MSRA-SR: Image Super-resolution Transformer with Multi-scale Shared Representation Acquisition

Xiaoqiang Zhou [1,2], Huaibo Huang[2,3], Ran He[2,3*], Zilei Wang[1], Jie Hu[4], Tieniu Tan[2,5]
[1] University of Science and Technology of China
[2] CRIPAC & MAIS, Institute of Automation, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences [4] OPPO [5] Nanjing University
xq525@mail.ustc.edu.cn, huaibo.huang@cripac.ia.ac.cn, {rhe, tnt}@nlpr.ia.ac.cn

Abstract

Multi-scale feature extraction is crucial for many computer vision tasks, but it is rarely explored in Transformer-based image super-resolution (SR) methods. In this paper, we propose an image super-resolution Transformer with Multi-scale Shared Representation Acquisition (MSRA-SR). We incorporate the multi-scale feature acquisition into two basic Transformer modules, i.e., self-attention and feed-forward network. In particular, self-attention with cross-scale matching and convolution filters with different kernel sizes are designed to exploit the multi-scale features in images. Both global and multi-scale local features are explicitly extracted in the network. Moreover, we introduce a representation sharing mechanism to improve the efficiency of the multi-scale design. Analysis on the attention map correlation indicates the representation redundancy in self-attention, which motivates us to design a shared self-attention across different Transformer layers. The exhaustive element-wise similarity matching is computed only once and then shared by later layers. Besides, the multi-scale convolution in different branches can be equivalently transformed into a single convolution with reparameterization trick. Extensive experiments on lightweight, classical and real-world image SR tasks verify the effectiveness and efficiency of the proposed method.

1. Introduction

Image super-resolution (SR) aims to restore high-resolution images from the low-resolution input images. As a fundamental image restoration task in low-level vision, image super-resolution has a broad range of applications, including photo enhancement [40], surveillance monitoring [53], and medical imaging [21], etc. However, due to the intricate degradation process and the diverse natural image content, single image super-resolution is still a challenging task in low-level vision.

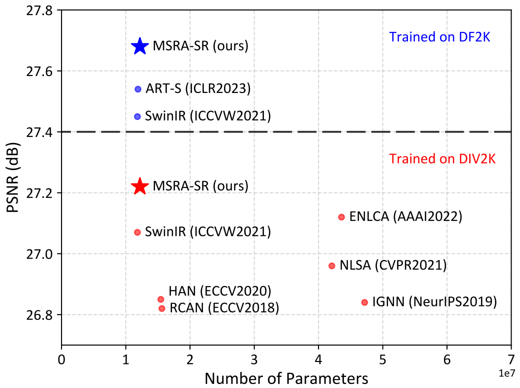


Figure 1: PSNR vs. #Params comparisons on Urban100 (×4).

In the last decade, the deep learning-based image SR methods have outperformed traditional image SR methods significantly. In the era of convolution neural network (CNN), Dong et al. pioneered a CNN-based image SR method called SRCNN [12]. Many great follow-up methods improve the CNN-based methods from the perspectives of network architecture [31, 78, 5, 74, 42, 47], degradation modeling [59, 75, 64] and training process [60, 62]. Recently, researchers adapt vision Transformers to the image super-resolution task and achieve appealing performance [35, 73]. However, the multi-scale feature acquisition, which is crucial for many computer vision tasks [32, 81], is rarely explored in Transformer-based image super-resolution methods. Besides, the inefficiency brought by exhaustive and repetitive calculation on self-attention hinders the application of Transformer-based image super-resolution methods.

In this paper, we introduce a multi-scale feature acquisition design and incorporate it into two basic Transformer modules, i.e., self-attention and feed-forward network. For the self-attention module, we improve the attention calcu-

[*] Ran He is the corresponding author.

lation from single-scale matching to cross-scale matching. The cross-scale attention (CSA) can not only compensate the limitation of self-attention in multi-scale dependency modeling, but also improve the receptive field of window self-attention efficiently. For the feed-forward network, the vanilla MLP layers can only enhance features with receptive field 1×1. In this paper, we propose a multi-scale depth-wise convolution layer (MS-DWC) and insert it into the feed-forward network. Convolutions with different kernel sizes can enhance the multi-scale local features. Apart from the multi-scale local representation extraction, we incorporate an efficient global attention into the Transformer block. In this way, both global and multi-scale local features are exploited explicitly.

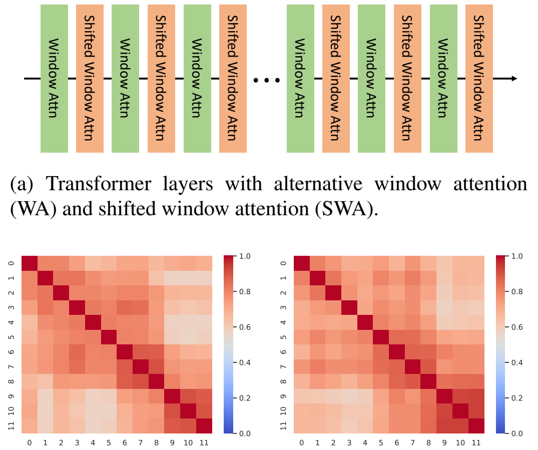
While multi-scale representation acquisition can improve the performance effectively, it also brings extra computational costs. Exhaustive and repetitive similarity matching in self-attention is computationally expensive, and the multi-branch depth-wise convolution in feed-forward network in is slow in inference.

To improve the efficiency of the multi-scale design and Transformer network, we present a representation sharing mechanism and also incorporate it into self-attention and feed-forward network. For the self-attention, we analyze the similarity map in self-attention across different layers, and find that there exists obvious representation redundancy. As shown in Fig. 2, the Pearson correlation coefficients of score maps in different layers are high. Based on such an investigation, we design a shared self-attention (SSA) for the image super-resolution Transformer. Specifically, the self-attention map is only computed once in the first layer and then shared by later layers. Exhaustive similarity matching among tokens is replaced by a lightweight transformation. In this way, expensive similarity matching is avoided and the computational cost is greatly reduced. For the multi-scale convolution in feed-forward network, while the multi-branch design is slow in inference, it can be transformed efficiently with the reparameterization trick. Convolutions with different kernel sizes can be reparameterized as the same kernel size in structure, and the weights can be aggregated into a single convolution with equivalence. In this way, the proposed representation sharing mechanism can boost the model's efficiency.

