This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Self-Calibrated Efficient Transformer for Lightweight Super-Resolution

Wenbin Zou[1,*], Tian Ye[2,*], Weixin Zheng[3,*], Yunchen Zhang[4], Liang Chen[1,†], Yi Wu[1]
Fujian Provincial Key Laboratory of Photonics Technology, Fujian Normal University, Fuzhou, China.[1]
School of Ocean Information Engineering, Jimei University, Xiamen, China.[2]
College of Physics and Information Engineering, Fuzhou University, Fuzhou, China.[3]
China Design Group Co., Ltd., Nanjing, China.[4]
alexzou14@foxmail.com, 201921114031@jmu.edu.cn, visinzheng@163.com,
cydiachen@cydiachen.tech, cl_0827@126.com, wuyi@fjnu.edu.cn

Abstract

Recently, deep learning has been successfully applied to the single-image super-resolution (SISR) with remarkable performance. However, most existing methods focus on building a more complex network with a large number of layers, which can entail heavy computational costs and memory storage. To address this problem, we present a lightweight Self-Calibrated Efficient Transformer (SCET) network to solve this problem. The architecture of SCET mainly consists of the self-calibrated module and efficient transformer block, where the self-calibrated module adopts the pixel attention mechanism to extract image features effectively. To further exploit the contextual information from features, we employ an efficient transformer to help the network obtain similar features over long distances and thus recover sufficient texture details. We provide comprehensive results on different settings of the overall network. Our proposed method achieves more remarkable performance than baseline methods. The source code and pre-trained models are available at https://github.com/AlexZou14/SCET.

1. Introduction

Single image super-resolution (SISR) [14] aims to recover a high-resolution (HR) image from its low-resolution (LR) observation, which is a challenging ill-posed problem because many latent HR images can be downsampled to an identical LR image. To address this significant problem, many image super-resolution (SR) methods [11,20,27] based on deep convolution architecture have been proposed and shown impressive performance. Thanks to the powerful representation capabilities of the deep convolution neural networks, numerous previous approaches can learn the complex non-linear mapping from paired LR-HR images.

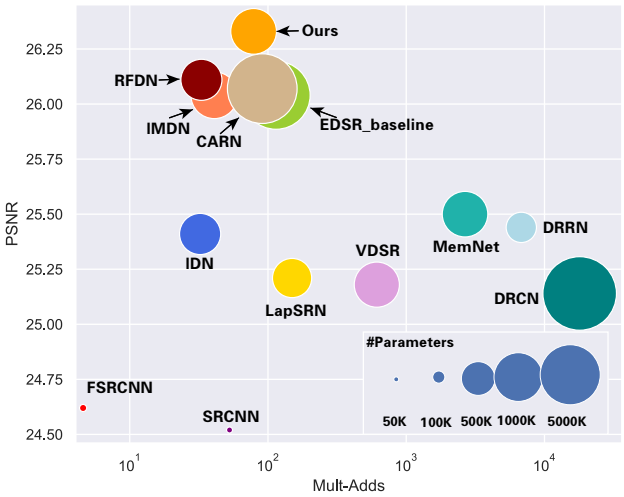

Figure 1. Trade-off between performance vs: number of operations and parameters on Urban100 $\times 4$ dataset. Multi-adds are calculated on 720p HR image. The results show the superiority of our model among existing methods.

Dong *et al.* [11] firstly propose the super-resolution convolutional neural network (SRCNN) that outperforms the previous work. On this basis, various SR algorithms [12, 20, 21, 34] have been proposed with superior performances, and those methods have a large margin compared with traditional methods. It is widely known that deeper networks based on residual learning [16] generally achieve better performances. Based on this cognition, deeper networks with larger frameworks, e.g. enhanced deep super-resolution network (EDSR) [27] and residual channel attention network (RCAN) [50], have been proposed and achieved excellent performance. However, previous CNN-based SR networks have a large number of parameters, re-

*Equal contribution
†Corresponding author

sulting in the limitation of the application of SR technology in edge devices.

A straightforward solution to this problem is to design lightweight and efficient networks via reducing the amount of the parameters, *e.g.*, building shallow networks with a single path [12, 23], recursive operation [21, 34], information distillation mechanism [18, 19], and neural architecture search (NAS) [6, 7]. However, most of these methods focus on local contextual information and do not consider global similar textures, leading to problems such as artifacts in the recovered image. The limited receptive field of convolution operation is difficult to capture globally similar features, resulting in a poor trade-off between performance and complexity.

The image restoration methods based on the transformer architecture have made remarkable progress recently. Yet, there are few studies on the lightweight SR transformer network, which attracts us to explore the following exciting topic:

*How to design a **lightweight** transformer to **effectively** perform single image super-resolution?*

Previous distillation-based solutions achieve impressive SISR performance. However, the above solutions have redundant parameters as the channel-splitting design of extract features progressively in a single basic block. Furthermore, they still have scope for improvement in performance as the spatial and channel modeling ability is relatively weak.

According to the above analysis, the core idea of our approach is how to make lightweight networks with both spatial modeling and channel modeling capabilities. Due to the complexity limitations, it is obviously more efficient to model dependencies in the channel and spatial dimensions respectively. Thus, we propose two complementary components, the SC module and the efficient transformer module to endow the network with powerful modeling capabilities in the spatial dimension and channel dimension respectively.

