IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 October 2023, accepted 22 October 2023, date of publication 27 October 2023, date of current version 3 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3328229

RESEARCH ARTICLE

SRFormer: Efficient Yet Powerful Transformer Network for Single Image Super Resolution

ARMIN MEHRI[1,2], PARICHEHR BEHJATI[1], DARIO CARPIO[3], AND ANGEL DOMINGO SAPPA[1,3], (Senior Member, IEEE)

[1]Computer Vision Center, Autonomous University of Barcelona, 08193 Barcelona, Spain
[2]Camaleonic Analytics SL, 08014 Barcelona, Spain
[3]ESPOL Polytechnic University, Guayaquil EC090112, Ecuador

Corresponding authors: Armin Mehri (amehri@cvc.uab.es) and Angel Domingo Sappa (asappa@espol.edu.ec)

This work was supported in part by the Air Force Office of Scientific Research under Award FA9550-22-1-0261, in part by the Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ''ERDF A way of making Europe'', and in part by the ESPOL Project CIDIS-12-2022. The authors acknowledge the support of the Generalitat de Catalunya CERCA Program to CVC's general activities, and the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR01499.

ABSTRACT Recent breakthroughs in single image super resolution have investigated the potential of deep Convolutional Neural Networks (CNNs) to improve performance. However, CNNs based models suffer from their limited fields and their inability to adapt to the input content. Recently, Transformer based models were presented, which demonstrated major performance gains in Natural Language Processing and Vision tasks while mitigating the drawbacks of CNNs. Nevertheless, Transformer computational complexity can increase quadratically for high-resolution images, and the fact that it ignores the original structures of the image by converting them to the 1D structure can make it problematic to capture the local context information and adapt it for real-time applications. In this paper, we present, SRFormer, an efficient yet powerful Transformer-based architecture, by making several key designs in the building of Transformer blocks and Transformer layers that allow us to consider the original structure of the image (i.e., 2D structure) while capturing both local and global dependencies without raising computational demands or memory consumption. We also present a Gated Multi-Layer Perceptron (MLP) Feature Fusion module to aggregate the features of different stages of Transformer blocks by focusing on inter-spatial relationships while adding minor computational costs to the network. We have conducted extensive experiments on several super-resolution benchmark datasets to evaluate our approach. SRFormer demonstrates superior performance compared to state-of-the-art methods from both Transformer and Convolutional networks, with an improvement margin of $0.1 \sim 0.53dB$. Furthermore, while SRFormer has almost the same model size, it outperforms SwinIR by 0.47% and inference time by half the time of SwinIR. The code will be available on GitHub.

INDEX TERMS Single image super resolution, transformers, convolutional neural network.

I. INTRODUCTION

The Super Resolution has been studied since 1974, when Gerchberg [1] introduced the notion of Super Resolution (SR) to improve optical system resolution over and above diffraction, since then the idea of super resolution has been defined as a way to obtain high resolution (HR) images from its degraded low resolution (LR) image with high visual quality, more realistic textures and enhanced in details of the given low-resolution input image.

Although super resolution being explored for decades, single image super resolution is still an active yet challenging topic in Computer Vision due to its complex nature and high practical values in improving image details and textures. The recent success of image super resolution has the potential to significantly improve the quality of media content, resulting in better user experiences. For example, the digital zoom algorithm used in mobile cameras and the image enhancement techniques used in digital devices. Furthermore,

The associate editor coordinating the review of this manuscript and approving it for publication was Miaohui Wang.

this core technology can be applied to a wide range of Computer Vision tasks, which leads to improvements in various Vision tasks, such as object detection [2], [3], medical imaging [4], [5], security and surveillance imaging [6], [7], face recognition [8], [9].

There are several reasons that make image super resolution remains challenging: *i*) Super Resolution is fundamentally an ill-posed inverse problem. There are multiple solutions for the same low-quality image instead of a unique single solution. *ii*) The complexity of the problem increases, as the up-scale factor increases. The retrieval of missing scene details becomes even more complicated with greater factors, which often leads to the reproduction of incorrect information; and *iii*) there are fundamental uncertainties among the LR and HR data since the down-sampling of different HR images may lead to a similar LR image [10].

Formerly, different methods were utilized to tackle super resolution problems, such as statistical methods, prediction-based methods, patching methods, edge-based methods, and sparse representation methods. However, researchers have lately been using Deep Learning (DL) approaches to solve the problems of image super resolution due to advanced progress in computer computational power.

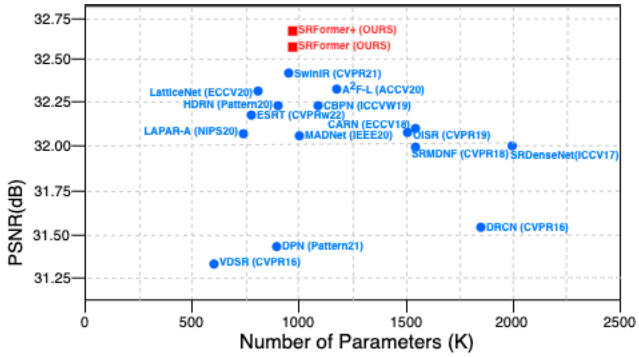


FIGURE 1. PSNR *vs*. Model size trade-off on Urban100 (×4). SRformer achieves superior performance among all the CNN and Transformer networks.