Overall, the contribution of this paper can be summarized as follows:

- We present a multi-scale shared image super-resolution Transformer, named MSRA-SR, to exploit the multi-scale feature acquisition and representation sharing mechanism in image super-resolution.

- The multi-scale representation acquisition is incorporated into the self-attention and feed-forward network. Both global and multi-scale local features are exploited explicitly with cross-scale attention and multi-scale depth-wise convolution.



(a) Transformer layers with alternative window attention (WA) and shifted window attention (SWA).

(b) Pearson correlation coefficient heatmap for WA.

(c) Pearson correlation coefficient heatmap for SWA.

Figure 2: Correlation of the self-attention maps in different layers is high. The redundancy analysis motivates to design a shared self-attention.

- The representation sharing mechanism is introduced to improve the model's efficiency. The shared self-attention is proposed to reduce the repetitive calculation on self-attention. And the efficiency of the multi-branch convolution is improved by structure reparameterization.

- Extensive experiments on lightweight, classical and real-world image SR tasks verify the effectiveness and efficiency of the proposed method.

2. Related Work

2.1. Single Image Super-Resolution

Over the past decade, deep learning has achieved tremendous success in many computer vision areas [22, 82, 46, 18, 19, 68, 69, 20, 80, 63, 16, 83]. As a fundamental task in low level vision, single image super-resolution (SR) methods in these years can be divided into two categories: CNN-based and Transformer-based methods [57, 36, 45, 51, 77, 34, 24]. For CNN-based methods, SRCNN [11] is the pioneering work that firstly introduces CNN into the image super-resolution task. After that, researchers propose neat and effective backbone networks, such as EDSR [38], RCAN [78], and ESRGAN [60]. Different kinds of visual attention (e.g., channel attention [78], spatial attention [10] and layer attention [52]) are explored. Introducing residual or dense connections in the backbone network are also verified to be effective [3, 42]. Some work focus on designing an efficient self-attention network and

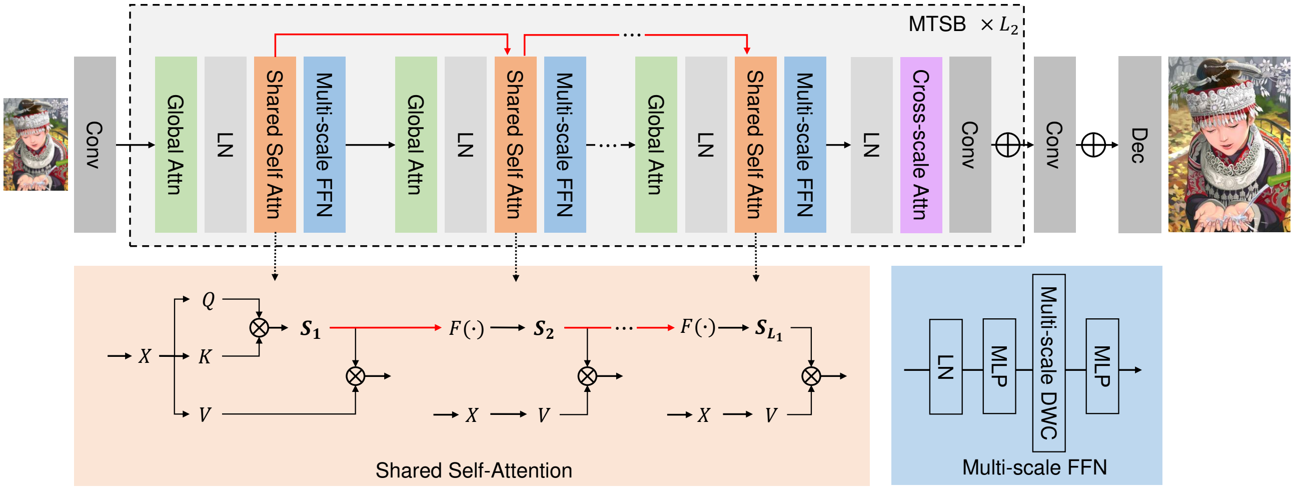


Figure 3: The architecture of the proposed MSRA-SR.

achieves promising performance by modeling long-range dependency [41, 50]. Recently, Transformer-based SR methods are proposed to adapt vision Transformer to low-level vision tasks [7, 61, 71, 73]. These methods improve the Transformer block design and integrate the Transformer blocks into the backbone network. Among them, SwinIR [35] integrates the Swin Transformer block [43] into the backbone and achieves great performance in image restoration. ART [73] incorporates dense and sparse attention modules to the image super-resolution Transformer.

2.2. Vision Transformer

The Transformer is firstly proposed by NLP researchers and achieves superior performance in many NLP tasks. Recently, there are more and more works incorporating the Transformer into computer vision tasks. ViT [14] is the pioneering work of Vision Transformers. After that, there are some works focusing on designing a general vision Transformer backbone for different kinds of vision tasks [76, 67, 26, 30, 25]. To adapt the Transformer network to the image inputs, hierarchical architectures [43, 58], efficient self-attention modules [43, 70] and diverse positional encodings [55, 8, 13] are proposed. While there are emerging some Transformer-based method for image restoration [7, 56]. However, there is just a few methods for Transformer-based image SR [35, 73]. And the multi-scale feature acquisition and representation sharing is rarely explored in image super-resolution Transformer.

2.3. Multi-scale Feature Learning

Multi-scale feature learning has always been a hot topic in computer vision. In the last decade, many CNN-based multi-scale feature learning methods are proposed, such as [39, 23]. These methods design a hierarchical architecture and use multi-scale convolutions to extract local features. Recently, with the prosperity of Transformer [55], more and more vision tasks incorporate the vision Transformer as backbone. The vanilla vision Transformer [15] is single-scale and is unsuitable for some high-resolution vision tasks, such as image segmentation [43]. To model the interactions among features of different scales, some image recognition methods improve the features in self-attention from single-scale to multi-scale and cross-scale, such as PVT [58], Focal Transformer [67], CrossViT [6], and MViT [17]. However, there is rare work on investigating the multi-scale feature learning in image super-resolution Transformer. P2T [65] and DualFormer [37] are designed for high-level vision, while the low-level vision task image SR focus on detail texture and structure instead. Differently, value features in our CSA are processed with pixel-unshuffle instead of downsample to preserve details. In this paper, we incorporate the multi-scale feature acquisition into self-attention and feed-forward network, and design a representation sharing mechanism to improve the efficiency of the multi-scale design.

