This ACCV 2022 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2022
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Multi-granularity Transformer for Image Super-resolution

Yunzhi Zhuge[1][0000−0002−4288−4516] and Xu Jia[2][0000−0003−3168−3505]

[1] The University of Adelaide, Australia
[2] School of Artificial Intelligence, Dalian University of Technology
xjia@dlut.edu

Abstract. Recently, transformers have made great success in computer vision. Thus far, most of those works focus on high-level tasks, e.g., image classification and object detection, and fewer attempts were made to solve low-level problems. In this work, we tackle image super-resolution. Specifically, transformer architectures with multi-granularity transformer groups are explored for complementary information interaction, to improve the accuracy of super-resolution. We exploit three transformer patterns, *i.e.*, the window transformers, dilated transformers and global transformers. We further investigate the combination of them and propose a **Mu**lti-**g**ranularity Trans**former** (MugFormer). Specifically, the window transformer layer is aggregated with other transformer layers to compose three transformer groups, namely, Local Transformer Group, Dilated Transformer Group and Global Transformer Group, which efficiently aggregate both local and global information for accurate reconstruction. Extensive experiments on five benchmark datasets demonstrate that our MugFormer performs favorably against state-of-the-art methods in terms of both quantitative and qualitative results.

1 Introduction

Single Image Super-resolution (SISR) is a low-level computer vision task where high-resolution (HR) images are recovered from their low-resolution (LR) counterparts. It often serves as an important pre-processing or intermediate step in many computer vision techniques to solve other problems, *e.g.*, Semantic Segmentation [36], Object Detection [15] and Text Recognition [39]. As one LR input can be associated with multiple HR images, SISR is an ill-posed problem.

Early methods of SISR rely on interpolation (*e.g.*, bicubic interpolation and discrete wavelet transform) and regularisation. In the past decade, convolutional neural networks (CNNs) have become the standard model for computer vision tasks due to their representational capability. Many CNN based SISR methods [21][49][48] have been proposed. These techniques learn a non-linear mapping between LR input and HR output; outperform traditional methods by a large margin. However, convolution operations are limited to information processing in local neighborhood, restricting the capability of CNN-based models to capturing long-range relationships among pixels, which could provide important internal prior for this task.

Another series of techniques combine Non-local Attention with CNN models to achieve SISR [21][48][27][26]. In RNAN [48], each attention block contains one trunk

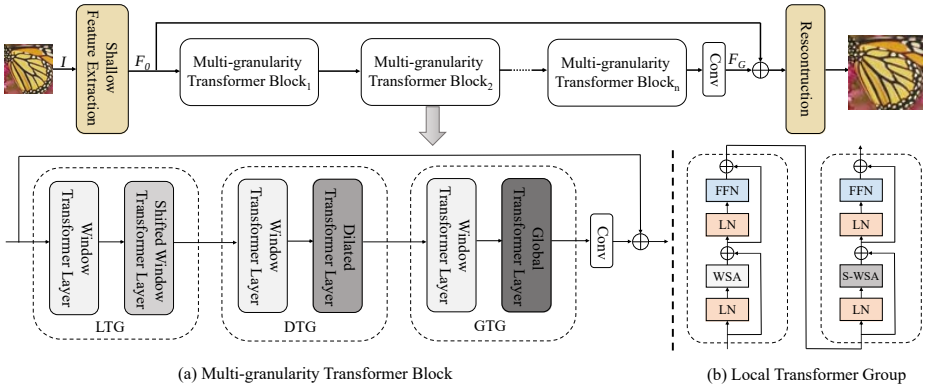
branch and one local mask branch which concentrate on local structures and long-range dependencies respectively. CSNLN [27] studies cross-scale feature correlation of images by learning to mine long-range dependencies between LR features and larger-scale HR patches. However, limited by the computational cost of Non-local attention, those methods are typically inefficient and hard to deploy.

Transformer [34] models have been dominating the field of natural language processing and many studies have demonstrated their ability on vision problems [13][3][33]. For instance, Swin Transformer [23] outperforms state-of-the-art methods by a large margin on image classification, object detection and segmentation. Inspired by the success of vision transformer, some recent works also apply them to low-level vision tasks [5][19][42][50]. Among them, IPT [5] is the pioneering work in which a standard transformer architecture is utilized; aiming at solving multiple restoration problems with large-scale pre-training. Uformer [42] and SwinIR [19] are built on Swin Transformers. The former combines a Swin Transformer with Unet [30], while the latter leverages deep residual connections [20][41][49] to stack several Swin Transformer layers inside each residual block. Although these methods achieve robust results on a series tasks including image super-resolution, image denoising and deraining, directly applying transformers to SISR still has some limitations. (1) The receptive field of local self-attention is restricted to a patch window, preventing the model from building long-range dependencies. (2) Although global self-attention is able to capture context of arbitrary distances, the computational cost can hardly be afforded.