Deep learning ConvNet-based approaches have consistently improved significantly to the classical methods over the last decade. Numerous deep convolutional neural networks introduced [11], [12], [13], [14], [15], [16] as well as many lightweight networks and techniques to reduce the computational complexity of the networks, such as using filter pruning [17], knowledge distillation [18] to minimize computing time by narrowing the network. However, these techniques often lead to poor performance due to several reasons such as lower network capacity, long inference time, and a large number of operations due to several iterations through the forward process.

In addition, ConvNet-based approaches suffer from two main issues that come from the fundamentals of the convolution layer. First, there is no content dependency in the interactions between images and convolution kernels. The same convolution kernel is used to restore various image regions, which is not the ideal solution. Second, convolution is effective for capturing local context information but ineffective for capturing long-range dependency [19].

Transformer [20] introduced to tackle the aforementioned problems of convolution layer, by designing a self-attention mechanism to capture global interactions between contexts, has shown promising performance in several Vision and NLP tasks [21], [22], [23]. However, the self-attention mechanism computational cost increases quadratically when dealing with spatial resolution and also ignores the local 2D structure information of the image by processing images as a 1D structure [24]. Furthermore, these methods usually need to occupy heavy GPU memory, which greatly limits their flexibility and application scenarios for low-capacity devices.

In this paper, we propose a novel lightweight approach for a single image super resolution task, namely SRFormer by bringing the strengths of both the convolution layer and Transformer layer together to address the aforementioned problems. By advancing both Convolution and Transformer together, SRFormer is able to capture both local context information and global interactions between contexts while staying computationally efficient. The combination of both CNN and Transformer together with the precise design of our SRFormer architecture, allows our model to perform exceptionally well on benchmark datasets with faster training and inference times compared to other Transformer based networks. It is worth mentioning that, SRFormer trained with only a single GPU for 3 days, while SwinIR trained on 8 GPUs for almost 2 days to achieve their results. Also, SRFormer has the advantage of multi-scale training, which can generate SR images with different scale factors [×2, ×3, ×4] in one training phase, while other methods need to train separately for each scale factor. As illustrated in Fig. 1, the proposed SRFormer yields to **21%** improvement on average of all benchmark datasets for scale factor 4 when compared to the SwinIR [19]– SOTA Transformer-based model, which shows the efficiency of the proposed model. The extension of this work on cross-spectrum applications can be found at [25].

The main contributions of our work can be summarized as follows:

- We present SRFormer, an efficient yet powerful Transformer based network for single image super resolution task, which is faster in training and inference time while generating more accurate SR images.
- We present a lightweight Dual Attention layer, which significantly improves the reconstruction quality by generating a global attention map from two local attention weights, which obtain individually by two branches in parallel while it's not memory hunger.
- We present a low-cost Gated MLP Feature Fusion module that yields a powerful representation by aggregating multi-stage feature representation from Transformer blocks with minor computation complexity.
- Extensive experiments show that SRFormer achieves state-of-the-art on various benchmark datasets for Single

Image Super Resolution (SISR) tasks compared to CNN/Transformer based networks.

The rest of the paper is organized as follows: Section II discusses the related work, including CNN- and Transformer-based super resolution methods. Section III describes the proposed SRFormer and its core components in detail. Experimental comparisons against several state-of-the-art methods are presented in Section IV. The model investigation presents in section V. Section VI concludes the paper.

II. RELATED WORK

In this section, the most recent state-of-the-art SR deep learning CNN and Transformer based approaches are detailed.

A. DEEP LEARNING BASED IMAGE SUPER-RESOLUTION

Single Image Super Resolution aims to restore the well-detailed image from its low-quality version. Dong et al. [10] introduced Super-Resolution Convolutional Neural Network (SRCNN), which is the first work using CNN to tackle the SR task. The SRCNN presents a shallow neural network that receives an upsampled image as an input that cost extra computation. Later on, to address this drawback, Fast Super-Resolution Convolutional Neural Network (FSRCNN) [26] and Efficient Sub-Pixel Convolutional Neural network (ESPCN) [27] have been proposed by receiving the LR image as input to reduce the large computational and run-time cost and upsampling the features near the output of the network by a single transposed convolution layer. Even though the strength of deep learning shows up from deep layers, the above-mentioned methods are referred to as shallow networks. Therefore, Kim et al. [28] use residual learning to ease the training challenges and increase the depth of their network by adding 20 convolutional layers. Then, [29] proposed a memory block in MemNet for deeper networks and solves the problem of long-term dependency with 84 layers. Lim et al. [30] introduce Enhanced Deep Super-Resolution network (EDSR) by expanding the network size and enhancing the residual block by omitting the batch normalization from a residual block. Zhang et al. [31] propose Residual Dense Network (RDN) with residual and dense skip connections to fully use hierarchical features.

Furthermore, in recent years the interest in building lightweight and efficient models has increased in super resolution tasks to reduce the high computational cost of this task. Ahn et al. [32] design an efficient network that is suitable for the mobile scenario. Later, [33] introduces Multi-Attentive Feature Fusion Super-Resolution Network (MAFFSRN) by proposing multi-attention blocks to improve the performance. LatticeNet [34] introduces an economical structure to adaptively combine Residual Blocks. Recently, OverNet [16] introduced by designing an efficient network structure by introducing a multi-loss function to boost the network performance. Also, a neural architecture search (NAS)-based strategy has been also proposed in SISR to construct efficient networks—Multi-Objective Reinforced Evolution in Mobile Neural Architecture Search (MoreMNAS) [35] and Fast, Accurate and Lightweight Super-Resolution (FALSR) [36] are some examples of using NAS strategy in their network. However, due to the limitation in NAS strategy, the performance of these models is limited.