3. Method

We propose an image super-resolution Transformer with multi-scale shared representation acquisition (MSRA-SR) for the single image super-resolution task. Fig. 3 shows the overall framework and shared self-attention. Fig. 4 presents the cross-scale attention and Fig. 5 illustrates the multi-scale shared depth-wise convolution.

3.1. Network Architecture

For an input low-resolution image X_{LR}, the image super-resolution network $H_{SR}(\cdot)$ aims to restore a high-resolution counterpart Y_{SR} from the input image, and the model parameters can be optimized with the groundtruth

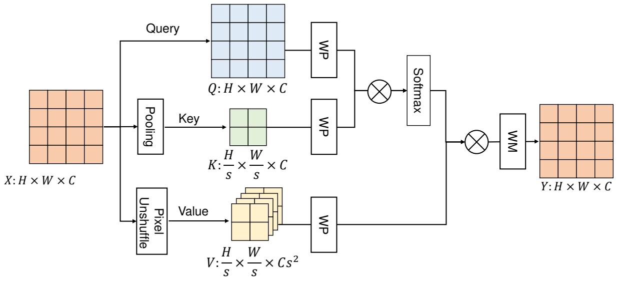


Figure 4: Illustration of the cross-scale attention (CSA). WP and WM denote window partition and window merging, respectively.

supervision Y_{HR}. Following the design of SwinIR [35], $H_{SR}(\cdot)$ consists of three sub-networks, i.e., shallow feature extraction network $H_{SF}(\cdot)$, deep feature extraction network $H_{DF}(\cdot)$ and image reconstruction network $H_{REC}(\cdot)$.

The shallow feature extraction network $H_{SF}(\cdot)$ is a 3×3 convolutional layer, and it transforms the input image $X_{LR} \in \mathbb{R}^{H \times W \times C_{in}}$ to a shllow feature $F_S \in \mathbb{R}^{H \times W \times C}$ as

$$F_S = H_{SF}(X_{LR}), \tag{1}$$

where H, W and C denote the feature height, width, and channels, respectively.

The deep feature extraction network $H_{DF}(\cdot)$ is built up with many multi-scale shared Transformer blocks (MSTB) and residual connections. As shown in Fig. 3, the MSTB extracts multi-scale features during the attention calculation and convolution process, including global attention, cross-scale attention and multi-scale convolution. The sharing mechanisms on self-attention and convolution are introduced to improve the efficiency of multi-scale design. Generally, the feature extraction process in the $H_{DF}(\cdot)$ can be described as

$$F_D = H_{DF}(F_S). \tag{2}$$

The shallow feature F_S contains the low-frequency image content, and the deep feature F_D is learnt to predict the missing high-frequency image details. They complement with each other and are aggregated in a residual manner. The image reconstruction network $H_{REC}(\cdot)$, which contains several non-strided and sub-pixel convolution layers, restores the super-resolution result Y_{SR} through

$$Y_{SR} = H_{REC}(F_S + F_D). \tag{3}$$

All of the learnable parameters in $H_{SR}(\cdot)$ are optimized with the L_1 loss between the prediction Y_{SR} and groundtruth Y_{HR}.

3.2. Multi-scale Representation Acquisition

As visual objects captured in images are often vary in scale, it is crucial to exploit multi-scale representation for achieving high-quality image super-resolution. In this paper, we present a novel multi-scale representation acquisition for the image super-resolution Transformer. In addition to the conventional single-scale local window attention, we propose the incorporation of an efficient global attention and multi-scale local feature extraction to enhance the acquisition of multi-scale representations.

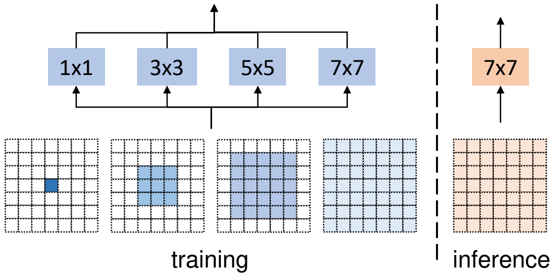


Figure 5: Illustration of the multi-scale depth-wise convolution (MS-DWC). The reparameterization trick helps to improve the inference efficiency.

Local and Global. The first way for multi-scale representation acquisition is complementing single-scale local features with global dependency modeling. Vanilla self-attention and many existing global self-attention methods are not suitable for image super-resolution task since the image resolution is usually high. Instead, we resort to channel attention for the sake of efficiency and effectiveness.

Channel attention can be considered as a means of global dependency modeling, as it aggregates all token information via global average pooling (GAP) in the spatial domain. The process of channel attention can be written as

$$Y = X \cdot \psi(W(GAP(X))), \tag{4}$$

where ψ denotes the sigmoid function, W is a two-layer MLP, and $\cdot$ means the channel-wise multiplication. More importantly, the computation complexity of channel attention is $\Omega(C^2)$, which is unrelated with the token number and therefore pretty efficient for the image super-resolution task. Channel attention is inserted into the Transformer block as shown in Fig. 3.

Single-scale to Multi-scale. The second way for multi-scale representation acquisition is extending single-scale local feature extraction to a multi-scale design. Since feedforward network and self-attention are two basic modules in the Transformer network, we introduce the multi-scale designment on these two modules simultaneously.

For the self-attention module, we propose a cross-scale attention (CSA) to cooperate with the window self-attention. Window self-attention is an efficient local attention, but it can only deal with single-scale local feature enhancement. As shown in Fig. 3 and Fig. 4, the cross-scale attention takes a feature $X \in \mathbb{R}^{H \times W \times C}$ as input, and projects it to the query, key, and value space respectively.

Specifically, Query $Q(X) \in \mathbb{R}^{H \times W \times C}$ keeps the same feature dimension shape with X. Key $K(X) \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times C}$ is a downscaled feature with scale factor s. Value $V(X) \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times Cs^2}$ is a informative downscaled feature which embeds the spatial feature into the channel dimension. The cross-scale attention (CSA) is calculated as

$$Y = \phi(\frac{Q(X) \cdot K(X)^T}{C}) \cdot V(X), \tag{5}$$

where ϕ is the softmax function, and C is the channel number. In practice, we use three scale factors $s = [1, 2, 4]$. The three scale cross-attention features are aggregated with a 1×1 convolution. We adopts the window partition in the projected features for computational efficiency. There are four advantages brought by CSA. Firstly, the similarity matching is performed in a cross-scale manner, enhancing the feature extraction in different scales. Secondly, the receptive field of downscaled key and value features in the window attention is bigger than that in vanilla window attention. Thirdly, there is no information loss in the downsampled value token as it embeds the local features into the channel dimension. Lastly, the computational cost of cross-scale attention is comparable with the window attention.