Self-Calibrated Module. We propose the SC module as the efficient extractor to explore the valuable spatial features from low-resolution input. With the help of the spatial attention mechanism, it adaptively pays more attention to the detailed textures. Therefore, the SC module provides strong spatial clues for the following transformer module.

Efficient Transformer Module. We construct a linear-complexity transformer module to perform channel-wise self-attention mechanism, which efficiently models the dependence in the channel dimension from input features. The combination of two proposed modules provides complementary clues in the channel and spatial dimensions for the HR image reconstruction.

Based on above components, we propose a lightweight **S**elf-**C**alibrated **E**fficient **T**ransformer (**SCET**) network to solve the SISR problem efficiently. For instance, our method achieves higher performance than the state-of-the-art lightweight SR method A^2F-M [42] with 0.53 dB PSNR gain on the $\times 4$ Manga109 [31] dataset, the number of parameters in SCET only 68.3% of A^2F-M. The SCET method is a competing entry in NTIRE 2022 Efficient Super-Resolution challenge [25].

The key contributions of this work are as follows:

- We introduce the efficient transformer design to the lightweight SISR task, effectively exploiting to the property that the transformer module can capture long-range dependencies, avoiding the problem of wrong textures generated by current lightweight SR methods.
- We design the SC module as the high-performance extractor. Compared with the information distillation mechanism in the IMDB block [18], the SC module employs a more efficient feature propagation strategy, achieving better performance with fewer parameters and less computational effort.
- As shown in Figure 1, our SCET occupies fewer parameters and takes fewer Multi-Adds, while significantly improving the performance of SISR networks at low resource consumption.

2. Related Work

2.1. Deep SR models

In recent years, deep CNN is employed in various low-level vision tasks, such as image denoising [1], deblurring [32], and so on. Dong *et al.* [11] make a big step forward by proposing a three-layer fully convolutional network SRCNN. On this basis, Kim *et al.* design deeper network VDSR [20] and DRCN [21] via residual learning. Subsequently, Tai *et al.* [34] later develop a deep recursive residual networks (DRRN) by introducing recursive blocks and then propose a persistent memory network (MemNet) [35] by utilizing memory block. However, the above methods use the bicubic interpolation to preprocess the LR image, which inevitably losses some details and bring large computation. To solve this problem, Dong *et al.* [12] propose FSRCNN by adopting a deconvolution layer to upsample images at the end of the network to decrease computations. Then, Shi *et al.* [33] introduce an efficient sub-pixel convolutional layer instead of deconvolution. On this basis, Lim *et al.* [27] propose a deeper and wider network EDSR by stacking residual blocks (eliminating batch normalization layers). The significant performance gain indicates the fact that the depth and width of the network occupy important places in image SR. Furthermore, some other networks, e.g.

non-local neural network (NLRN) [41], RCAN [50], and second-order attention network (SAN) [9], improve the performances by modeling the correlation of features in space or channel dimensions. Yet, these networks sacrifice the portability of the network, leading to the highly cost in memory storage and computational complexity.

2.2. Lightweight SR models

During these years, many lightweight networks have been working on SR problem. They can be approximately divided into three classes: the architectural design-based methods [2, 19, 23], the knowledge distillation-based methods [15], and the NAS-based methods [6, 7]. The first class mainly focuses on the recursive operation and channel splitting. Deeply-recursive convolutional network (DRCN) [21] and deep recursive residual network (DRRN) [34] are proposed to share parameters via introducing the recursive layers. However, the reduction of computational operation and the amount of parameters are still unsatisfying. Ahn *et al.* design a cascading residual network (CARN) [2], that accomplishes a cascading mechanism based on residual learning. Lightweight feature fusion network (LFFN) [8] uses multi-path channel learning to incorporate multi-scale features. NAS [53], which is an emerging approach to automatically design efficient networks, is introduced to the SR task [6,7]. However, the performances of NAS-based methods are limited by the search space and strategies. IMDN [18] extracts hierarchical features step by step through splitting operations and further improves the efficiency of the model. On this basis, RFDN [28] has further improved the information multi-distillation block in IMDN and won the first place at the Efficient Super-Resolution Challenge in AIM 2020 [47]. Inspired by SCNet [29], Zhao *et al.* [51] employ a self-calibrated convolution with pixel attention block, which further reduces the network parameters and improves the network operation speed. Therefore, we employ the self-calibrated convolution scheme in our SCET network for efficient SR.

2.3. Vision Transformer

The breakthroughs from Transformer in the NLP area lead to sigificant interest in the computer vision community. It has been successfully applied in image recognition [13, 24, 38], object detection [4, 10] and segmentation [40,43]. Currently, most Vision Transformer split the image into a sequence of patches and then flatten them into vectors to learn their interrelationships through self-attention. Therefore, the Vision Transformers possesses the strong capability to learn long-term dependencies between image pixel. Owing to its powerful learning capabilities, Transformer is introduced to low-level vision tasks [5, 26, 39, 44] and obtained excellent performance recently. However, the self-attention mechanism in the Transformer introduces a huge amount of computation and GPU resource consumption, which is not friendly to lightweight networks. Therefore, building efficient Vision Transformer has become a hot research topic in recent years.

