The quantitative results are shown in Table 4. All the experiments are conducted on DIV2K validation set blurred by *Gaussian8* with the scale factor of $\times 2$.

Effect of Multiple Receptive Fields Depthwise Separable Convolution. MConv diversifies the receptive field of the pixel-wise branch to enrich degradation information. Based on the enriched degradation information, we can enhance the degradation classi-

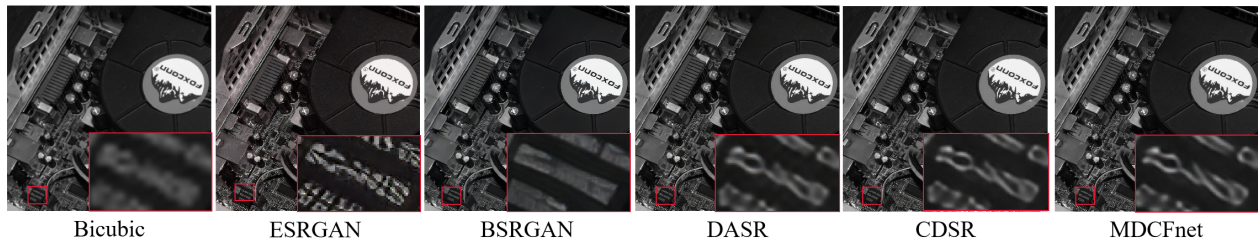


Figure 6: Visual comparison on RealsRSet with scale factor as 4.

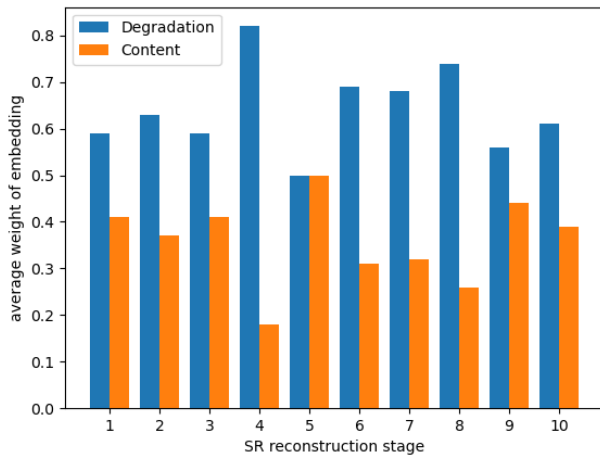


Figure 7: The average weights of degradation and content embeddings at each SR stage. The experiment is conducted for $\times 2$ SR on Set5 with Gaussian8.

fication capability of E_d to improve SR performance. To demonstrate its effect, we replace standard convolution layers in LPE_L by MConv. In addition, other components in CDSR are maintained. Shown in Table 4 "Model1", the degradation classification accuracy (Acc.) of E_d increases from 54.00% to 65.40%, and the SR performance shows an increase of 0.12dB on PSNR.

Effect of Local and Global Combination Module. In order to enhance the degradation classification capability of E_c , the LGC is proposed to combine content embeddings with degradation embeddings in the patch-wise branch. To demonstrate that improving the degradation classification capability of E_c benefits SR, we further add LGC to "Model1". Shown in Table 4 "Model2", the degradation classification accuracy (Acc.) of E_c increases from 18.20% to 33.20%, and the SR performance shows a moderate increase of 0.05dB on PSNR though LGC only contains a trainable 2-layer MLP.

Effect of Multi-Stage Fusion. In order to handle the domain gap between the degradation embedding space and the SR feature space varying at each SR stage, MSF is proposed to fuse the degradation embedding with the content embedding individually at each SR stage rather than before multi-stage SR. To study its effect, based on "Model2", we use a 2-layer MLP to fuse the degrada-

tion embedding with the content embedding at each SR stage. As demonstrated in Table 4 "Model3", the SR performance shows an increase of 0.10dB on PSNR.

Effect of Adaptive Embedding Fusion. AEF is designed to adaptively calculate the weights of degradation and content embeddings. Based on the AEF, we can fuse the degradation embedding with the content embedding flexibly. To demonstrate its effect, we replace the 2-layer MLP by the AEF. As shown in Table 4 MDCFnet, the SR performance shows an increase of 0.08dB on PSNR. Furthermore, we show the average weights of degradation and content embeddings at each SR stage in Figure 7. It indicates that the domain gap between the degradation embedding space and the SR feature space indeed varies at each SR stage, due to the variation of the weights of the content embedding at each SR stage.