In this work we propose a novel transformer-based method, named **Mu**lti-**g**ranularity Trans**former** (MugFormer), which efficiently aggregate both local and global information in an image to enhance details while maintaining relatively low computational cost. Specifically, a MugFormer block, the major component of the model, is composed of three groups of transformer layers, *i.e*., Local Transformer Groups (LTGs), Dilated Transformer Groups (DTGs) and Global Transformer Groups (GTGs). An LTG comprises a Window Transformer Layer and a shifted Window Transformer Layer. The Window Transformer Layer computes self-attention among pixels within a local neighborhood, hence the LTGs are able to aggregate sufficient local information. In a DTG, the shifted Window Transformer Layer is replaced by a Dilated Transformer Layer which is able to capture long-distance relationship among pixels, hence able to aggregate non-local information in a larger range. A GTG consists of a Window Transformer Layer and a Global Transformer Layer. In the Global Transformer Layer, self-attention is computed based on patches instead of pixels, *i.e*., both queries and keys are computed features for patches. In this way, the GTGs can efficiently integrate information at larger scale from the whole image. By stacking several MugFormer blocks, the model is able to make full use of external and internal priors to recover details in an image. A series of ablative models are designed and compared to demonstrate the effectiveness of the proposed method. In the comparisons with state-of-the-art methods, it shows favorable performance against them.

In summary, the main contributions of our work are three-fold:

- We propose three transformer groups, *i.e*., Local Transformer Group (LTG), Dilated Transformer Group (DTG) and Global Transformer Group (GTG), which are capable of capturing context of different ranges for accurate image SR.



(a) Multi-granularity Transformer Block (b) Local Transformer Group

Fig. 1. The architecture of the proposed MugFormer. (a) is the structure of a Multi-granularity Transformer Block and (b) shows the detailed components inside a Local Transformer Group.

- We analysis the characteristics of transformer block and explore the complementarity between them. Exhaustive experiments demonstrate that our arrangement, *i.e.*, from local to global could obtain optimal results.
- Our **Mu**lti-**g**ranularity Trans-**former** (MugFormer) outperforms the state-of-the-art methods on multiple image SR benchmarks.

2 Related work

2.1 Image Super-resolution

The rapid development of digital devices has increased the demand for high-quality graphics. Image restoration algorithms are often used on edge devices to overcome optical hardware bottlenecks. Early approaches are model-based which formulate image restoration as optimization problems [14][12]. Recently, convolutional neural networks have been applied to image restoration and inspired many leaning-based approaches [11], the results of which are much better than the model-based ones. Dong *et al*. [11] first proposed to solve image super resolution. In [18][20][49], deep residual and dense connections are exploited to learn hierarchical features for more effective representation. Considering that images might fall into uneven distribution, some recent works [49][27][51][26] use attention mechanisms to focus learning on challenging areas and capture realtionships across longer distances. Syed *et al*. [45] proposed a multi-scale architecture which extracts and aggregates spatially-precise information in high-resolution and contextual information in low-resolution simultaneously. In [26], deep feature pixels are partitioned into groups and then attention is only calculated within the group, which significantly reduces the computational cost while forcing the module to focus on informative area.

2.2 Vision Transformer

Transformer [34] was first applied on natural language processing (NLP) tasks [10][29]. Motivated by the success of transformers in NLP, a variety of transformer models have

been proposed to solve visual tasks, *e.g*., image classification [13][33][8], object detection [3][52] [**?**], image segmentation [35][43] and video understanding [1]. Seminally, ViT[13] first uses stacked transformer encoders to classify images. To do this input images are partitioned into non-overlapping patches which are used as tokens. This adaptation of transformers surpasses state-of-the-art convolutional architectures on image classification. However, the astonishing performance of this work is due to a hyper-scale training dataset(JFT-300M). Touvron *et al*. [33] introduced an additional distillation token with hard-label distillation that achieves comparable results when training on much smaller datasets(ImageNet-1K). Some recent works[38][37][**?**][7] focus on general backbone design, leading to more flexibility in downstream tasks, *e.g*., object detection and semantic/instance segmentation.

Inspired by the successful adaptations of transformers to vision task, several works have applied transformers to image restoration [5] [42][19]. IPT [5] aims to solve different restoration problems with a multi-task transformer framework. Controversially, IPT is pre-trained on ImageNet and its performance is influenced more by training data quantity than network architecture. Uformer [42] explores several designs which combine window self-attention [23] with UNet [30]. SwinIR [19] stacks window transformer blocks into a deep network. However, both Uformer and SwinIR calculate self-attention in non-overlapping windows, failing to capture patterns between distant tokens. These studies apply traditional transformer blocks to image restoration tasks without modification. In this work, we analyse the characteristics of transformer structures and how they relate to image SR. From this we propose a multi-granularity transformer architecture for image SR.

3 Methodology

In standard transformers the computational complexity of self-attention grows quadratically with input size, which becomes a serious problem when applied to image restoration tasks. Image restoration tasks such as image super-resolution often require high resolution images as inputs for decent performance, making the application of self-attention computationally infeasible. Window-based self-attention [23] is utilized in SwinIR [19] and Uformer [42] to achieve trade-offs between accuracy and computational complexity. However, window-based self-attention limits feature extraction to a local receptive field. Although shifted window partitioning and layer stacking help expanding the receptive field of a window, interactions among distant elements in the input fail to be modelled, especially for high-resolution inputs. In this work, we propose multi-granularity transformer blocks which are able to extract local information and global context in a unified network. The multi-granularity transformer block is composed of three transformer groups with different granularity, *i.e*., Local Transformer Group, Dilated Transformer Group and Global Transformer Group. Each transformer group captures different patterns of texture from local to global which complement one another.