3. Self-Calibrated Efficient Transformer

In this section, we present the overall architecture of the proposed Self-Calibrated Efficient Transformer (SCET) firstly. Then, we introduce the lightweight self-calibrated (SC) module, which consists of several stacked self-calibrated convolutions with pixel attention (SCPA) blocks to efficiently extract texture information from images. Finally, we describe the efficient transformer module.

3.1. Overview of Network Framwork

Considering that complex network structure blocks may bring a large number of parameters and complexity, we choose a simple network structure, as shown in Figure 2. Our SCET mainly consists of two parts: SC module and efficient transformer module. Specifically, the SC module is used to efficiently extract image texture features and the Efficient Tranformer module is used to recover similar textures across long ranges.

Given an input low-resolution image $I_{LR} \in \mathbb{R}^{H\times W\times 3}$, SCET first applies a convolution to obtain shallow feature $F_0 \in \mathbb{R}^{H\times W\times C}$, where $H \times W$ denotes the spatial dimension and C is the number of channels. It can be formulated as:

$$F_0 = H_{conv}(I_{LR}), \tag{1}$$

where H_{conv} denotes 3×3 convolution operation. Next, inspired by PAN [51], we employed an SC module composed of SCPA blocks to efficiently extract the deep texture feature. It can be expressed as:

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

where H_{SC} denotes SC module, F_{SC} denotes the output of SC module. To obtain global similarity information, we use the efficient transformer module to further recover similar textures across long distances. Inspired by Restormer [45] that the amount of computation can reduce from $\mathcal{O}(W^2H^2)$ to $\mathcal{O}(C^2)$ by applying self-attention to compute cross-covariance across channels, we employ the multi-Dconv head transposed attention (MDTA) to generate an attention map encoding the global context implicitly. Besides, we adopt a gated-Dconv feed-forward network (GDFN) to focus on the fine texture details complimentary. It can be written as:

$$F_{out} = H_{ET}(F_{SC}) = H_{GDFN}(H_{MDTA}(F_{SC})), \tag{3}$$

where H_{ET}, H_{MDTA}, and H_{GDFN} denote the efficient transformer, MDTA and GDFN, respectively. F_{out} denote

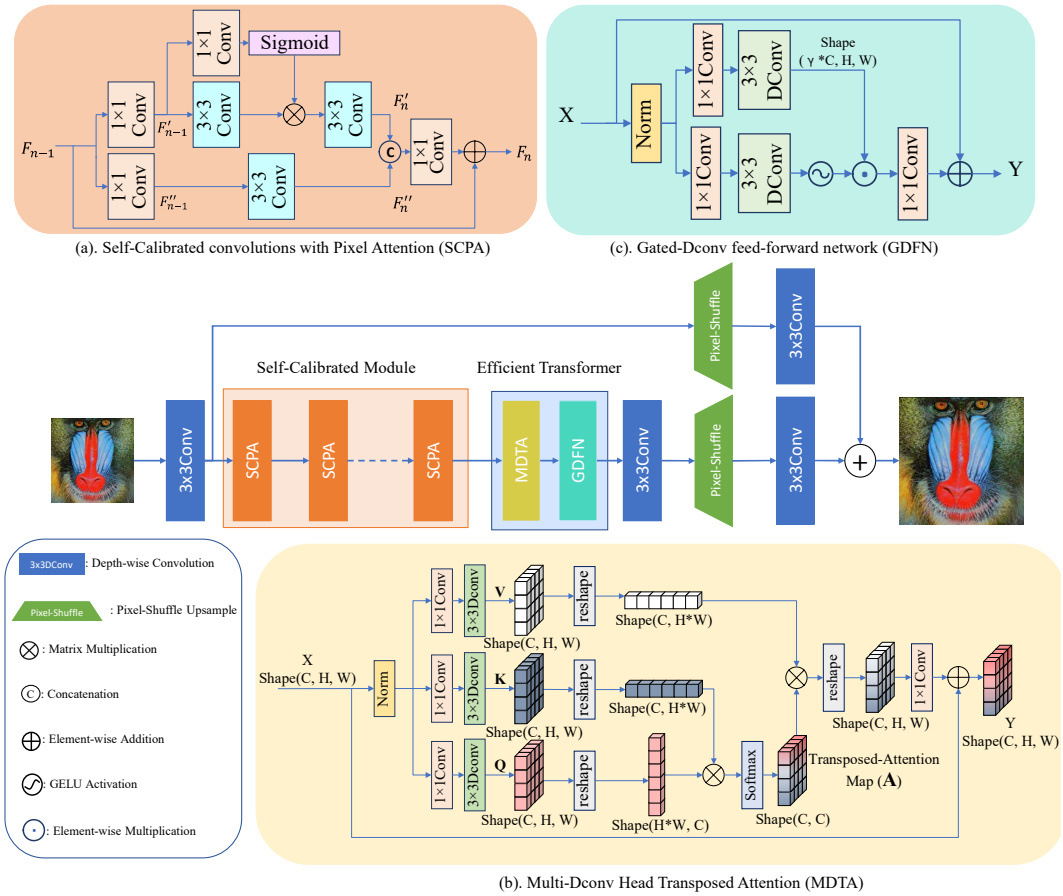

Figure 2. The architecture of self-calibrated efficient transformer (SCET) network. Here, the core modules of network are: (a) Self-calibrated convolution with pixel attention (SCPA), (b) Multi-Dconv head transposed attention (MDTA), and (c) Gated-Dconv feed-forward network (GDFN).

the output of efficient transformer. Finally, we utilize the pixel-shuffle to upsample the features to the HR size. In addition, we added a global residual path to make full use of the shallow feature information. It can be expressed as:

$$I_{SR} = H^1_{up}(F_{out}) + H^2_{up}(F_0) = H_{SCET}(I_{LR}), \quad (4)$$

where H^1_{up} and H^2_{up} denote the upsampling operation of the backbone network and the upsampling operation of the global residual path. H_{SCET} denotes the proposed SCET network. I_{SR} denotes the final restored image.