B. VISION TRANSFORMER

Transformer networks show breakthrough performance in the Natural Language Process (NLP). In contrast to ConvNets, Transformer networks have the advantage of capturing long-range dependency in the input with global self-attention. The core idea of the Transformer is the self-attention module, which is capable of capturing long-term information between sequence elements.

The impressive performance Transformer based network in the NLP domain inspires the Computer Vision community to adopt the Transformer for Vision tasks. The first work in this direction has been done by Alex et al. who propose ViT [22] as a Vision Transformer, which replaces the standard CNN with Transformer and directly trains on the medium-size flattened patches with large-scale data pre-training.

Since introducing the first work, many Transformer based architectures have been proposed for the Vision tasks in image recognition [37], object detection [21], [23], segmentation [38], [39], and action recognition [40], [41]. In addition, Transformer based models have been studied for low-level vision problems such as super resolution [19], [42], [43], image colorization [44], denoising [45], and image restoration [46]. For instance, DEtection TRansformer (DETR) [23] is a transformer network designed for object detection, which can predict a set of objects and model their relationships. SwinIR [19] was introduced by Jingyun et al. for low-level vision tasks by using Swin Transformer [21] by applying self-attention within local image regions to solve the low-level vision problems.

Although the Transformer based networks achieve excellent performance in low-level Vision tasks, these methods still depend on providing heavy GPU resources to train the model, which is not feasible or available to most researchers. Also, the computational complexity of self-attention in Transformers can increase quadratically with the number of tokens to mix (i.e., image patches), thereby prohibiting its application to high-resolution images. Therefore, in contrast to recent work in the super resolution domain, we present a Transformer based network that can learn long-range dependency and local context information while remaining computationally efficient without the need for heavy GPU resources.

III. PROPOSED METHOD

In this section, the overall network architecture of the proposed SRFormer is described. Later, detailed information on the Dual Attention layer is provided.

A. OVERALL PIPELINE

The primary goal is to design an efficient Transformer architecture, which can generate well-detailed high-quality

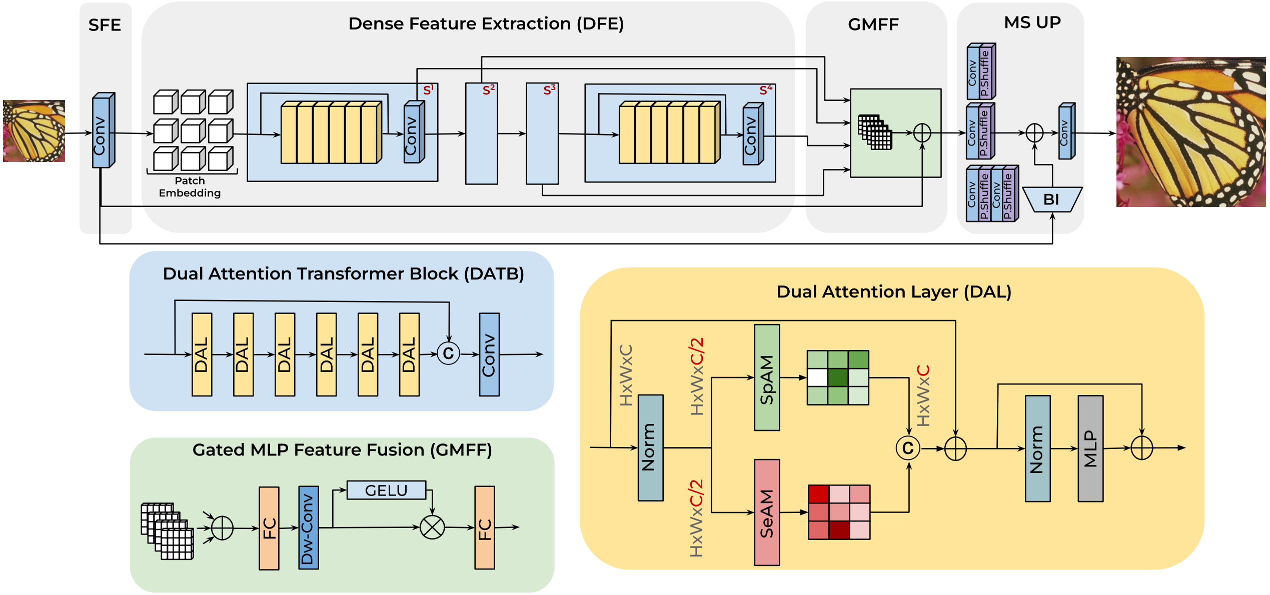


FIGURE 2. The overall network architecture of the proposed SRFormer.

images while remaining computationally efficient. Thus, we utilize the basic Transformer structure but specially designed for efficient network structure with significant performance gains compared to existing CNN and Transformer networks. The overall architecture of the SRFormer is shown in Fig.2. In particular, the proposed SRFormer consists of four modules: Shallow Feature Extraction (SFE), Dense Feature Extraction (DFE), Gated MLP Feature Fusion (GMFF), and Multi-Scale Up-Sampling (MS-UP) modules. We defined I_{LR} and I_{SR} as the low- and high-quality input and output of our network.