4.4. Performance on Real Degradation

To further demonstrate the effectiveness of our method, we conduct experiments on the real images. UDP methods trained on anisotropic Gaussian kernels with noise is used for evaluation. Visualization results are shown in Figure 6. There are obvious artifacts in ESRGAN [WYW*18] and BSRGAN [ZLVGT21], though they may generate clearer textures. In the UDP methods, our MDCFnet can produce sharper edges and visual pleasing SR results.

5. Conclusion

In this paper, we propose the MDCF that adaptively fuses the degradation embedding with the content embedding at each SR stage rather than before multi-stage SR. Specifically, the AEF is proposed to adaptively calculate the weights of degradation and content embeddings by a self-attention-like mechanism. To separate degradation and content embeddings, the DPLG is applied. Specially, DPLG diversifies receptive fields to enrich the degradation embedding and combines local and global features to optimize the content embedding. Extensive experiments on real images and several benchmarks demonstrate that the proposed MDCFnet can outperform the state-of-the-art UDP methods on PSNR and SSIM and achieve visually favorable results. In the future, we will try to apply similar unsupervised manners in other low-level vision tasks, such as deblurring and denoising.

References

- [AKS18] AHN N., KANG B., SOHN K.-A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 252–268. 5
- [AT17] AGUSTSSON E., TIMOFTE R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2017), pp. 126–135. 4
- [BKS19] BELL-KLIGLER S., SHOCHER A., IRANI M.: Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems* 32 (2019). 1, 5, 6
- [BRGM12] BEVILACQUA M., ROUMY A., GUILLEMOT C., MOREL M.-L. A.: Low-complexity single-image super-resolution based on non-negative neighbor embedding. In *Proceedings of the British Machine Vision Conference* (2012), BMVA press, pp. 135.1–135.10. 5
- [BYT18] BULAT A., YANG J., TZIMIROPOULOS G.: To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 185–200. 2
- [CFGH20] CHEN X., FAN H., GIRSHICK R., HE K.: Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020). 2, 4
- [CH21] CHEN X., HE K.: Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15750–15758. 2
- [CKNH20] CHEN T., KORNBILTH S., NOROUZI M., HINTON G.: A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (2020), PMLR, pp. 1597–1607. 2, 4
- [DLHT15] DONG C., LOY C. C., HE K., TANG X.: Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* 38, 2 (2015), 295–307. 2
- [DSRB14] DOSOVITSKIY A., SPRINGENBERG J. T., RIEDMILLER M., BROX T.: Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems* 27 (2014). 2
- [FWX*22] FU J., WANG H., XIE Q., ZHAO Q., MENG D., XU Z.: Kxnet: A model-driven deep neural network for blind super-resolution. *arXiv preprint arXiv:2209.10305* (2022). 1, 2, 4
- [GLZD19] GU J., LU H., ZUO W., DONG C.: Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 1604–1613. 1, 2, 5, 6
- [HCL06] HADSELL R., CHOPRA S., LECUN Y.: Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (2006), vol. 2, IEEE, pp. 1735–1742. 2
- [HFV*20] HE K., FAN H., WU Y., XIE S., GIRSHICK R.: Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 9729–9738. 2
- [HLW*20] HUANG Y., LI S., WANG L., TAN T., ET AL.: Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems* 33 (2020), 5632–5643. 1, 2, 6
- [HSA15] HUANG J.-B., SINGH A., AHUJA N.: Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 5197–5206. 5
- [HSU18] HARIS M., SHAKHAROVICH G., UKITA N.: Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 1664–1673. 2, 6
- [KLL16a] KIM J., LEE J. K., LEE K. M.: Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1646–1654. 2, 6
- [KLL16b] KIM J., LEE J. K., LEE K. M.: Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1637–1645. 2
- [KSK21] KIM S. Y., SIM H., KIM M.: Koalnet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 10611–10620. 6
- [LCS*21] LIANG J., CAO J., SUN G., ZHANG K., VAN GOOL L., TIMOFTE R.: Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 1833–1844. 2
- [LHY*22] LUO Z., HUANG H., YU L., LI Y., FAN H., LIU S.: Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 17642–17652. 1, 2, 6
- [LSK*17] LIM B., SON S., KIM H., NAH S., MU LEE K.: Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2017), pp. 136–144. 6
- [LSZ*21] LIANG J., SUN G., ZHANG K., VAN GOOL L., TIMOFTE R.: Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 4096–4105. 2, 4
- [LTH*17] LEDIG C., THEIS L., HUSZÁR F., CABALLERO J., CUNNINGHAM A., ACOSTA A., AITKEN A., TEJANI A., TOTZ J., WANG Z., 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* (2017), pp. 4681–4690. 2
- [LZG*21] LIANG J., ZHANG K., GU S., GOOL L. V., TIMOFTE R.: Flow-based kernel prior with application to blind super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2021). 1
- [MFTM01] MARTIN D., FOWLKES C., TAL D., MALIK J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (2001), vol. 2, IEEE, pp. 416–423. 5
- [SCH*16] SHI W., CABALLERO J., HUSZÁR F., TOTZ J., AITKEN A. P., BISHOP R., RUECKERT D., WANG Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1874–1883. 2
- [SCI18] SHOCHER A., COHEN N., IRANI M.: “zero-shot” super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 3118–3126. 2, 5, 6
- [TAVG*17] TIMOFTE R., AGUSTSSON E., VAN GOOL L., YANG M.-H., ZHANG L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2017), pp. 114–125. 4
- [TJW*21] TAO G., JI X., WANG W., CHEN S., LIN C., CAO Y., LU T., LUO D., TAI Y.: Spectrum-to-kernel translation for accurate blind image super-resolution. *Advances in Neural Information Processing Systems* 34 (2021), 22643–22654. 1, 2
- [WWD*21] WANG L., WANG Y., DONG X., XU Q., YANG J., AN W., GUO Y.: Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10581–10590. 1, 2, 4, 5, 6