For the feed-forward network, as shown in Fig. 3 and Fig. 5, we insert depth-wise convolutions with different kernel sizes (MS-DWC) into the original MLP layers. There are three advantages brought by this design. Firstly, convolutions with different kernel sizes help to exploit the multi-scale local features, while the receptive field of MLP is only 1×1. Besides, the convolution can serves as a relative positional encoding while MLP cannot. Lastly, the computational cost brought by depth-wise convolutions is small compared with MLP.

3.3. Representation Sharing Mechanism

While the multi-scale representation acquisition can enhance the network's effectiveness on feature extraction, it also leads to a lower efficiency from two aspects. Firstly, the calculation on different feature scales brings extra computational costs. Besides, the multi-branch calculation in MS-DWC is unfriendly to GPU parallelization and increases the actual runtime. To improve the efficiency of the multi-scale design, we propose a representation sharing mechanism and apply it to self-attention and convolution.

Shared Self-attention. Self-attention is a core module in Transformer to model the element-wise long-range dependency. A Transformer network consists of many Transformer blocks and each Transformer block contains a self-attention module. By analyzing the similarity map in self-attention across different layers, we find that there exists obvious representation redundancy. We calculate the Pearson correlation coefficient between two score maps in different layers, and visualize it in a correlation map. As shown in Fig. 2, the correlation coefficient is high across self-attention similarity map in different layers. Such a investigation motivates us to reduce the redundancy in self-attention and improve the efficiency.

We design a shared self-attention (SSA) and apply it in the Transformer network. The repetitive and exhaustive self-attention similarity map calculation is conducted only once in the first layer and then shared by later layers. As shown in Fig. 3, the similarity map in later layers are inherited from the first layer with a lightweight learnable transformation function $F(\cdot)$. We use a 3×3 depth-wise convolution as the transformation to enhance the self-attention similarity map S progressively.

Given N visual tokens with C channels, the computational cost for vanilla self-attention is

$$\Omega(\text{SA}) = 2N^2C + 4NC^2. \tag{6}$$

And the computational cost for shared self-attention is

$$\Omega(\text{SSA}) = 9NC + N^2C + 2NC^2. \tag{7}$$

The computational cost of SSA is about half of the SA. When the token number N is 64 with channel dimension C equals to 180, the computational cost of shared self-attention is just 51% of the vanilla self-attention.

Structure Reparameterized Convolution. To improve the efficiency of multi-scale depth-wise convolutions (MS-DWC), we take use of the reparameterization trick to merge multi-branch feature extraction into a single branch. As shown in Fig. 5, convolutions with different kernel sizes can be reparameterized as the same kernel size in structure, and the weights can be aggregated with equivalence. During the training phase, the feed-forward network contains multi-branch convolutions with different kernel sizes. In the inference phase, the learned kernel weights are aggregated into a single depth-wise convolution. Multi-branch inference is avoided and the computational cost is reduced from $\Omega(84NC)$ to $\Omega(49NC)$. The actual inference time for a feed-forward network also decreases from 0.15s to 0.02s with structure reparameterization. In this way, the proposed representation sharing mechanism can boost the efficiency of the multi-scale design.

4. Experiments

4.1. Experimental Setup

Datasets and Metrics Following the setting in SwinIR [35], we use DIV2K [38], and DF2K (DIV2K + Flickr2K [1]) as the training sets. DIV2K [38] consists of 800 training images and 100 validation images. Following [35], we use the 800 training images to train our model. DF2K is a merged training set of DIV2K and Flicker2K [1], and includes 3,450 (800+2,650) images in total. We test our

Table 1: Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for lightweight image SR on benchmark datasets. Best and second best performance are in red and blue colors, respectively.

Method	Scale	#Params	#Mult-Adds	Set5 [4]		Set14 [72]		BSD100 [48]		Urban100 [27]		Manga109 [49]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CARN [2]	×2	1,592K	222.8G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
IMDN [28]	×2	694K	158.8G	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
LAPAR-A [33]	×2	548K	171.0G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
LatticeNet [44]	×2	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
HPUN-L [54]	×2	714K	151.1G	38.10	0.9608	33.78	0.9201	32.28	0.9010	32.49	0.9318	39.08	0.9779
SwinIR [35]	×2	878K	195.6G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
MSRA-SR (Ours)	×2	769K	196.0G	38.23	0.9614	34.01	0.9211	32.33	0.9017	32.98	0.9358	39.24	0.9783
CARN [2]	×3	1,592K	118.8G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
IMDN [28]	×3	703K	71.5G	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
LAPAR-A [33]	×3	544K	114.0G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
LatticeNet [44]	×3	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	-	-
HPUN-L [54]	×2	723K	69.3G	34.58	0.9282	30.46	0.8445	29.19	0.8073	28.39	0.8582	33.96	0.9467
SwinIR [35]	×3	886K	87.2G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
MSRA-SR (Ours)	×3	777K	91.5G	34.65	0.9291	30.60	0.8470	29.24	0.8093	28.86	0.8664	34.29	0.9489
CARN [2]	×4	1,592K	90.9G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
IMDN [28]	×4	715K	40.9G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
LAPAR-A [33]	×4	659K	94.0G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
LatticeNet [44]	×4	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
HPUN-L [54]	×4	734K	39.7G	32.38	0.8969	28.72	0.7847	27.66	0.7393	26.36	0.7947	30.83	0.9124
SwinIR [35]	×4	897K	49.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
MSRA-SR (Ours)	×4	789K	53.6G	32.46	0.8984	28.86	0.7876	27.72	0.7419	26.65	0.8037	31.08	0.9157