B. SHALLOW FEATURE EXTRACTION

The convolution layer proves that can perform well at early visual processing, which leads to improved performance of the network [47]. Therefore, a single 3 × 3 convolutional layer is applied on the given low-quality input image I_{LR} to extract the initial features and map the input image space to a higher dimensional feature space to generate a better SR image. Therefore, we extract the shallow features F_0 as:

$$\boldsymbol{F}_0 = Conv_{3\times3}(\boldsymbol{I}_{LR}), \tag{1}$$

C. DENSE FEATURE EXTRACTION

Next, the extracted shallow feature passes through the Dense Feature Extraction F_{DFE} as an input. DFE is built up with a set of Transformer blocks. The input is first processed by input embedding such as patch embedding for Vision Transformers (ViTs):

$$\boldsymbol{I}_{EMB} = InputEmb(F_0), \tag{2}$$

where I_{EMB} denotes the embedding tokens with the length of N sequence and C embedding dimension. Our Dense Feature Extraction module takes embedding tokens as input to our Transformer blocks. Specifically, Dense Feature Extraction contains several Transformer blocks, which include i^{th} Transformer layers and a 1 × 1 Conv layer at the end of each block with the benefit of waterfall residual connection to transfer the information from the previous stage to the current stage. The shallow features from the SFE process through different Transformer stages extract more abstract features and spotlight the high-level information (further details provided in section III-G). Thus, we extract the feature as follows:

$$\boldsymbol{F}_{DFE} = H_{DFE}(I_{EMB}), \tag{3}$$

where $H_{DFE}(.)$ is Dense Feature Extraction module with several Transformer blocks, which can be seen as

$$\boldsymbol{F}_i = Conv_{1\times1}(C[H_{DATB}(F_{i-1}), X_{i-1}],\ i = 1, 2, \ldots, K, \tag{4}$$

where $H_{DATB}(.)$ denotes the i_{th} Transformer blocks. C denotes the concatenation operation between the input feature of each *DATB* block and the output. Concatenating a convolutional layer within each stage of the Transformer block helps to transfer inductive bias from the convolution operation into the Transformer-based network and provides a more solid foundation for the later aggregation of shallow and deep features together.

D. GATED MLP FEATURE FUSION

The aim of the Gated MLP Feature Fusion (GMFF) design is to highlight the location information in the stacked feature map of different stages of Transformer blocks. GMFF consists of N stacked residual *DATB* as shown in Fig 2. GMFF first, accumulates the multi-stage features from different Transformer stages to create multi-stage representations of the input image. Then, passes the features through the lightweight MLP network. However, in contrast to a standard MLP network, we propose a novel MLP module by using a 3 × 3 Depthwise Conv layer inside the module to leak the spatial information in order to boost the network performance since highlighting such features are important in super resolution task to achieve high performance. Also, the gating mechanism is used by formulating the element-wise product of two parallel routes of linear transformation layer that one is activated with the GELU [48]. Thus, Gated MLP Feature Fusion can be formulated as follow:

$$\boldsymbol{F}_{GMFF} = MLP(GELU(Conv_{3\times 3}(MLP(F_i)))) + F_0, \quad (5)$$

where F_{GMFF} denotes the output of our feature aggregation of multi-stage Transformer block with the initial features, which is later used by the Multi-Scale Up-Sampling module. In the ablation study, we will show the effectiveness of our proposed Gated MLP Feature Fusion compared to the standard MLP network.

E. MULTI SCALE UP-SAMPLING

Given the feature from previous modules, which contains an aggregation of low- and high-level information, our model generates a high-quality image I_{SR}. Multi-Scale Up-Sampling (MSUP) module takes the features directly from GMFF module to be able to reconstruct the high-quality output. MSUP consists of several convolutional and pixel-shuffle layers to upsample the features to the corresponding sizes in one training phase instead of training for each interested scale factor separately. Furthermore, we incorporate a global connection path H_{UP} with only a bicubic interpolation to grant access to the original LR information and facilitate the back-propagation of the gradients. The Multi Scale Up-Sample module can be formulated as:

$$\boldsymbol{I}_{SR} = H^{\uparrow}_{Rec}(F_0 + F_{GMFF} + H_{UP}(I_{LR})), \quad (6)$$

where $H_{Rec}(\cdot)$ and I_{SR} denote the up-sampling module and high quality reconstructed image respectively:

F. LOSS FUNCTION

To keep the consistency with previous works, we use L_1 loss as a cost function during training to optimize the parameters of the proposed SRFormer.

$$L_1(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \|I_{SR} - I_{HR}\|_1 \,, \quad (7)$$

where I_{SR} is obtained by taking a low-quality image as the input of our model and I_{HR} is the corresponding ground truth.

In the next subsections, more details about our Transformer layer are given.

G. DUAL ATTENTION LAYER

This section presents the proposed Dual Attention layer by completely revising the token mixer (i.e., self-attention). As well known, self-attention is playing an important role to achieve high performance in Natural Language Processing (NLP) and Computer Vision Transformer based networks. However, self-attention can be problematic due to several reasons, especially when it comes to working with spatial resolution, which involves high-resolution images. The computational complexity of self-attention increases quadratically to the number of tokens to mix. Besides that, self-attention treats images as flattened sequences, which neglects the original structure of images therefore it ignores the adaptability in channel dimensions, which has proven important for visual tasks. Also, self-attention does not take into account the local contextual information due to the nature of self-attention. Thus, we introduce the Dual Attention layer to overcome the aforementioned shortages by generating a global attention map with less computational cost compared to the existing token mixer. Dual Attention generates a global attention map by aggregating two local attention maps, which are separately obtained by using two different branches, CNN-based Attention Module, and Transformer self-attention in parallel. By doing so, unlike the previous token mixer, Dual Attention can also consider both long-range dependency and local contextual information with less computational complexity.