3.1 Overview

The overall pipeline of our proposed MugFormer is shown in Fig. 1. It is composed of three main parts: a shallow CNN stem, a multi-granularity feature enhancement module and an upsampling module. Following the practices of previous work [20][47][19], a single convolutional layer is used to extract shallow features:

$$F_0 = H_{SF}(I), \tag{1}$$

where $F_0 \in \mathbb{R}^{H\times W\times C}$ and $I \in \mathbb{R}^{H\times W\times C'}$ are the extracted shallow features and input image respectively. $H_{SF}(\cdot)$ is a 3×3 convolutional layer. F_0 is then fed to a more powerful feature extraction module, producing feature maps $F_{HF} \in \mathbb{R}^{H\times W\times C}$ as

$$F_{HF} = H_{MF}(F_0), \tag{2}$$

where $H_{MF}(\cdot)$ is the multi-granularity feature enhancement module, which is formed by stacking K Multi-granularity Transformer Blocks (MGTBs). The feature extraction procedure of k-th MGTB can be described as

$$\begin{aligned} F_k = H_{MGTB_k}(F_{k-1}) = H_{MGTB_k} \\ (H_{MGTB_{k-1}}(\dots H_1(F_0)\dots)), \end{aligned} \tag{3}$$

where $MGTB_k(\cdot)$ is the k-th MGTB, F_{k-1} and F_k represent the input and output respectively. At the end of each MGTB we attach a convolutional layer with structured inductive bias after the transformers. Due to the important role of skip connections in both CNNs [49] and transformers [34], they are also adopted here to stabilise training and promote information propagation:

$$F_{HD} = H_{CONV}(F_K) + F_0, \tag{4}$$

where H_{CONV} represents a 3×3 convolutional layer and F_{HD} is the output of the last MGTB. The specific design of Multi-granularity Transformer Block will be detailed later in Sec. 3.2. Finally, the high-resolution output I_{HR} is obtained via the reconstruction module as

$$I_{HR} = H_{REC}(F_{HD}), \tag{5}$$

where $H_{REC}(\cdot)$ represents the reconstruction operation which first upsamples features via a sub-pixel layer [31] and then produces the final super-resolution result with an additional convolutional layer. The network is trained end-to-end with L_1 reconstruction loss.

3.2 Multi-granularity Transformer Block

Recent works [7][4][44] has demonstrated that using both global and local information in transformer models helps achieving better performance. Motivated by them, we further explore the interaction of tokens in transformers. Specifically, we examine three patterns, *i.e.*, window transformers, dilated transformers and global transformers. As shown in Fig. 1(a), a MGTB is composed of a Local Transformer Group, a Dilated

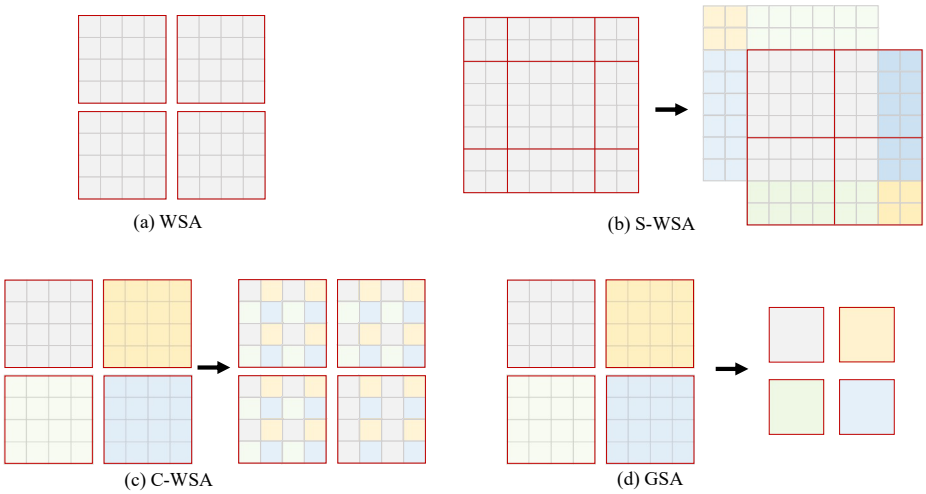


Fig. 2. Schematic views of the Self-attentions used in the proposed three transformer groups.

Transformer Group, a Global Transformer Group and a convolutional layer. Each group contains two consecutive transformer layers and one convolutional layer is added before the residual connection.

Local Transformer Group. Motivated by previous work which reduce the computational cost of transformers by executing self-attention calculation within non-overlapped windows [23], we follow the practice of Swin Transformer [23] to build our Local Transformer Group, which is composed of Window Transformer Layer and shifted Window Transformer Layer. As shown in Fig. 2(a), window self-attention (WSA) evenly divides an input of size $H \times W \times C$ into $\frac{HW}{M^2}$ windows, each of which corresponds to an $M \times M$ patch. Self-attention is computed for all pixels in local neighborhood. Thus, the computational cost of the self-attention is $\mathcal{O}(4HWC^2 + 2HWCM^2)$. This modification makes computation linear with respect to input resolution rather than quadratic, which is the case in standard self-attention $\mathcal{O}(H^2W^2C)$. In a separated window, the details of MSA can be formulated as