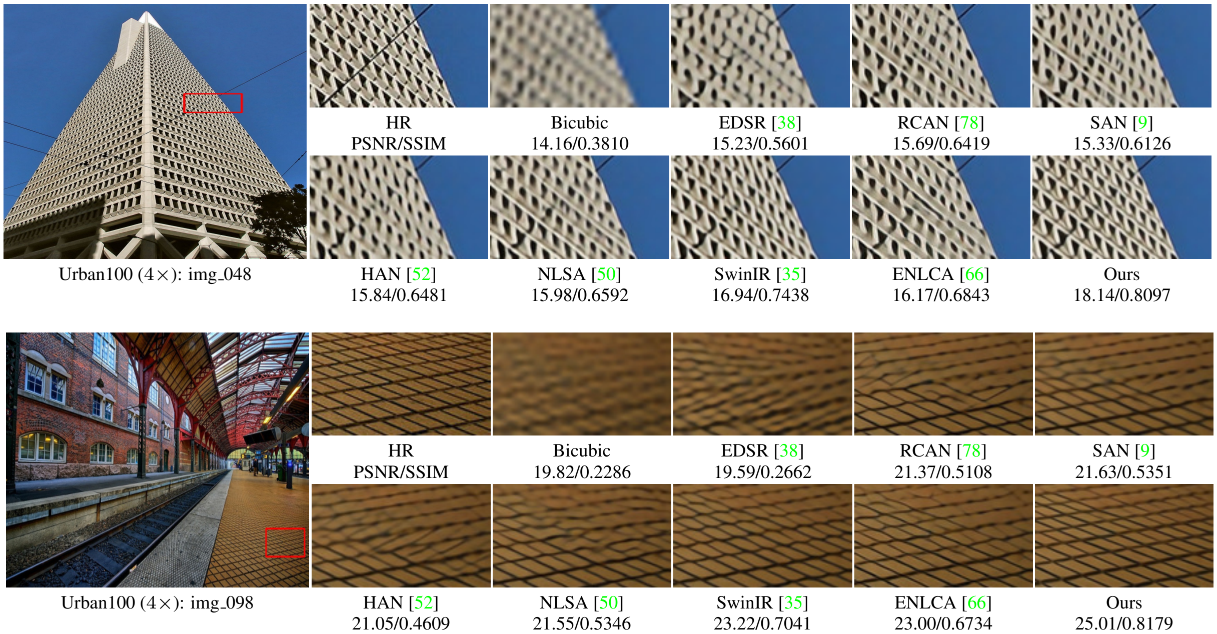


Figure 6: Visual results for classical image SR (4×) with BI degradations on the Urban100 dataset.

model on 5 standard image super-resolution benchmarks: Set5 [4], Set14 [72], BSD100 [48], Urban100 [27] and Manga109 [49]. We use the bicubic degradation for the lightweight and classical image SR tasks. As for evaluation metrics, PSNR and SSIM on the Y channel of transformed YCbCr space are used to evaluate the performance.

Implementation Details Following SwinIR [35], the training patch size is 48×48 for classical image SR on DIV2K and 64×64 for other experiments, respectively. For lightweight image SR task, the Transformer block depth, Transformer block number, window size, channel number, mlp ratio and attention head number are set to 4, 4, 16, 60, 2 and 6, respectively. While for classical image SR task, the above parameters are 4, 6, 16, 192, 2 and 6, respectively. Global Attn is a conventional channel attention. Shared Self Attn is same with the window attention in SwinIR, with our newly proposed sharing mechanism. The mini-batch size is 64 for lightweight SR and 32 for classical SR. The total training iterations are 500K. The model is trained with Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The model is marked with a symbol "+" when self-ensemble strategy [38] is adopted in testing. Our implementations are based on the MindSpore framework.

Table 2: Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for classical image SR with bicubic degradation model on benchmark datasets. Best and second best performance are in red and blue colors, respectively.

Method	Scale	Training Dataset	Set5 [4]		Set14 [72]		BSD100 [48]		Urban100 [27]		Manga109 [49]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN [78]	×2	DIV2K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SAN [9]	×2	DIV2K	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
HAN [52]	×2	DIV2K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
NLSA [50]	×2	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
SwinIR [35]	×2	DIV2K	38.35	0.9620	34.14	0.9227	32.44	0.9030	33.40	0.9393	39.60	0.9792
ENLCA [66]	×2	DIV2K	38.37	0.9618	34.17	0.9229	32.49	0.9032	33.56	0.9398	39.64	0.9791
MSRA-SR (Ours)	×2	DIV2K	38.40	0.9621	34.26	0.9233	32.48	0.9034	33.77	0.9421	39.78	0.9804
MSRA-SR+ (Ours)	×2	DIV2K	38.44	0.9622	34.31	0.9237	32.51	0.9036	33.90	0.9428	39.86	0.9807
IPT [7]	×2	ImageNet	38.37	-	34.43	-	32.48	-	33.76	-	-	-
DFSA [47]	×2	DIV2K+Flickr2K	38.38	0.9620	34.33	0.9232	32.50	0.9036	33.66	0.9412	39.98	0.9798
SwinIR [35]	×2	DIV2K+Flickr2K	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
ART-S [73]	×2	DIV2K+Flickr2K	38.48	0.9625	34.50	0.9258	32.53	0.9043	34.02	0.9437	40.11	0.9804
MSRA-SR (Ours)	×2	DIV2K+Flickr2K	38.50	0.9625	34.63	0.9262	32.55	0.9044	34.15	0.9446	40.18	0.9809
MSRA-SR+ (Ours)	×2	DIV2K+Flickr2K	38.55	0.9627	34.69	0.9264	32.57	0.9047	34.27	0.9454	40.24	0.9811
RCAN [78]	×3	DIV2K	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SAN [9]	×3	DIV2K	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
HAN [52]	×3	DIV2K	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
NLSA [50]	×3	DIV2K	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
SwinIR [35]	×3	DIV2K	34.89	0.9312	30.77	0.8503	29.37	0.8124	29.29	0.8744	34.74	0.9518
MSRA-SR (Ours)	×3	DIV2K	34.95	0.9316	30.82	0.8514	29.41	0.8136	29.50	0.8787	34.95	0.9528
MSRA-SR+ (Ours)	×3	DIV2K	35.00	0.9320	30.93	0.8524	29.44	0.8141	29.62	0.8804	35.13	0.9536
IPT [7]	×3	ImageNet	34.81	-	30.85	-	29.38	-	29.49	-	-	-
DFSA [47]	×3	DIV2K+Flickr2K	34.92	0.9312	30.83	0.8507	29.42	0.8128	29.44	0.8761	35.07	0.9525
SwinIR [35]	×3	DIV2K+Flickr2K	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
ART-S [73]	×3	DIV2K+Flickr2K	34.98	0.9318	30.94	0.8530	29.45	0.8146	29.86	0.8830	35.22	0.9539
MSRA-SR (Ours)	×3	DIV2K+Flickr2K	35.01	0.9322	30.99	0.8541	29.48	0.8148	30.04	0.8853	35.34	0.9542
MSRA-SR+ (Ours)	×3	DIV2K+Flickr2K	35.06	0.9325	31.04	0.8546	29.50	0.8154	30.18	0.8882	35.46	0.9547
RCAN [78]	×4	DIV2K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SAN [9]	×4	DIV2K	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
HAN [52]	×4	DIV2K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
NLSA [50]	×4	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
SwinIR [35]	×4	DIV2K	32.72	0.9021	28.94	0.7914	27.83	0.7459	27.07	0.8164	31.67	0.9226
ENLCA [66]	×4	DIV2K	32.67	0.9004	28.94	0.7892	27.82	0.7452	27.12	0.8141	31.33	0.9188
MSRA-SR (Ours)	×4	DIV2K	32.77	0.9035	29.08	0.7927	27.86	0.7477	27.22	0.8193	31.85	0.9243
MSRA-SR+ (Ours)	×4	DIV2K	32.88	0.9043	29.14	0.7938	27.90	0.7484	27.30	0.8212	32.03	0.9255
IPT [7]	×4	ImageNet	32.64	-	29.01	-	27.82	-	27.26	-	-	-
DFSA [47]	×4	DIV2K+Flickr2K	32.79	0.9019	29.06	0.7922	27.87	0.7458	27.17	0.8163	31.88	0.9266
SwinIR [35]	×4	DIV2K+Flickr2K	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
ART-S [73]	×4	DIV2K+Flickr2K	32.86	0.9029	29.09	0.7942	27.91	0.7489	27.54	0.8261	32.13	0.9263
MSRA-SR (Ours)	×4	DIV2K+Flickr2K	32.95	0.9040	29.13	0.7954	27.93	0.7493	27.68	0.8291	32.22	0.9269
MSRA-SR+ (Ours)	×4	DIV2K+Flickr2K	33.05	0.9048	29.18	0.7962	23.96	0.7499	27.80	0.8335	32.41	0.9280