As shown in Fig 2, we design our Dual Attention in a way that it splits the channel features equally for both attention module branches (SpAM and SeAM). From the Norm layer tensor X, both of our branches receive half of the input tensor to create the local attention maps individually. SeAM is a self-attention Transformer, which first generates the query (Q), key (K), and value (V) projections enriched with the local context. We apply SeAM only across channels rather than spatial dimensions. Our SeAM uses only depth-wise convolutions to emphasize the channel-wise spatial context before computing feature covariance to produce the attention map. Thus, Q, K, V are computed as:

$$Q = W_d^Q Y, \;\; K = W_d^K Y, \;\; V = W_d^V Y \quad (8)$$

where $W_d^{(\cdot)}$ is the 3 × 3 bias-free depth-wise convolution. Next, query and key projections reshape in a way that their dot-product interaction generates a transposed attention map. Thus, the attention map generates as follows:

$$Attention(Q, K, V) = W_d(V.Softmax(K.Q/\alpha)) + X \quad (9)$$

where X is the input feature map and α is a learnable scaling parameter that is used to regulate the magnitude of the dot product of K and Q before applying the Softmax function. Similar to previous works [19], [20], [49], we perform the attention function for h times to learn separate attention maps in parallel in our SeAM module.

The second branch of the Dual Attention layer is the Spatial Attention Module (SpAM), which is an almost parameter-free attention mechanism. SpAM receives the other half of the input tensor to generate the local attention map. The goal of the SpAM module is to encode the spatial information, which represents the importance of each pixel in the input feature with a negotiable cost. Given half of the input tensor information, the channels of the input tensor are reduced by mean and max operations, of which the shape is $1 \times H \times W$. The obtained features concatenated and then passed through a convolution layer with a kernel size of 7×7. After, a sigmoid activation layer applies to the output feature to generate the attention weights of shape $1 \times H \times W$ which are later multiplied with the input tensor to refined tensors of shape $C \times H \times W$. Thus, the SpAM can be formulated as follow:

$$X = Sigmoid(Conv_{7\times 7}[F_{Mean}(X), F_{Max}(X)]) * X \qquad (10)$$

where $F_{Mean}(\cdot)$ and $F_{Max}(\cdot)$ denotes for mean and max operations. Later, generated local attention maps from SpAM and SeAM are concatenated together to obtain a unified global attention map with less computational cost. Thus, the generated attention map contains both long-range dependency and local context information with enrich of spatial features.

Following that, a multi-layer perceptron (MLP) with two fully connected layers and a GELU non-linearity activation function between them is employed for further feature modifications. The norm layer is also added before MLP, and both modules contain the residual connection between them. Thus, the entire procedure inside of our Dual Attention is as follows:

$$\begin{aligned} X &= (Norm(SpAM(X/2) + SeAM(X/2))) + X \\ Y &= MLP(Norm(X)) + X \end{aligned} \qquad (11)$$

where $Norm(\cdot)$ stands for the normalization layer and Y for the output feature map.

IV. EXPERIMENTAL RESULTS

A. SETTING

1) DATASETS

following prior works [34] and [70], $DIV2K$ dataset has been used for training and validating the network. $DIV2K$ splits into 800 high-quality images for the training phase, 100 validation images, and 100 test images. SRFormer is trained with all training images and validated with validation image sets. To evaluate the proposed method, five standard benchmark datasets have been used, namely, $Set5$ [50], $Set14$ [51], $B100$ [52], $Urban100$ [53], $Manga109$ [54].

2) EVALUATION PROTOCOL

Two widely used quantitative metrics have been considered to measure the performance of our SRFormer in order to maintain consistency with previous works. Peak Signal-to-Noise Ratio (PSNR) is measured in decibels (dB) and the Structural Similarity index (SSIM), is computed between generated SR images and the corresponding ground truths. Keeping up with the SR community, the RGB reconstruction results are first transformed to $YCbCr$ space, and then just the luminance channel is considered to compute the PSNR and SSIM in our experiments.

3) DEGRADATION MODELS

In order to demonstrate the efficiency of the proposed model, following the work of [31], three different degradation models were created to simulate LR images and make fair comparisons with available methods. Degradation data were obtained as follows: Firstly, a bicubic (BI) down-sampling dataset with scaling factors [×2, ×3, ×4] has been created. Secondly, Blur-Downsampled (BD) has been created by applying Gaussian kernel 7×7, and $\sigma = 1.6$ to HR images and then downsampled images with scaling factor ×3. Aside from the BD, a more challenging degradation model has been created, referred to as Downsample-Noisy (DN). DN degradation model is down-sampling HR images with bicubic followed by adding 30% Gaussian noise.

4) IMPLEMENTATION DETAILS

In the training phase, RGB patches are provided as inputs with the size of 64×64 from each of the randomly selected 32 low-quality training images. Data augmentation is applied on patches by means of horizontal random flips and 90 degree rotation. AdamP [71] optimizer has been employed with the initial learning rate 10^{-3} and its halved every 4×10^5 steps. $L1$ is used as a loss function to optimize the model. Also, the configurations of our transformer encoder are as follows, we used 4 Transformer blocks within 6 Transformer layers for each block, Embedding dimension set to 64, and MLP ratio of 2 for all Transformer blocks. Also, a Conv1×1 is used inside each Transformer block. SRFormer was developed by using the PyTorch framework and trained on a single NVIDIA RTX 3090 GPU to achieve its performance.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, SRFormer and SRFormer+ are compared to other lightweight state-of-the-art SR methods. Self-ensemble method [72] is also used to further boost the performance of the proposed SRFormer (denoted as SRFormer+).