3.2. Self-Calibrated Module

Most CNN-based lightweight SR networks extract hierarchical features step-by-step to reduce parameters and computational effort, making the insufficient use of low-frequency information resulting in poor image recovery. We employ the SC module constructed from SCPA for feature extraction and recovery. Instead of the step-by-step approach, the SC module allows the network to purposefully recover missing textures through pixel attention. As depicted in Figure 2, our SC module consists of several SCPA blocks. It can be expressed as:

$$F_{out} = H^n_{SCPA}(H^{n-1}_{SCPA}(\cdots H^0_{SCPA}(F_{in})\cdots)), \quad (5)$$

where H^n_{SCPA} denotes the function of the n-th SCPA blocks. F_{in} and F_{out} denote the input and output of the SC module, respectively. Next, we describe specifically the SCPA block in the SC module, as shown in Figure 2 (a). We define F_{n-1} and F_n as the input and output of the n-th SCPA blocks, respectively. The SCPA block consists of two branches, one for the computation of pixel attention information and the other for the recovery of spatial domain information directly. Specifically, the SCPA block first uses pixel convolution of the two branches to reduce the half number of channels. It can be written as:

$$F'_{n-1} = H^1_{pconv}(F_{n-1}), \quad (6)$$
$$F''_{n-1} = H^2_{pconv}(F_{n-1}), \quad (7)$$

where H^1_{pconv} and H^2_{pconv} denote the pixel convolution of upper and lower branch, respectively. F'_{n-1} and F''_{n-1} only have half of the channel number of F_{n-1}. Then, the upper branch computes the attention information by a pixel

attention, and the lower channel branch through a 3×3 convolution to recover the spatial domain information. It can be expressed as:

$$F_{PA} = H_{conv}(F'_{n-1}) \odot \sigma(H_{pconv}(F'_{n-1})), \quad (8)$$
$$F'_n = H^1_{conv}(F_{PA}), \quad (9)$$
$$F''_n = H^2_{conv}(F''_{n-1}), \quad (10)$$

where σ and $\odot$ denote the function of sigmoid and element-wise multiplication, respectively. F_{PA} denotes the pixel attention map. Finally, the output features of the two branches are concatenated together, and then the attention information and spatial domain information are fused together by a pixel convolution to recover the missing texture information in a targeted manner. It can be expressed as:

$$F_n = H_{pconv}(concat(F'_n, F''_n)) + F_n, \quad (11)$$

where $concat$ denotes the operation of concatenation. In order to accelerate training, local residual path is used to produce the final output feature F_n.

3.3. Efficient Transformer

To further improve the performance of our network, we use the efficient transformer module to obtain global contextual information, allowing the network to recover more high frequency texture details. Our efficient transformer consists of MDTA and GDFN. Next, we introduce each module in the efficient transformer in detail.

The major computational overhead in the Transformer lies in the self-attention layer and tends to grow quadratically with the input size. To alleviate this problem, we employ MDTA to compute the cross-covariance over the channel dimensions, as shown in Figure 2 (b). Specifically, we use pixel convolution and depth-wise convolution in three branches to generate query (**Q**), key (**K**) and value (**V**) from the input features $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$. It can be expressed as:

$$\mathbf{Q} = H^1_{dconv}(H^1_{pconv}(LN(\mathbf{X}))), \quad (12)$$
$$\mathbf{K} = H^2_{dconv}(H^2_{pconv}(LN(\mathbf{X}))), \quad (13)$$
$$\mathbf{V} = H^3_{dconv}(H^3_{pconv}(LN(\mathbf{X}))), \quad (14)$$

where H_{dconv}, H_{pconv} and LN denote depth-wise convolution, pixel convolution, and the layer normalization, respectively. Then, we apply the reshape operation to obtain $\hat{\mathbf{Q}} \in \mathbb{R}^{HW\times C}$, $\hat{\mathbf{K}} \in \mathbb{R}^{C\times HW}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{HW\times C}$. Next, their dot-product interaction generates a transposed-attention map **A** of size $\mathbb{R}^{C\times C}$. It can be defined as:

$$\mathbf{A} = \mathbf{V} \cdot Softmax(\mathbf{K} \cdot \mathbf{Q}/\alpha), \quad (15)$$
$$\mathbf{Y} = H_{pconv}(\mathbf{A}) + \mathbf{X}, \quad (16)$$

where $Softmax$ denotes the function of softmax to generate probability map. α is a learnable scaling parameters to control the magnitude of the dot product of **K** and **Q**. Unlike the existing Transformer which calculates self-attention on the spatial domain, MDTA can effectively reduce the amount of computation.