- [WYW*18] WANG X., YU K., WU S., GU J., LIU Y., DONG C., QIAO Y., CHANGE LOY C.: Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops* (2018), pp. 0–0. [2](#), [7](#)
- [XCY*21] XI D., CHEN Z., YAN P., ZHANG Y., ZHU Y., ZHUANG F., CHEN Y.: Modeling the sequential dependence among audience multi-step conversions with multi-task learning in targeted display advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 3745–3755. [2](#)
- [XTT*20] XU Y.-S., TSENG S.-Y. R., TSENG Y., KUO H.-K., TSAI Y.-M.: Unified dynamic convolutional network for super-resolution with variational degradations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12496–12505. [2](#)
- [XZS*20] XI D., ZHUANG F., SONG B., ZHU Y., CHEN S., HONG D., CHEN T., GU X., HE Q.: Neural hierarchical factorization machines for user’s event sequence analysis. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (2020), pp. 1893–1896. [2](#)
- [ZEP10] ZEYDE R., ELAD M., PROTTER M.: On single image scale-up using sparse-representations. In *International conference on curves and surfaces* (2010), Springer, pp. 711–730. [5](#)
- [ZLL*18] ZHANG Y., LI K., LI K., WANG L., ZHONG B., FU Y.: Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 286–301. [2](#)
- [ZLL*22] ZHOU Y., LIN C., LUO D., LIU Y., TAI Y., WANG C., CHEN M.: Joint learning content and degradation aware feature for blind super-resolution. In *Proceedings of the 30th ACM International Conference on Multimedia* (2022), pp. 2606–2616. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#)
- [ZLVGT21] ZHANG K., LIANG J., VAN GOOL L., TIMOFTE R.: Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 4791–4800. [7](#)
- [ZTK*18] ZHANG Y., TIAN Y., KONG Y., ZHONG B., FU Y.: Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 2472–2481. [2](#)
- [ZZZ18] ZHANG K., ZUO W., ZHANG L.: Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 3262–3271. [2](#)
- [ZZZ19] ZHANG K., ZUO W., ZHANG L.: Deep plug-and-play super-resolution for arbitrary blur kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 1671–1681. [2](#)