1) RESULTS ON BICUBIC DEGRADATION

We present comparisons between the proposed method (SRFormer and SRFormer+) and several of the most recent lightweight SOTA CNN and Transformer based models: VDSR [28], DRCN [55], CARN [32], CBPN [56], FALSR [57], LAPAR-A [59], LatticeNet [61], MADNet [62], HDRN [63], DPN [64], A^{2}F [65], ESRT [42], and SwinIR [19] on the Bicubic (BI) degradation model for scale factors [×2, ×3, ×4]. Also, the number of network parameters and Multi-Adds operations are presented in Table 1 to demonstrate the complexity of the model and have a fair comparison with the existing methods. As can be seen, SRFormer produces superior outcomes in practically all circumstances when compared to the other methods mentioned above. This shows that SRFormer is capable of continuously

TABLE 1. **Average PSNR/SSIM comparison with state-of-the-art CNN- and Transformer-based methods with the same range of network parameters on the Bicubic (BI) degradation for scale factors [×2, ×3, ×4] (Transformer based methods separated with horizontal line). Red is the Best and Blue is the second best performance. We assume that the generated SR image is *720P* to calculate Multi-Adds (MAC). SRFormer with self-ensemble results are Highlighted.**

Scale	Method	Params	FLOPs	Set5 [50]		Set14 [51]		B100 [52]		Urban100 [53]		Manga109 [54]	
				PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
×2	VDSR [28]	665K	613G	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140	37.22	0.9750
	DRCN [55]	1,774K	17,974G	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140	37.22	0.9750
	CARN [32]	1,592K	223G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
	CBPN [56]	1,036K	240.7G	37.90	0.9590	33.60	0.9171	32.17	0.8989	32.14	0.9279	–	–
	FALSR-A [57]	1,021K	234.7G	37.82	0.9595	33.55	0.9168	32.12	0.8987	31.93	0.9256	–	–
	SRMDNF [58]	1,513K	348G	37.79	0.9600	33.32	0.9150	32.05	0.8980	31.33	0.9200	–	–
	LAPAR-A [59]	548K	171G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
	OISR-LF-s [60]	1,370K	316.2G	38.02	0.9605	33.69	0.9178	32.20	0.9000	32.21	0.9290	–	–
	LatticeNet [61]	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.24	0.9302	–	–
	MADNet [62]	878K	187.1G	37.94	0.9604	33.46	0.9167	32.10	0.8988	31.74	0.9246	–	–
	HDRN [63]	878K	316.2G	37.75	0.9590	33.49	0.9150	32.03	0.8980	31.87	0.9250	38.07	0.9770
	DPN [64]	832K	140G	37.52	0.9586	33.08	0.9129	31.89	0.8958	30.82	0.9144	–	–
	A^{2}F-L [65]	1,363K	306.1G	38.09	0.9607	33.78	0.9192	32.23	0.9002	32.46	0.9313	38.95	0.9772
	ESRT [42]	677K	–	38.03	0.9600	33.75	0.9184	32.25	0.9001	32.58	0.9318	39.12	0.9774
	SwinIR [19]	878K	195.6G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
	SRFormer (Ours)	958K	183.8G	38.12	0.9611	33.92	0.9221	32.35	0.9023	32.82	0.9398	39.23	0.9801
	SRFormer+ (Ours)	958K	–	**38.18**	**0.9621**	**33.98**	**0.9232**	**32.41**	**0.9036**	**32.88**	**0.9409**	**39.29**	**0.9821**
×3	VDSR [28]	665K	613G	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279	37.22	0.9750
	DRCN [55]	1,774K	17,974G	33.82	0.9226	29.76	0.8311	28.80	0.7963	27.15	0.8276	32.24	0.9343
	CARN [32]	1,592K	119G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
	SRMDNF [58]	1,530K	156G	34.12	0.9250	30.04	0.8370	28.97	0.8030	27.57	0.8400	–	–
	LAPAR-A [59]	544K	114G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
	OISR-LF-s [60]	1,550K	160.1G	34.39	0.9272	30.35	0.8426	29.11	0.8053	28.24	0.8544	–	–
	LatticeNet [61]	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	–	–
	MADNet [62]	930K	88.4G	34.26	0.9262	30.29	0.8410	29.04	0.8033	27.91	0.8464	–	–
	HDRN [63]	878K	187.1G	34.24	0.9240	30.23	0.8400	28.96	0.8040	27.93	0.8490	33.17	0.9420
	DPN [64]	832K	114.2G	33.71	0.9222	29.80	0.8320	28.84	0.7981	27.17	0.8282	–	–
	A^{2}F-L [65]	1,367K	136.1G	34.54	0.9283	30.41	0.8436	29.14	0.8062	28.40	0.8574	33.83	0.9463
	ESRT [42]	770K	–	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
	SwinIR [19]	886K	87.2G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
	SRFormer (Ours)	958K	81.6G	34.67	0.9301	30.59	0.8470	29.26	0.8095	28.72	0.8652	34.06	0.9488
	SRFormer+ (Ours)	958K	–	**34.72**	**0.9313**	**30.66**	**0.8484**	**29.32**	**0.8105**	**28.79**	**0.8686**	**34.11**	**0.9502**
×4	VDSR [28]	665K	613G	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524	28.83	0.8809
	DRCN [55]	1,774K	17,974G	31.54	0.8850	29.19	0.7720	27.32	0.7280	25.12	0.7560	29.09	0.8845
	SRDenseNet [66]	2,015K	390G	32.00	0.8931	28.50	0.7782	27.53	0.7337	26.05	0.7819	30.41	0.9071
	CARN [32]	1,592K	91G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
	CBPN [56]	1,197K	97.9G	32.21	0.8944	28.63	0.7813	27.58	0.7356	26.14	0.7869	–	–
	SRMDNF [58]	1,555K	89G	31.96	0.8930	28.35	0.7770	27.49	0.7340	25.68	0.7730	–	–
	LAPAR-A [59]	659K	94G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
	OISR-LF-s [60]	1,520K	114.2G	32.14	0.8947	28.63	0.7819	27.60	0.7369	26.17	0.7888	–	–
	LatticeNet [61]	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	–	–
	MADNet [62]	1,002K	54.1G	32.11	0.8939	28.52	0.7799	27.52	0.7340	25.89	0.7782	–	–
	HDRN [63]	867K	316.2G	32.23	0.8960	28.58	0.7810	27.53	0.7370	26.09	0.7870	30.43	0.9080
	DPN [64]	832K	140G	31.42	0.8849	28.07	0.7688	27.30	0.7256	25.25	0.7546	–	–
	A^{2}F-L [65]	1,374K	77.2G	32.32	0.8964	28.67	0.7839	27.62	0.7379	26.32	0.7931	30.72	0.9115
	ESRT [42]	751K	–	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
	SwinIR [19]	897K	49.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
	SRFormer (Ours)	958K	42.3G	32.56	0.9018	28.86	0.7884	27.73	0.7429	26.61	0.8013	31.01	0.9168
	SRFormer+ (Ours)	958K	–	**32.62**	**0.9037**	**28.91**	**0.7904**	**27.82**	**0.7441**	**26.68**	**0.8025**	**31.10**	**0.9184**