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_0, , ...H_n)W^O, \tag{6}$$

$$Head_j = \text{Attention}(QW_j^Q, KW_j^K, VW_j^V), \tag{7}$$

where $Q, K, V \in \mathbb{R}^{n \times d}$ are the embeddings of key, query and value, and $W^O \in \mathbb{R}^{C \times C}$, $W_j^Q \in \mathbb{R}^{C \times d}$,$W_j^K \in \mathbb{R}^{C \times d}$ and $W_j^V \in \mathbb{R}^{C \times d}$ are linear projection matrices. n represents number of heads in the attention layer and d, which is equal to $\frac{C}{n}$, is the dimension of each head. Previous work added relative positional bias to each head which changes the attention map representation to:

$$\text{Attention}(Q, V, K) = \text{Softmax}(QK^T/\sqrt{d} + B)V, \tag{8}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ denote the $query$, key and $value$ in the self-attention module respectively; d is the dimension of $query$ and key. $B \in \mathbb{R}^{G^2 \times G^2}$ is a relative positional

bias matrix. In our task, the sizes of testing images and training images are inconsistent and using a fixed-sized matrix will result in sub-optimal performance [8]. Thus, we adopt dynamic positional bias(DPB) [40] of which the relative position bias is generated dynamically via MLPs.

As shown in Fig. 1(b), window self-attention (WSA) is followed by a feed-forward network (FFN) composed of two Linear layers using ReLU as an activation function. Layer normalisation (LN) is applied to inputs before WSA and the FFN. The whole process of Window Transformer Layer can be represented as

$$\begin{aligned} Z &= \mathrm{WSA}(\mathrm{LN}(Z)) + Z, \\ Z &= \mathrm{FFN}(\mathrm{LN}(Z)) + Z. \end{aligned} \tag{9}$$

In WSA, self-attention is computed in each window and therefore lacks between window interactions. To alleviate this shortcoming we utilise shifted window self-attention (S-WSA) [23]. This method adopts a window partitioning strategy to establish dependency across windows. As shown in Fig. 2(b), window partition starts from the top-left which displaces the windows by $(\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor)$ pixels from the regularly partitioned windows. Although S-WSA can enable some connection between windows, it only models interactions between locally connected windows; failing to capture important long-range dependencies.

Table 1. Ablation study on the compositions of each transformer block. We report the PSNR results on Manga109($4\times$). The performance increases significantly when both three transformers are adopted, comparing with the baseline which only contains WSA-WSA/GSA-GSA Transformer Groups.

WSA-WSA Transformer Group	✓							
GSA-GSA Transformer Group		✓						
Local Transformer Group			✓			✓		✓
Dilated Transformer Group				✓			✓	✓
Group Transformer Group					✓	✓	✓	✓
PSNR	31.65	31.48	31.72	31.73	31.70	31.80	31.78	**31.86**

Dilated Transformer Group. To resolve the limited receptive field in Local Transformer Group, we sample features to obtain a dispersed input for self-attention; similar to atrous convolution [6]. This forms the basis for the Dilated Transformer Group which includes pixels from further away into the computation of self-attention. As shown in Fig. 1(a), a Dilated Transformer Group contains one Window Transformer Layer and one Dilated Transformer Layer. In the computation of dilated self-attention (DSA), the receptive field is expanded by sampling tokens with a large interval rate. For an input of spatial dimension $\lfloor H, W \rfloor$, we sample $\frac{HW}{M^2}$ windows at $I = (\frac{H}{M}, \frac{W}{M})$ intervals. Fig. 2(c) shows an example of DSA when $I = (2, 2)$. Tokens from different locations are effectively reallocated, enabling long-range connections between distant features. It is worth mentioning that DSA establishes hierarchical connections of tokens while maintaining equal computational cost as WSA.

Global Transformer Group. Although Local and Dilated Transformer Groups enable inter-token relationship modelling for short and long distant regions respectively, the self-attention calculation is still limited to a patch. To further expand the receptive field, we introduce global self-attention (GSA). GSA computes self-attention between each patch in the image, as shown in Fig. 2(d). Similar to PVT2 [37], GSA reduces the computational cost by decreasing the spatial scale of key and $value$ with global average pooling before calculating self-attention. We set the down-sampling size R to the window size in WSA. Thus, the computation cost of GSA is $\mathcal{O}(HWCR^2)$. GSA efficiently allows dependency modelling between tokens extracted across the whole image; an important ability for discovering and modelling large scale similarities in image patterns.

3.3 Comparisons with SwinIR

SwinIR is based on the Swin Transformer, which is composed of local window self-attention layers and different windows are connected via window partitioning. However, there are essential differences between Swin Transformer and SwinIR. Swin Transformer is a hierarchical architecture in which the receptive field size expands as features are down-sampled. When the input resolution is 224×224, the features in the last stage have been down-sampled to 7×7, which is equal to the window size. In contrast, the resolution throughout SwinIR remains unchanged. This makes SwinIR prone to relying on local dependencies due to its limited receptive field, especially for high resolution scenes.

Our method investigates information of different ranges by novel combinations of WSA, S-WSA, DSA and GSA, which effectively exploits the local and global context for better results.

4 Experiments

In this section, we elaborate on the datasets, implementation details and experiments to evaluate the efficacy of MugFormer.

4.1 Datasets and Evaluation Metrics.

Following [20][49][27], 800 images from DIV2K [32] training set are used to train MugFormer. We choose 5 standard benchmarks: Set5 [2], Set14 [46], B100 [24], Urban100 [17] and Manga109 [25] as our testing sets with three upscaling factors: $\times 2$, $\times 3$ and $\times 4$. We transform SR outputs into YCbCr space and evaluate performance with PSNR and SSIM metrics on Y channel.