To further recover the accurate structural information, we also adopt the gated-Dconv feed-forward Network. Instead of the feed-forward network in the existing Transformer, GDFN has more operational operations to help the network focus on recovering high frequency details using contextual information, as shown in Figure 2 (c). Given the input feature $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, GDFN can be formulated as:

$$\mathbf{X}^1_G = \phi(H_{dconv}(H_{pconv}(LN(\mathbf{X})))), \quad (17)$$
$$\mathbf{X}^2_G = H_{dconv}(H_{pconv}(LN(\mathbf{X}))), \quad (18)$$
$$\mathbf{Y}_G = \mathbf{X}^1_G \odot \mathbf{X}^2_G, \quad (19)$$
$$\mathbf{Y} = H_{pconv}(\mathbf{Y}_G), \quad (20)$$

where LN and ϕ denote layer normalization and the function of GELU. GDFN controls the information flow through the respective hierarchical levels in our method, thereby allowing each level to focus on the fine details complimentary to the other levels.

Overall, our efficient transformer effectively helps the network to obtain global contextual information to recover high frequency texture details.

3.4. Loss Function

Our SCET is optimized with mean absolute error (MAE, also known as L1) loss function for a fair comparison. Given a training set $\{I^i_{LR}, I^i_{HR}\}$, that contains $\mathcal{N}$ LR inputs and their HR counterparts. The goal of training SCET is to minimize the L_1 loss function:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||H_{SCET}(I^i_{LR}) - I^i_{HR}||_1, \quad (21)$$

where Θ denotes the parameter set of SCET and $||\cdot||_1$ is L_1 norm. The loss function is optimized by using stochastic gradient descent (SGD) algorithm. More training details of our method are presented in Section 4.

4. Experiments

4.1. Settings

In this subsection, we clarify the experimental setting about datasets, degradation models, evaluation metrics, and training settings.

Dataset. Following the previous methods [18, 19, 28, 42, 51], we conduct the training process on a widely used dataset, DIV2K [36] and Flickr2K [37], which contains

Table 1. Average PSNR/SSIM for scale factor $\times 2$, $\times 3$ and $\times 4$ on datasets Set5, Set14, B100, Urban100, and Manga109. Best and second best results are red and blue

Method	Scale	Params	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	×2	-	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN [11]		8K	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
VDSR [20]		666K	37.53/0.9587	33.03/0.9124	31.90/0.8960	30.76/0.9140	37.22/0.9750
DRRN [34]		298K	37.74/0.9591	33.23/0.9136	32.05/0.8973	31.23/0.9188	37.88/0.9749
DRCN [21]		1,774K	37.63/0.9588	33.04/0.9118	31.85/0.8942	30.75/0.9133	37.55/0.9732
IDN [19]		553K	37.83/0.9600	33.30/0.9148	32.08/0.8985	31.27/0.9196	38.01/0.9749
CARN [2]		1,592K	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
IMDN [18]		694K	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
PAN [51]		261K	38.00/0.9605	33.59/0.9181	32.18/0.8997	32.01/0.9273	38.70/0.9773
RFDN [28]		534K	38.05/0.9606	33.68/0.9184	32.16/0.8994	32.12/0.9278	38.88/0.9773
A^2F-M [42]		999K	38.04/0.9607	33.67/0.9184	32.18/0.8996	32.27/0.9294	38.87/0.9774
SCET (Ours)		683K	38.06/0.9615	33.78/0.9198	32.24/0.9006	32.38/0.9299	39.86/0.9821
Bicubic	×3	-	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN [11]		8K	32.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
VDSR [20]		666K	33.66/0.9213	29.77/0.8314	28.82/0.7976	27.14/0.8279	32.01/0.9340
DRCN [21]		1,774K	33.82/0.9226	29.76/0.8311	28.80/0.7963	27.15/0.8276	32.24/0.9343
DRRN [34]		298K	34.03/0.9244	29.96/0.8349	28.95/0.8004	27.53/0.8378	32.71/0.9379
IDN [19]		553K	34.11/0.9253	29.99/0.8354	28.95/0.8013	27.42/0.8359	32.71/0.9381
CARN [2]		1,592K	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.50/0.9440
IMDN [18]		703K	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
PAN [51]		261K	34.40/0.9271	30.36/0.8423	29.11/0.8050	28.11/0.8511	33.61/0.9448
RFDN [28]		541K	34.41/0.9273	30.34/0.8420	29.09/0.8050	28.21/0.8525	33.67/0.9449
A^2F-M [42]		1003K	34.50/0.9278	30.39/0.8427	29.11/0.8054	28.28/0.8546	33.66/0.9453
SCET (Ours)		683K	34.53/0.9278	30.43/0.8441	29.17/0.8075	28.38/0.8559	34.29/0.9503
Bicubic	×4	-	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [11]		8K	30.48/0.8626	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
VDSR [20]		666K	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524	28.83/0.8870
DRCN [21]		1,774K	31.53/0.8854	28.02/0.7670	27.23/0.7233	25.14/0.7510	28.93/0.8854
DRRN [34]		298K	31.68/0.8888	28.21/0.7720	27.38/0.7284	25.44/0.7638	29.45/0.8946
IDN [19]		553K	31.82/0.8903	28.25/0.7730	27.41/0.7297	25.41/0.7632	29.41/0.8942
CARN [2]		1,592K	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.47/0.9084
IMDN [18]		715K	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
PAN [51]		272K	32.13/0.8948	28.61/0.7822	27.59/0.7363	26.11/0.7854	30.51/0.9095
RFDN [28]		550K	32.24/0.8952	28.61/0.7819	27.57/0.7360	26.11/0.7858	30.58/0.9089
A^2F-M [42]		1010K	32.28/0.8955	28.62/0.7828	27.58/0.7364	26.17/0.7892	30.57/0.9100
SCET (Ours)		683K	32.27/0.8963	28.72/0.7847	27.67/0.7390	26.33/0.7915	31.10/0.9155

3450 LR-HR RGB image pairs. We augment the training data with random horizontal flips and rotations. For testing, we use five standard benchmark datasets: Set5 [3], Set14 [46], B100 [30], Urban100 [17], and Manga109 [31].