4.2. Comparison with State-of-the-Arts

Lightweight image SR. Firstly, we provide the comparison of MSRA-SR lightweight version against state-of-the-art lightweight image SR methods, including: CARN [2], IMDN [28], LAPAR [33], LatticeNet [44], HPUN-L [54] and SwinIR [35]. In addition to PSNR and SSIM, the total numbers of parameters as well as multiply-accumulate operations (Mult-Adds) evaluated on a 1024 × 720 HR image are reported to compare the model size and computational cost of different methods. As listed in Table 1, our MSRA-SR achieves the best results on different benchmark datasets with similar numbers of parameters and multiply-accumulate operations. In particular, our method reaches a maximum PSNR increase of 0.22dB in Urban100 (×2) and 0.32dB in Manga109 (×3) compared with SOTA methods, verifying the model's effectiveness and efficiency.

Table 3: Number of parameters, Mult-Adds, and performance with scaling factor × 4 in classical image SR task. The Mult-Adds are calculated with HR image size 1024 × 720 × 3.

Method	RCAN	SwinIR	ART-S	MSRA-SR
Params(M)	15.6	11.9	11.9	12.2
Mult-Adds (G)	832	638	941	645
PSNR(dB)@Urban100	26.82	27.45	27.54	27.68
PSNR(dB)@Manga109	31.22	32.03	32.13	32.22

Classical image SR. Table 2 shows the overall performance comparisons with SOTA methods on classical image SR task with bicubic degradation. The compared methods include RCAN [78], SAN [9], HAN [52], NLSA [50], CRAN [79], DFSA [47], IPT [7], SwinIR [35], ENLCA [66] and ART-S [73]. We choose ART-S for a fair comparison with our method and SwinIR. The detailed model size and computational cost are listed in Table 3.

In Table 2, the proposed method MSRA-SR achieves the best or second best performance on all benchmarks with all scaling factors. For the first setting which adopts DIV2K as the training set, our method gets more obvious performance gains on Urban100 and Manga109, which are mainly composed of images with complicated structures. In particular, our method reaches a maximum PSNR increase of 0.14dB in Set14 (×4), 0.21dB in Urban100 (×2), 0.21dB

Table 4: Detailed ablations. ‡means the layouts (model depth and width) are adjusted for a fair comparison. Final choices are bolded.

Model	#Param (K)	FLOPs (G)	Urban100 PSNR (dB)	Manga109 PSNR (dB)
w/o Global Attn (GA) ‡	690	205.2	32.84	39.13
w/o Cross-Scale Attn (CSA) ‡	728	188.4	32.88	39.16
GA+CSA	769	196.0	32.94	39.21
w/o Shared Attn	915	229.3	32.94	39.20
w Shared Attn	769	196.0	32.94	39.21
w/o MS-DWC	671	177.4	32.85	39.16
w/o MS-DWC, +DWC(K=7)	769	196.0	32.91	39.19
MS-DWC	845	208.9	32.94	39.21
MS-DWC, + Reparam.	769	196.0	32.94	39.21
window=8 ‡	824	179.7	32.84	39.14
window=16	769	196.0	32.94	39.21
$F(\cdot)$=identity	766	184.1	32.77	32.10
$F(\cdot)$=1×1 DWC	766	185.5	32.79	32.11
$F(\cdot)$**=3×3 DWC**	769	196.0	32.94	39.21
$F(\cdot)$=5×5 DWC	774	216.8	32.95	39.21

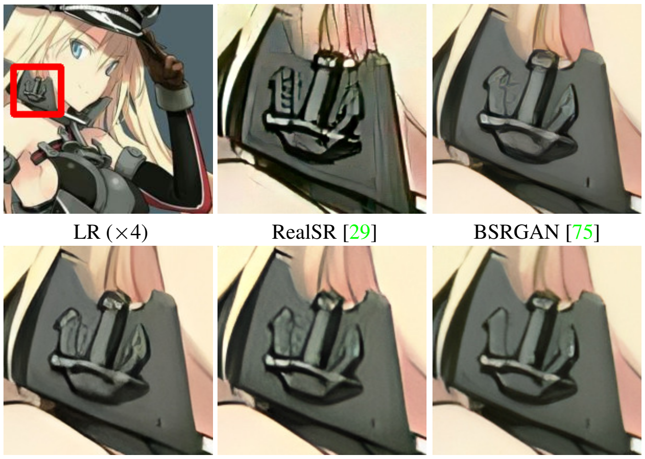
in Manga109 (×3) compared with SOTA methods. For the second setting that uses DIVK2+Flickr2K as the training set, our method also achieves the best performance on different benchmarks. Following SwinIR [35], we also report the performance of MSRA-SR with self-ensemble (marked with "+") for the sake of completeness.