4.2 Implementation Details.

The implementation details of our MugFormer is specified here. In Local, Dilated and Global Transformer Groups, we set the number of attention heads and dimensions to 6 and 180 respectively. These hyperparameters maintaining comparable parameters and

FLOPs with SwinIR. We set the window size of WSA, S-WSA, DSA, and the downsampling size of GSA to 8. The number of Multi-granularity Transformer Blocks is set to 6.

During training, paired images are augmented by randomly applying rotations of $90°$, $180°$ or $270°$ and horizontally flipping. Each mini-batch contains 16 LR patches and with size 64×64. We optimize the model using Adam with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-8$. The initial learning rate is set to $1e-4$ and is reduced by half when iterations reach $\{250000, 400000\}$. The training is done on two Nvidia TITAN RTX GPUs.

Table 2. Ablation study on influence of the transformer blocks orders w.r.t. the peformence.

Methods	PSNR	SSIM
GTG$\longrightarrow$ LTG $\longrightarrow$ DTG	31.79	0.9226
GTG$\longrightarrow$ DTG $\longrightarrow$ LTG	31.78	0.9224
LTG$\longrightarrow$ GTG $\longrightarrow$ DTG	31.83	0.9230
DTG$\longrightarrow$ GTG $\longrightarrow$ LTG	31.83	0.9228
DTG $\longrightarrow$ LTG $\longrightarrow$ GTG	31.84	0.9230
LTG $\longrightarrow$ DTG $\longrightarrow$ GTG	**31.86**	**0.9233**

Table 3. Ablation studies on investigating the impact of transformer blocks.

Group Numbers	Parameters(M)	FLOPs(G)	Set5	Set14	B100	Urban100	Manga109
n=2	4.51	159.92	32.63	28.79	27.75	26.75	31.33
n=3	6.40	220.18	32.70	28.86	27.81	26.90	31.57
n=4	8.29	280.45	32.75	28.89	27.83	26.99	31.70
n=5	10.19	340.71	32.82	28.94	27.86	27.07	31.79
n=6	12.10	400.97	32.86	29.03	27.88	27.16	31.8

4.3 Ablation Study

MugFormer is primarily composed of three transformer groups, discussed in Sec. 3. In this section we conduct ablative experiments to analyze and verify the effectiveness of the proposed architectural units. To begin with, we analyse the efficacy of each transformer group. Then, we illustrate that our arrangement, *i.e.*, from local to global, achieves optimal results.

Multi-granularity Transformer Group. To demonstrate that each transformer group contributes to the final results, we conduct a series experiments on Manga109 and the results are shown in Tab. 1. WSA-WSA Transformer Group and GSA-GSA Transformer Group denote that the transformer group is composed of two consecutive window transformer layers or global transformer layers respectively. WSA-WSA Transformer Group

Table 4. Quantitative comparisons (PSNR/SSIM) with BI degradation on benchmark datasets. Best and second best results are highlighted with red and blue colors, respectively.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR[20]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
DBPN[16]	×2	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
RDN[49]	×2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780
RCAN[47]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
NLRN[21]	×2	38.00	0.9603	33.46	0.9159	32.19	0.8992	31.81	0.9249	–	–
RNAN[48]	×2	38.17	0.9611	33.87	0.9207	32.32	0.9014	32.73	0.9340	39.23	0.9785
SAN[9]	×2	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
RFANet[22]	x2	38.26	0.9615	34.16	0.9220	32.41	0.9026	33.33	0.9389	39.44	0.9783
HAN[28]	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
NLSA[26]	×2	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
SwinIR[19]	×2	38.35	0.9620	34.14	0.9227	32.44	0.9030	33.40	0.9393	39.60	0.9792
Ours	×2	38.38	0.9622	34.19	0.9232	32.46	0.9031	33.43	0.9395	39.64	0.9785
Ours+	×2	38.43	0.9624	34.28	0.9236	32.50	0.9233	33.52	0.9399	39.71	0.9789
EDSR[20]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RDN[49]	×3	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484
RCAN[47]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
NLRN[21]	×3	34.27	0.9266	30.16	0.8374	29.06	0.8026	27.93	0.8453	-	-
RNAN[48]	×3	34.66	0.9290	30.52	0.8462	29.26	0.8090	28.75	0.8646	34.25	0.9483
SAN[9]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
RFANet[22]	x3	34.79	0.9300	30.67	0.8487	29.34	0.8115	29.15	0.8720	34.59	0.9506
HAN[28]	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
NLSA[26]	×3	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
SwinIR[19]	×3	34.89	0.9312	30.77	0.8503	29.37	0.8124	29.29	0.8744	34.74	0.9518
Ours	×3	34.93	0.9318	30.87	0.8520	29.40	0.8132	29.38	0.8756	34.89	0.9521
Ours+	×3	34.98	0.9321	30.92	0.8527	29.42	0.8135	29.48	0.8769	35.01	0.9528
EDSR[20]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
DBPN[16]	×4	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
RDN[49]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
RCAN[47]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
NLRN[21]	×4	31.92	0.8916	28.36	0.7745	27.48	0.7306	25.79	0.7729	-	-
RNAN[48]	×4	32.49	0.8982	28.83	0.7878	27.72	0.7421	26.61	0.8023	31.09	0.9149
SAN[9]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
RFANet[22]	×4	32.66	0.9004	28.88	0.7894	27.79	0.7442	26.92	0.8112	31.41	0.9187
HAN[28]	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
NLSA[26]	×4	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
SwinIR[19]	×4	32.72	0.9021	28.94	0.7914	27.83	0.7459	27.07	0.8164	31.67	0.9226
Ours	×4	32.86	0.9037	29.03	0.7931	27.88	0.7468	27.16	0.8168	31.86	0.9233
Ours+	×4	32.92	0.9041	29.09	0.7940	27.90	0.7474	27.23	0.8194	31.99	0.9245