Degradation models. We downscale HR images with the scaling factors (×2, ×3, and ×4) using Bicubic degradation models [48, 49].

Evaluation metrics. The SR images are evaluated with PSNR and SSIM [52] on Y channel of transformed YCbCr space. Besides, we use Multi-Adds (the size of a query image is 1280×720) and model parameters to evaluate the computational complexity of a model.

Training Settings. We give the implementation details of the proposed SCET. The numbers of the SCPA blocks and feature channels in the self-calibrated module are flexible and configurable, which set 16 and 64, respectively. During training, We train our model SCET on the crop training dataset with LR and HR, the ground turth patch size is random crop into 416×416. We use the Adam [22] optimizer with the 2×10^{-4} learning rate to training 1,000,000 iteration and decay the learning rate with the cosine strategy. Weight decay is 10^{-4} for all the training periodic. We implement our model on the PyTorch platform. Training the SCET roughly takes two days with one RTX2080Ti GPU for the whole training.

4.2. Comparisons with State-of-the-art Methods

Results with Bicubic degaradation. It is widely used to simulate LR images with Bicubic degradation in image SR settings. To verify the effectiveness of our SCET, we compare SCET with 10 SOTA image SR methods: SRCNN [11], VDSR [20], DRCN [21], DRRN [34], IDN [19], CARN [2], IMDN [18], PAN [51], RFDN [28], and A^2F-M [42]. All the quantitative results for various scaling factors are reported in Table 1. Compared with other methods, our SCET, with fewer parameters and computation complexity, performs the best results on five datasets with various scaling factors.

Visual Results of Recent Methods. To further illustrate the superiority of SCET, we also show the visual results of various methods (Bicubic upsampling, SRCNN [11], VDSR [20], CARN [2], IDN [19], IMDN [18], PAN [51], RFDN [28], and our SCET) in Figure 3. We can see that most baseline models cannot reconstruct the lattices accu-