In Fig. 6, we show some visual comparisons on Urban100 with scale ×4. All the methods are trained on DIV2K and tested without self-ensemble for a fair comparison. The results show that our method can not only restore the correct structure of the low-resolution images, but also generate clear details. In image "img_048", existing methods suffer from blurry artifacts, and our method restores more vivid texture details compared with other methods. In image "img_098", the bricks restored by our method have clearer details and correct structure.

Real-world image SR. To verify the effectiveness of the proposed method on real-world applications, we follow the setting in SwinIR [35] and train our MSRA-SR for real-world image SR task. Briefly, we use the model in classical image SR (i.e., middle size) and adopt the degradation process in BSRGAN [75] for low-resolution image synthesis. We use the DIV2K, Flickr2K and OST datasets for training and evaluate the performance on RealSRSet. Following SwinIR, our MSRA-SR is trained with L_1 loss for 1,000K iterations in the first stage, and with GAN loss for 600K iterations in the second stage. Fig. 7 shows the super-resolution results by different methods on real-world LR images.

4.3. Ablation Study

For ablation study, we train MSRA-SR on DIV2K for lightweight image SR (×2) and test it on the Urban100 and Manga109 datasets. All the model variants are trained for 300K iterations with a mini-batch size 32. We analyze the module effectiveness by removing a single module iteratively from the whole model architecture. The overall ablation results are reported in Table 4.


LR (×4) RealSR [29] BSRGAN [75]

Real-ESRGAN [59] SwinIR [35] **MSRA-SR** (ours)

Figure 7: Visual comparison of real-world image SR (×4) methods on real-world images.

Multi-scale Representation Acquisition. We introduce both efficient global attention and local multi-scale feature extraction to enhance the multi-scale representation acquisition. As shown in Table 4, channel attention serves as an efficient global attention (Global Attn) and can complement with the convolution and self-attention. The multi-scale depth-wise convolutions (MS-DWC) with different kernel sizes also contribute to the performance improvement. The results verify that introducing multi-scale depth-wise convolution into the feed-forward network helps the multi-scale feature extraction. Cross-scale attention are also verified to be effective in multi-scale feature acquisition.

Representation Sharing Mechanism. For the shared self-attention, we find that directly applying attention sharing across different layers even improves the performance slightly (from 39.20dB to 39.21dB). The performance variations in Table 4 and Fig. 2 verify the redundancy of self-attention map calculation. Shared Attn reduces the redundancy and uses depthwise convolutions to incorporate local prior to self attention. The feature dependency is refined progressively. Besides, the efficiency brought by shared self-attention is also obvious. The model size and computational cost is reduced by about 20%. For the multi-scale depth-wise convolution, incorporating structure reparameterization in inference stage reduces the runtime time of a feed-forward network from 0.15s to 0.02s.

We also validate other variants for self-attention map transformation function $F(\cdot)$. As shown in Table 4, inserting a depth-wise convolution between self-attention maps performs better than using an identity mapping, verifying the necessity of adding transformation to the self-attention map. The performance saturates when increasing kernel size from 3 to 5. Therefore, we use a 3 × 3 depth-wise convolution as transformation function $F(\cdot)$ for efficiency.

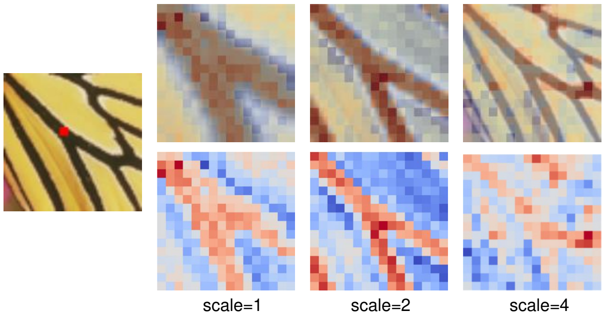


Figure 8: Attention map visualization on cross-scale attention.

Visualization on the Cross-Scale Attention In this section, we provide the attention map visualization of cross-scale attention (CSA) in Fig. 8. As mentioned in the main paper, the similarity matching in CSA is performed in a cross-scale manner, enhancing the feature extraction in different scales. Besides, the receptive field of downscaled key and value features in the window attention is bigger than that in vanilla window attention. When the downscale s is bigger, the receptive field in the image space is bigger correspondingly. From the visualization result, in different scales, different visual tokens are activated in the similarity matching, verifying the effectiveness of cross-scale attention in capturing features from different scales.

5. Conclusion

In this paper, we have proposed a multi-scale shared representation acquisition Transformer (MSRA-SR) for image super-resolution. The multi-scale feature acquisition is integrated into self-attention and feed-forward network. Both global and multi-scale local features are exploited explicitly with cross-scale attention and multi-scale depth-wise convolution. To improve the efficiency of the multi-scale design, we further design a representation sharing mechanism. In shared self-attention, the attention map is computed only once in the first layer and then shared by later layers. Convolutions with different kernel sizes in different branches can be reparameterized as the same kernel size in structure, and the weights can be aggregated into a single convolution with equivalence. Extensive experiments shows that our MSRA-SR achieves competitive performance on lightweight, classical and real-world image SR tasks.