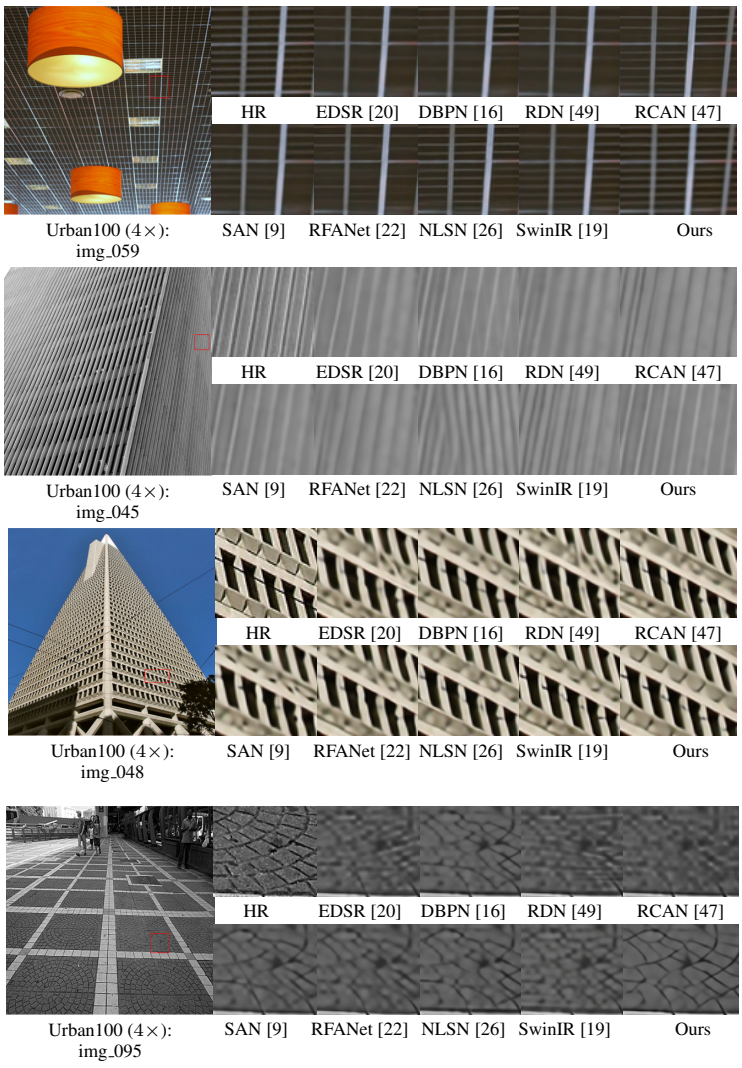


Fig. 3. Visual comparisons for $4\times$ SR with BI degradation on Urban100 dataset.

compute self-attention in windows without connections between them; only extracting local information. In contrast, GSA-GSA Transformer Group contains group transformer layers that capture global context. In the first two columns, we observe that both transformers using purely local or global information perform relatively poorly on Manga109(4x), the PSNR of which are 31.65dB and 31.48 dB respectively. In columns 3-5, the transformer blocks are made with LTGs, DTGs and GTGs respectively. Steady improvement in PSNR is observed as transformer layers model increasing global relationships. Columns 6-7 contain results from transformer blocks consisting of two LTGs/DTGs and one GTG. We find that both combinations increase PSNR, demonstrating that context of different ranges are important when learning to perform single image SR. In the last column where all transformer groups are combined, PSNR is improved from 31.80dB to 31.86dB eventually.

Table 5. Quantitative comparisons (PSNR/SSIM) with BD degradation on benchmark datasets. Best and second best results are highlighted with red and blue colors, respectively.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×3	28.78	0.8308	26.38	0.7271	26.33	0.6918	23.52	0.6862	25.46	0.8149
VDSR [18]	×3	33.25	0.9150	29.46	0.8244	28.57	0.7893	26.61	0.8136	31.06	0.9234
RDN [49]	×3	34.58	0.9280	30.53	0.8447	29.23	0.8079	28.46	0.8582	33.97	0.9465
RCAN [47]	×3	34.70	0.9288	30.63	0.8462	29.32	0.8093	28.81	0.8647	34.38	0.9483
SAN [9]	×3	34.75	0.9290	30.68	0.8466	29.33	0.8101	28.83	0.8646	34.46	0.9487
RFANet [22]	×3	34.77	0.9292	30.68	0.8473	29.34	0.8104	28.89	0.8661	34.49	0.9492
HAN [28]	×3	34.76	0.9294	30.70	0.8475	29.34	0.8106	28.99	0.8676	34.56	0.9494
Ours	×3	34.92	0.9309	30.91	0.8501	29.42	0.8127	29.29	0.8726	34.95	0.9513
Ours+	×3	34.95	0.9311	30.95	0.8505	29.44	0.8130	29.38	0.8740	35.07	0.9518

Influence of transformer block numbers. We show the influence of transformer block numbers in Tab. 3. It can be observed that PSNR, parameters and FLOPs increase steadily as the number of transformer blocks growths, demonstrating that a trade-off between performance and computational cost can be achieved by adaptively changing the number of transformer blocks.