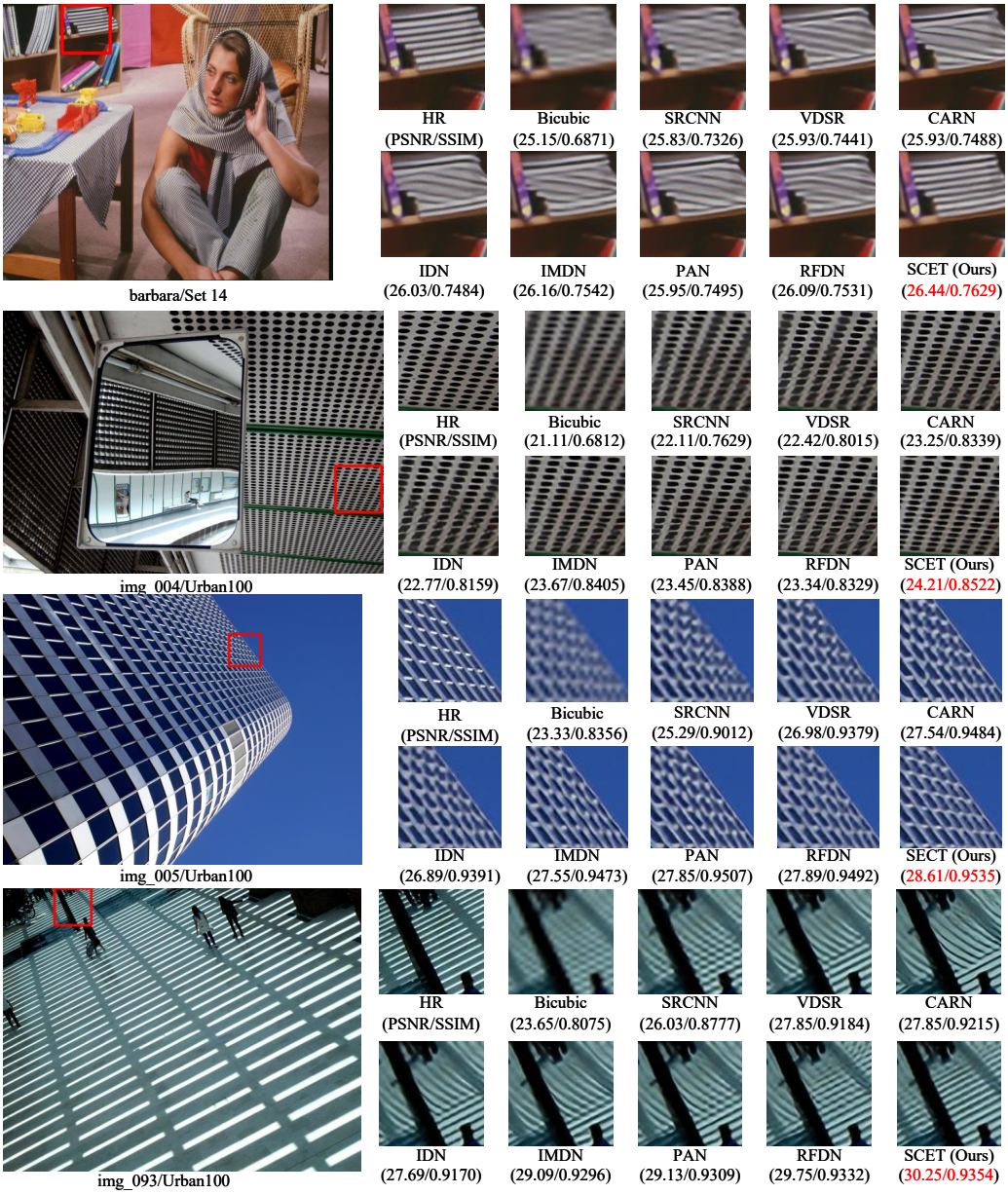

Figure 3. Qualitative comparison with the leading algorithms: SRCNN [11], VDSR [20], CARN [2], IDN [19], IMDN [18], PAN [51], and RFDN [28] on $\times 4$ task. From the figure, we can see that our method can generate finer details of the image and achieve outstanding performance.

rately and thus suffer from serious aliasing. In contrast, our SCET obtains sharper results and recovers more high-frequency details. Take the image img_093/Urban100 for example, most compared methods output heavy aliasing. The early developed methods, i.e., Bicubic upsampling, SRCNN [11], VDSR [20] and CARN [2] lose most of the structure due to the limited network depth and abundant inefficient features. More recent methods, such as IDN [19], IMDN [18], PAN [51], and RFDN [28], can recover the main outlines but fail to recover shaper details. Compared with that, our SCET can restore more details and sharper edges and gain higher visual quality. That should be attributed to more efficient feature extraction and the ability to access global information.

Model Complexity. To further prove the ascendency of SCET in terms of complexity, we compare performance in the matter of parameters and computational complexity. As shown in Figure 1, SCET with limited operations and performance, achieves better performance than other large models. This shows that SCET has a good balance between model complexity and performance.

4.3. Ablation Study

In this subsection, we design a series of ablation experiments to analyze the effectiveness of each of the modules we propose. We use the DIV2K validation dataset for eval-

Table 2. Model Policy with deep and wide on network performance. The 'd' denotes the number of SCPA blocks. The 'w' denotes the number of feature channels.

Model	Params	Multi-Adds	PSNR	SSIM
d = 8, w = 32	98k	11.46G	28.32	0.7741
d = 8, w = 64	388k	44.85G	28.64	0.7894
d = 16, w = 32	172k	19.9G	28.58	0.7869
d = 16, w = 64	683k	78.72G	**28.72**	**0.8158**

Table 3. Ablation studies of different backbone. We report the PSNR (dB) values on DIV2K validation datasets (×4).

Backbone	Params	Multi-Adds	PSNR	SSIM
ResBlock	1274k	146.87G	28.29	0.7965
RCAB	1284k	146.87G	28.32	0.7984
IMDB	920k	106.05G	28.49	0.8033
RFDB	1336k	145.9G	28.57	0.8042
SCPA	683k	78.72G	28.72	0.8158

Table 4. Ablation studies of different transformer. We report the PSNR (dB) values on DIV2K validation datasets (×4).

Transformer	Component	Params	Multi-Adds	PSNR
Baseline	SCPA blocks	629K	72.59G	28.54
Self-Attention	MTA+FN	1002K	129.65G	28.62
	MDTA+FN	929K	107.09G	28.69
Feed-forward Network	MDTA+Resblock	721K	83.14G	28.59
	MDTA+RCAB	722K	84.21G	28.61
Overall	MDTA+GDFN	683K	78.72G	28.72

uation and performed 1,000,000 iterations of training on an input image patch of size 32 × 32.

Model Design Policy. We explore the impact of different depths and widths on network performance, as shown in Table 2. The depth represents the number of SPCA blocks and the width represents the numberof channels in our intermediate features. As can be seen from the experimental results, the width affects network performance and parameters more than the depth. Our model works best at d = 16 and w = 64. Therefore, our final model is set to d = 16 and w = 64.

Comparison of different backbone schemes. To illustrate the effectiveness of the SCPA as a backbone, we used the residual block, residual channel attention block (RCAB), information multi-distillation block (IMDB) and residual feature distillation block (RFDB) to replace the original SCPA blocks for the ablation experiments.

In Table 3, we give the comparison in terms of parameters, Multi-Adds, and the performance in PSNR. Note that all results are the mean values of PSNR calculated by 100 images on DIV2K validation dataset. Mult-Adds is computed by assuming that the resolution of HR image is 720p. It is observed that SCPA could achieve the best performance with the fewest parameters and Multi-Adds. SCPA can reduce parameters and calculations by nearly half in comparison to RFDB, obtaining a performance improvement of 0.15dB. This indicates that SCPA is more effective than traditional basic modules which employ a step-by-step approach to extract hierarchical features.

Comparison of different Transformer schemes. To illustrate the effectiveness of MDTA and GDFN in efficient transformer, we compare the effects of different approaches to self-attention and different feed-forward networks on the model. Note that our baseline model is set up as a residual network of cascading multiple SCPA blocks.