Acknowledgement

This work is partially funded by National Natural Science Foundation of China (Grant No.62006228, U21B2045), Youth Innovation Promotion Association CAS (Grant No.2022132) and CAAI-Huawei MindSpore Open Fund.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 5

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, pages 252–268, 2018. 6, 7

[3] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE TPAMI*, 44(3):1192 – 1204, 2020. 2

[4] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, pages 1–10, 2012. 6, 7

[5] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Real-world image denoising with deep boosting. *IEEE TPAMI*, 42(2):710 – 722, 2019. 1

[6] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *CVPR*, pages 357–366, 2021. 3

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 3, 7

[8] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *ICLR*, 2023. 3

[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 6, 7

[10] Tao Dai, Hua Zha, Yong Jiang, and Shu-Tao Xia. Image super-resolution via residual block attention networks. In *ICCVW*, pages 3879–3886, 2019. 2

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014. 2

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2015. 1

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, pages 12124–12134, 2022. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[16] Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. Facial-sketch synthesis: A new

challenge. *Machine Intelligence Research*, 19(4):257–287, 2022. 2

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021. 3

[18] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dual variational generation for low shot heterogeneous face recognition. In *NeurIPS*, pages 2670–2679, 2019. 2

[19] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dvg-face: Dual variational generation for heterogeneous face recognition. *IEEE TPAMI*, 44(6):2938–2952, 2021. 2

[20] Chaoyou Fu, Xiaoqiang Zhou, Weizan He, and Ran He. Towards lightweight pixel-wise hallucination for heterogeneous face recognition. *IEEE TPAMI*, 2022. 2

[21] Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009. 1

[22] Ran He, Man Zhang, Liang Wang, Ye Ji, and Qiyue Yin. Cross-modal subspace learning via pairwise constraints. *IEEE Transactions on Image Processing*, 24(12):5543–5556, 2015. 2

[23] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018. 3

[24] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet domain generative adversarial network for multi-scale face hallucination. *IJCV*, 127(6-7):763–784, 2019. 2

[25] Huaibo Huang, Xiaoqiang Zhou, Jie Cao, Ran He, and Tieniu Tan. Vision transformer with super token sampling. In *CVPR*, pages 22690–22699, 2023. 3

[26] Huaibo Huang, Xiaoqiang Zhou, and Ran He. Orthogonal transformer: An efficient vision transformer backbone with token orthogonalization. In *NeurIPS*, pages 14596–14607, 2022. 3

[27] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 6, 7

[28] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, pages 2024–2032, 2019. 6, 7

[29] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, pages 466–467, 2020. 8

[30] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. In *NeurIPS*, pages 18590–18602, 2021. 3

[31] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016. 1

[32] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, pages 517–532, 2018. 1

[33] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. In *NeurIPS*, pages 20343–20355, 2020. 6, 7

[34] Yawei Li, Vagia Tsiminaki, Radu Timofte, Marc Pollefeys, and Luc Van Gool. 3d appearance super-resolution with deep learning. In *CVPR*, pages 9671–9680, 2019. 2

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 1, 3, 4, 5, 6, 7, 8

[36] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *CVPR*, pages 5657–5666, 2022. 2

[37] Yuxuan Liang, Pan Zhou, Roger Zimmermann, and Shuicheng Yan. Dualformer: Local-global stratified transformer for efficient video recognition. In *ECCV*, pages 577–595, 2022. 3

[38] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 2, 5, 6

[39] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[40] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE TPAMI*, pages 1–19, 2022. 1

[41] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NeurIPS*, pages 1673–1682, 2018. 3

[42] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, pages 2359–2368, 2020. 1, 2

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002, 2021. 3

[44] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *ECCV*, pages 272–289, 2020. 6, 7

[45] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *CVPR*, pages 17642–17652, 2022. 2

[46] Xin Ma, Xiaoqiang Zhou, Huaibo Huang, Gengyun Jia, Zhenhua Chai, and Xiaolin Wei. Contrastive attention network with dense field estimation for face completion. *Pattern Recognition*, 124:108465, 2022. 2

[47] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV*, pages 4288–4297, 2021. 1, 7

[48] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and

its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–423, 2001. 6, 7

[49] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 6, 7

[50] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, pages 3517–3526, 2021. 3, 6, 7

[51] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, pages 5690–5699, 2020. 2

[52] 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 *ECCV*, pages 191–207, 2020. 2, 6, 7

[53] Pejman Rasti, Tonis Uiboupin, Sergio Escalera, and Gholamreza Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In *International conference on articulated motion and deformable objects*, pages 175–184, 2016. 1

[54] Bin Sun, Yulun Zhang, Songyao Jiang, and Yun Fu. Hybrid pixel-unshuffled network for lightweight image super-resolution. In *AAAI*, pages 2375–2383, 2023. 6, 7

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[56] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *ICCV*, pages 4692–4701, 2021. 3

[57] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *CVPR*, pages 4917–4926, 2021. 2

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021. 3

[59] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 1, 8

[60] 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 *ECCVW*, pages 63–79, 2018. 1, 2

[61] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 3

[62] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song. Unsupervised real-world image super resolution via domain-distance aware training. In *CVPR*, pages 13385–13394, 2021. 1

[63] Dong Wu, Manwen Liao, Weitian Zhang, Xinggang Wang, Xiang Bai, Wenqing Cheng, and Wenyu Liu. YOLOP: you only look once for panoptic driving perception. *Machine Intelligence Research*, 19(6):550–562, 2022. 2

[64] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. Animesr: Learning real-world super-resolution models for animation videos. In *NeurIPS*, pages 11241–11252, 2022. 1

[65] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *IEEE TPAMI*, 2022. 3

[66] Bin Xia, Yucheng Hang, Yapeng Tian, Wenming Yang, Qingmin Liao, and Jie Zhou. Efficient non-local contrastive attention for image super-resolution. In *AAAI*, pages 2759–2767, 2022. 6, 7

[67] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *NeurIPS*, pages 30008–30022, 2021. 3

[68] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Graph information bottleneck for subgraph recognition. In *ICLR*, 2021. 2

[69] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. Recognizing predictive substructures with subgraph information bottleneck. *IEEE TPAMI*, 2021. 2

[70] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. In *NeurIPS*, pages 12992–13003, 2021. 3

[71] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 3

[72] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *ICCS*, pages 711–730, 2010. 6, 7

[73] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023. 1, 3, 7

[74] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE TPAMI*, 44(10):6360 – 6376, 2021. 1

[75] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 1, 8

[76] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *ICCV*, pages 2998–3008, 2021. 3

[77] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, pages 649–667, 2022. 2

[78] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 1, 2, 6, 7

[79] Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. Context reasoning attention network for image super-resolution. In *ICCV*, pages 4278–4287, 2021. 7

[80] Xin Zhao, Jiayi Guo, Yueting Zhang, and Yirong Wu. Asymmetric bidirectional fusion network for remote sensing pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2

[81] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *NeurIPS*, pages 3499–3509, 2020. 1

[82] Xiaoqiang Zhou, Junjie Li, Zilei Wang, Ran He, and Tieniu Tan. Image inpainting with contrastive relation network. In *ICPR*, pages 4420–4427, 2021. 2

[83] Yangguang Zhu, Xian Sun, Wenhui Diao, Haoran Wei, and Kun Fu. Dualda-net: Dual-head rectification for cross domain object detection of remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2