Influence of transformer sequences. As is introduced in Sec. 3, self-attention is calculated in a hierarchical manner from local to global in each transformer block of MugFormer. To verify that this arrangement is optimal, we evaluate the PSNR of transformer groups with different group orderings on Manga109. As shown in Tab. 2, peak performance is reached when we cascading transformer groups by LTG, DTG and GTG. Other permutations result in PSNR dropping from 0.02dB to 0.07 dB.

4.4 Results with Bicubic (BI) Degradation

To verify the effectiveness of MugFormer, we compare our method to 10 others, *i.e.*, EDSR [20], DBPN [16], RDN [49], RCAN [47], NLRN [21], RNAN [48], SAN [9],

RFANet [22], NLSA [26] and SwinIR [19]. Following [20][9][19], we apply the self-ensemble strategy to further boost performance which is denoted as Ours+.

Quantitative results. In Tab. 4, we compare the PSNR and SSIM results for various scaling factors. Compared with other methods, MugFormer achieves the best results on all benchmark datasets at all scales, except for SSIM on Manga109 ($2\times$). In particular, our method increases the PSNR by 0.14dB and 0.19 dB over SwinIR on Set5 and Manga109 at $4\times$ scale. This illustrates that the proposed multi-granularity transformer block could efficiently contextualising information extracted from different ranges to achieve more accurate results. When applying the self-ensemble strategy, PSNR is further improved by 0.06dB and 0.13 dB respectively.

Qualitative Results. In Fig. 3, we show the outputs of a series SR method on the Urban100 benchmark dataset at $4\times$ scale. Our method achieves more visually pleasant results than other methods on a variety of patterns. The textures in 'img_045' are challenging, most other methods either suffer from distortions or fail to restore fine details. Our method recovers the high frequency details clearly and produces less blurring artifacts. 'img_048' contains repeated patterns that demonstrate MugFormer's ability of exploiting different context ranges to produce fine results. In 'img_095', MugFormer recover the edges of the irregular tiles clearly whereas the results produced by other methods are blurred in different levels. Several other methods, *e.g.*, RCAN, RFANet and SwinIR deform the original structure; while our method can keep the shape intact and produce more fidelity results than others.

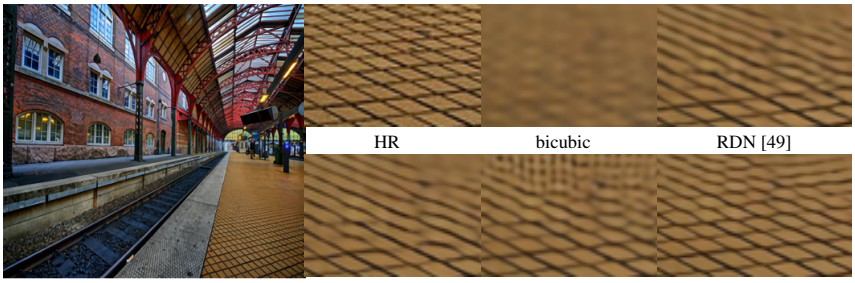


Fig. 4. Visual comparisons for $4\times$ SR with BI degradation on Urban100 dataset.

4.5 Results with Blur-downscale (BD) Degradation

We qualitatively and quantitatively compare the results on benchmark datasets with blur-down degradation (BD), which is a common practice in some recent works [47][49] [9]. In Fig. 4, we can observe that compared with bicubic upsampling, RDN [49], RCAN [47] and RFANet [22], out method restore the shape of tiles more clearly . In Tab. 5, we observe that our method outperforms all other methods by a large margin on the benchmark datasets. The performance gap between MugFormer and other methods is further enlarged when the self-ensemble strategy is used, *e.g.*, Ours+ exceeds

RFANet by 0.39dB on Urban100 and 0.58dB on Manga109. The above comparative analysis demonstrates that our method is robust to BD degradation.

4.6 Complexity Analysis

We compare the model size, FLOPs and PSNR with several SR methods to analyse model complexity in Tab. 6. MugFormer uses significantly less parameters and FLOPs than both EDSR and HAN while achieving better PSNR scores. Specifically, HAN would consume 1679.72 GFLOPs when the input size is $160 \times 160 \times 3$, which is 4 times more than ours. SwinIR has slightly less parameters than our method (11.94 M vs. 12.10 M), while its PSNR score is much worse (31.67dB vs. 31.86dB). The above analysis demonstrate that comparing with previous state-of-the-art methods, our MugFormer achieves better results without paying additional computational cost.

Table 6. Parameters, FLOPs and PSNR scores on Manga109 with 4× factor. The FLOPs are calculated with input size 160×160×3.