accumulating these hierarchical characteristics to build more robust representative features that are well-focused on spatial context information. This trait can be confirmed by the obtained SSIM scores, which are based on the visible

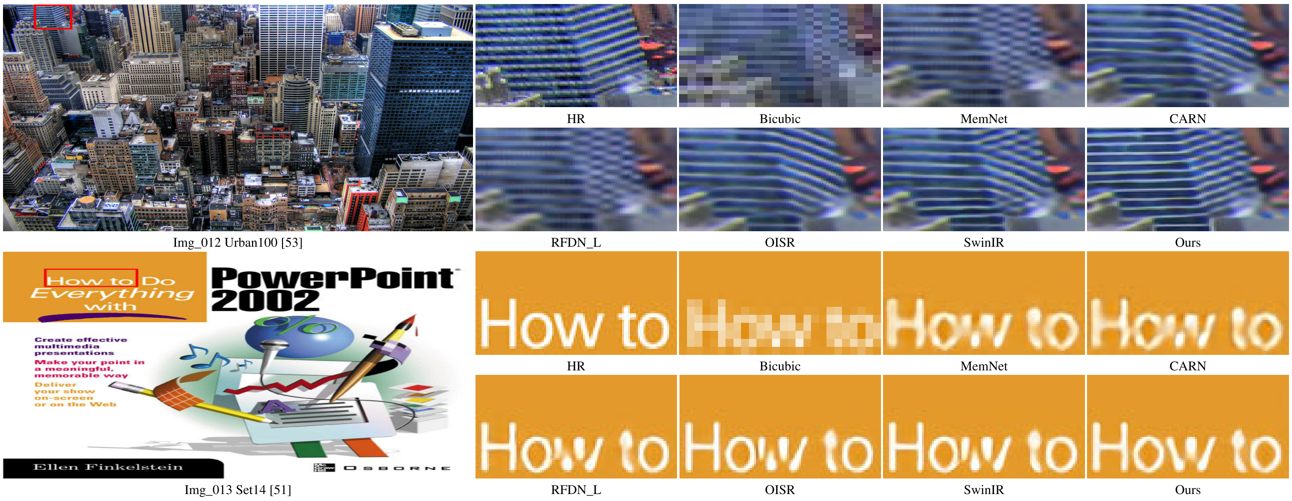


FIGURE 3. Visual results of BI degradation model for ×4 scale factor.

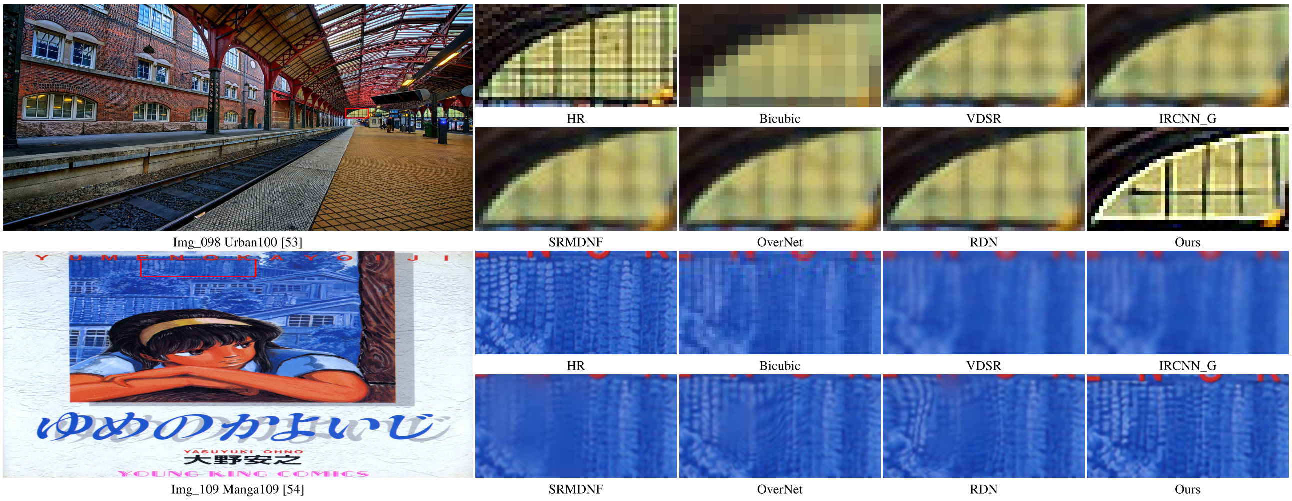


FIGURE 4. Visual results of BD degradation model for ×4 scale factor.