As shown in Table 4, it demonstrates that the MDTA provides favorable gain of 0.18 dB over the baseline. The MDTA can reduce the amount of computation by 20% compared to traditional self-attention. Moreover, it is shown that deep convolution can effectively improve the robustness of the efficient transformer. For feedback networks, the gating mechanism in GDFN that controls the information flowing can effectively help the network to obtain better performance. Compared to other feedforward network designs, the GDFN can improve performance by about 0.1 dB.

5. Conclusion

In this paper, we propose a lightweight SCET network for efficient super-resolution. In particular, we design a new Efficient Transfomer framework, which effectively combines the efficient pixel attention mechanism with the transformer to achieves excellent results with few parameters. Additionally, numerous experiments have shown that the proposed method achieves a commendable balance between visual quality and parameters amount, which are the vital factors that affect practical use of SISR.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China under Grant No. 61901117, U1805262, 61971165, in part by the Natural Science Foundation of Fujian Province under Grant No. 2019J05060, 2019J01271, in part by the Special Fund for Marine Economic Development of Fujian Province under Grant No. ZHHY-2020-3, in part by the research program of Fujian Province under Grant No. 2018H6007, the Special Funds of the Central Government Guiding Local Science and Technology Development under Grant No. 2017L3009, and the National Key Research and Development Program of China under Grant No. 2016YFB1001001.

References

[1] Chunwei Tian A, Yong Xu A B, and Wangmeng Zuo C. Image denoising using deep cnn with batch renormalization. *Neural Networks*, 121:461–473, 2020. 2

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. 03 2018. 3, 6, 7

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 6

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 3

[6] Xiangxiang Chu, Bo Zhang, Hailong Ma, Ruijun Xu, and Qingyuan Li. Fast, accurate and lightweight super-resolution with neural architecture search. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 59–64. IEEE, 2021. 2, 3

[7] Xiangxiang Chu, Bo Zhang, and Ruijun Xu. Multi-objective reinforced evolution in mobile neural architecture search. In *European Conference on Computer Vision*, pages 99–113. Springer, 2020. 2, 3

[8] Xiangxiang Chu, Bo Zhang, and Ruijun Xu. Multi-objective reinforced evolution in mobile neural architecture search. In *European Conference on Computer Vision*, pages 99–113. Springer, 2020. 3

[9] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11057–11066, 2019. 3

[10] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997, 2021. 3

[11] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 1, 2, 6, 7

[12] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. 08 2016. 1, 2

[13] 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. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *International journal of computer vision*, 40(1):25–47, 2000. 1

[15] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong. Image super-resolution using knowledge distillation. In *Asian Conference on Computer Vision*, pages 527–541. Springer, 2018. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 1

[17] J. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 6

[18] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th acm international conference on multimedia*, pages 2024–2032, 2019. 2, 3, 5, 6, 7

[19] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 723–731, 2018. 2, 3, 5, 6, 7

[20] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 1, 2, 6, 7

[21] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016. 1, 2, 3, 6

[22] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer ence*, 2014. 6

[23] W. Lai, J. Huang, N. Ahuja, and M. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5835–5843, 2017. 2, 3

[24] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 3

[25] Yawei Li, Kai Zhang, Luc Van Gool, Radu Timofte, et al. Ntire 2022 challenge on efficient super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 2

[26] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 3

[27] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 1, 2

[28] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In

European Conference on Computer Vision, pages 41–55. Springer, 2020. 3, 5, 6, 7

[29] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10096–10105, 2020. 3

[30] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, 2001. 6

[31] Yusuke Matsui, Kota Ito, Yuji Aramaki, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. 2015. 2, 6

[32] J. Pan, W. Ren, Z. Hu, and M. Yang. Learning to deblur images with exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1412–1425, 2019. 2

[33] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 2

[34] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2798, 2017. 1, 2, 3, 6

[35] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4549–4557, 2017. 2

[36] R. Timofte, E. Agustsson, L. V. Gool, M. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1110–1121, 2017. 5

[37] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 5

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[39] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. *arXiv preprint arXiv:2111.14813*, 2021. 3

[40] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 3

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[42] Xuehui Wang, Qing Wang, Yuzhi Zhao, Junchi Yan, Lei Fan, and Long Chen. Lightweight single-image super-resolution network with attentive auxiliary feature learning. In *Proceedings of the Asian conference on computer vision*, 2020. 2, 5, 6

[43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[44] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 3

[45] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021. 3

[46] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. 2010. 6

[47] Kai Zhang, Martin Danelljan, Yawei Li, Radu Timofte, Jie Liu, Jie Tang, Gangshan Wu, Yu Zhu, Xiangyu He, Wenjie Xu, et al. Aim 2020 challenge on efficient super-resolution: Methods and results. In *European Conference on Computer Vision*, pages 5–40. Springer, 2020. 3

[48] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. 6

[49] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3262–3271, 2018. 6

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. 2018. 1, 3

[51] Hengyuan Zhao, Xiangtao Kong, Jingwen He, Yu Qiao, and Chao Dong. Efficient image super-resolution using pixel attention. In *European Conference on Computer Vision*, pages 56–72. Springer, 2020. 3, 5, 6, 7

[52] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[53] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016. 3