Methods	Parameters(M)	FLOPs(G)	PSNR(dB)
EDSR [20]	43.09	1287.03	31.02
RCAN [47]	15.59	408.53	31.22
HAN [28]	64.19	1679.72	31.42
SwinIR [19]	11.94	410.86	31.67
Ours	12.10	400.97	31.86

5 Conclusion

Recent works have validated that vision transformer can achieve state-of-the-art results on low-level vision tasks. However, none of them have been able to exploit the full range of contextual information available when calculating self-attention. In this work, we propose a **mu**lti-**g**ranularity transformer (MugFormer) capable of solving single image super-resolution. Specifically, transformer blocks in MugFormer are composed of three transformer groups with different receptive field sizes: Local Transformer Group, Dilated Transformer Group and Global Transformer Group. Local Transformer Group and Dilated Transformer Group both exchange tokens between neighbouring windows; the former performs on adjacent windows while the latter shuffles them by sampling tokens across the image at fixed intervals. Finally, Global Transformer Group builds connection between all tokens. Experiments demonstrate that the integration of multiple transformer groups achieves state-of-the-art results on benchmark datasets.

Acknowledgement The research was partially supported by the Natural Science Foundation of China, No. 62106036, and the Fundamental Research Funds for the Central University of China, DUT21RC(3)026.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691 (2021)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proc. Brit. Mach. Vis. Conf. (2012)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proc. Eur. Conf. Comp. Vis. pp. 213–229. Springer (2020)
4. Chen, B., Li, P., Li, C., Li, B., Bai, L., Lin, C., Sun, M., Ouyang, W., et al.: Glit: Neural architecture search for global and local image transformer. In: Proc. IEEE Int. Conf. Comp. Vis. (2021)
5. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 12299–12310 (2021)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
7. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting spatial attention design in vision transformers. In: Proc. Adv. Neural Inf. Process. Syst. (2021)
8. Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C.: Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021)
9. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 11065–11074 (2019)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proc. Eur. Conf. Comp. Vis. pp. 184–199. Springer (2014)
12. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE Trans. Image Process. **22**(4), 1620–1630 (2012)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. Int. Conf. Learn. Repr. (2021)
14. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. Image Process. **15**(12), 3736–3745 (2006)
15. Haris, M., Shakhnarovich, G., Ukita, N.: Task-driven super resolution: Object detection in low-resolution images. arXiv preprint arXiv:1803.11316 (2018)
16. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 1664–1673 (2018)
17. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5197–5206 (2015)
18. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 1646–1654 (2016)
19. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. arXiv preprint arXiv:2108.10257 (2021)
20. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshop. pp. 136–144 (2017)

21. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: Proc. Adv. Neural Inf. Process. Syst. (2018)
22. Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G.: Residual feature aggregation network for image super-resolution. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2359–2368 (2020)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE Int. Conf. Comp. Vis. (2021)
24. 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: Proc. IEEE Int. Conf. Comp. Vis. vol. 2, pp. 416–423. IEEE (2001)
25. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. J. Multimedia Tools and Applications **76**(20), 21811–21838 (2017)
26. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 3517–3526 (2021)
27. Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T.S., Shi, H.: Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5690–5699 (2020)
28. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: Proc. Eur. Conf. Comp. Vis. pp. 191–207. Springer (2020)
29. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
31. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 1874–1883 (2016)
32. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshop. pp. 114–125 (2017)
33. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proc. Int. Conf. Mach. Learn. pp. 10347–10357. PMLR (2021)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. Adv. Neural Inf. Process. Syst. pp. 5998–6008 (2017)
35. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 5463–5474 (2021)
36. Wang, L., Li, D., Zhu, Y., Tian, L., Shan, Y.: Dual super-resolution learning for semantic segmentation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 3774–3783 (2020)
37. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvtv2: Improved baselines with pyramid vision transformer. arXiv preprint arXiv:2106.13797 (2021)
38. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proc. IEEE Int. Conf. Comp. Vis. (2021)
39. Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., Bai, X.: Scene text image super-resolution in the wild. In: Proc. Eur. Conf. Comp. Vis. pp. 650–666. Springer (2020)

40. Wang, W., Yao, L., Chen, L., Cai, D., He, X., Liu, W.: Crossformer: A versatile vision transformer based on cross-scale attention. arXiv preprint arXiv:2108.00154 (2021)
41. 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: Proc. Eur. Conf. Comp. Vis. Workshop (2018)
42. Wang, Z., Cun, X., Bao, J., Liu, J.: Uformer: A general u-shaped transformer for image restoration. arXiv preprint arXiv:2106.03106 (2021)
43. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203 (2021)
44. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. In: Proc. Adv. Neural Inf. Process. Syst. (2021)
45. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: Proc. Eur. Conf. Comp. Vis. pp. 492–511. Springer (2020)
46. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proc. Int. Conf. curves and surfaces. pp. 711–730. Springer (2010)
47. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proc. Eur. Conf. Comp. Vis. pp. 286–301 (2018)
48. Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y.: Residual non-local attention networks for image restoration. In: Proc. Int. Conf. Learn. Repr. (2019)
49. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2472–2481 (2018)
50. Zhao, D., Li, J., Li, H., Xu, L.: Hybrid local-global transformer for image dehazing. arXiv preprint arXiv:2109.07100 (2021)
51. Zhao, H., Kong, X., He, J., Qiao, Y., Dong, C.: Efficient image super-resolution using pixel attention. In: Proc. Eur. Conf. Comp. Vis. pp. 56–72. Springer (2020)
52. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)