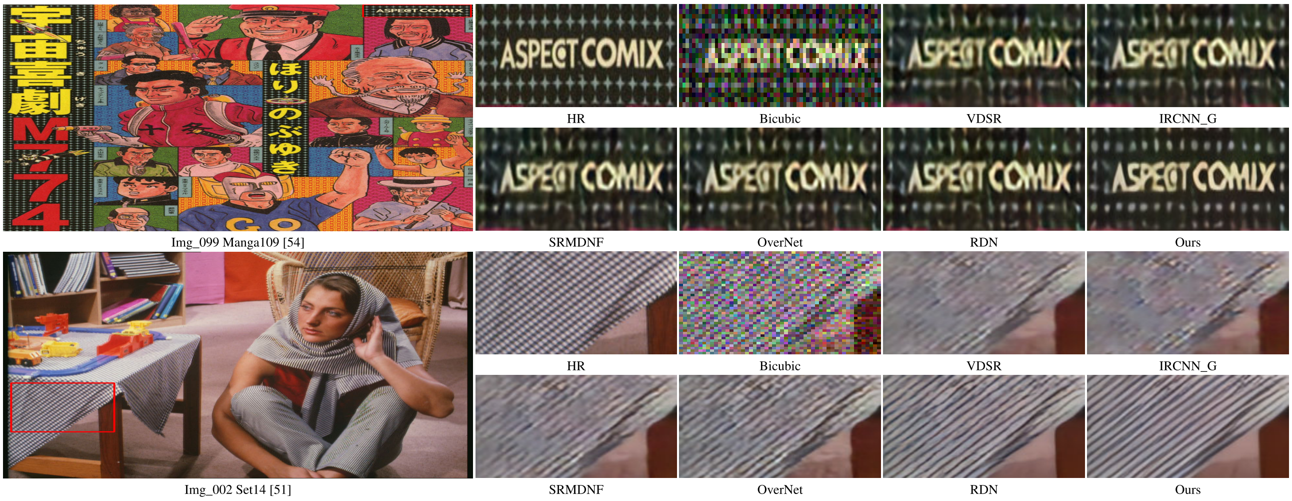


FIGURE 5. Visual results of DN degradation model for ×4 scale factor.

structures in the image and are therefore more accurate. Furthermore, it can be observed that using self-ensembles [72], the proposed SRFormer+ gains even more performance benefits. Several visual outcomes are presented in Fig. 3.

TABLE 2. Quantitative results with BD degradation model. Performance is shown for scale factor ×3. The best and second best results are highlighted in red and blue respectively. SRFormer with self-ensemble results are **Highlighted.**

Methods	Degrad.	Set5 [50]		Set14 [51]		B100 [52]		Urban100 [53]		Manga109 [54]	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRCNN [67]	BD	32.05	0.8944	28.80	0.8074	28.13	0.7736	25.70	0.7770	29.47	0.8924
VDSR [28]	BD	33.25	0.9150	29.46	0.8244	28.57	0.7893	26.61	0.8136	31.06	0.9234
IRCNN_G [68]	BD	33.38	0.9182	29.63	0.8281	28.65	0.7922	26.77	0.8154	31.15	0.9245
IRCNN_C [68]	BD	29.55	0.8246	27.33	0.7135	26.46	0.6572	24.89	0.7172	28.68	0.7701
SRMDNF [58]	BD	34.09	0.9242	30.11	0.8364	28.98	0.8009	27.50	0.8370	32.97	0.9391
RDN [31]	BD	34.57	0.9280	30.53	0.8447	29.23	0.8079	28.46	0.8581	33.97	0.9465
OverNet [16]	BD	34.59	0.9287	30.46	0.8310	29.13	0.8060	28.24	0.8485	–	–
CASGCN [69]	BD	34.62	0.9283	30.60	0.8458	29.30	0.8196	28.68	0.8611	34.27	0.9476
SRFormer (Ours)	BD	34.78	0.9306	30.76	0.8487	29.45	0.8215	28.79	0.8635	34.41	0.9505
SRFormer+ (Ours)	BD	**34.82**	**0.9316**	**31.83**	**0.8498**	**29.52**	**0.8238**	**28.84**	**0.8683**	**34.46**	**0.9517**

TABLE 3. Quantitative results with DN degradation models. Performance is shown for scale factor ×3. The best and second best results are highlighted in red and blue respectively. SRFormer with self-ensemble results are **Highlighted.**

Methods	Degrad.	Set5 [50]		Set14 [51]		B100 [52]		Urban100 [53]		Manga109 [54]	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SRCNN [67]	DN	25.01	0.6950	23.78	0.5898	23.76	0.5538	21.19	0.5737	23.75	0.7148
VDSR [28]	DN	25.20	0.7183	24.00	0.6112	24.00	0.5749	22.22	0.6096	24.20	0.7525
IRCNN_G [68]	DN	25.70	0.7379	24.45	0.6305	24.28	0.5900	22.90	0.6429	24.88	0.7765
IRCNN_C [68]	DN	26.18	0.7430	24.68	0.6300	24.52	0.5850	22.63	0.6205	24.74	0.7701
SRMDNF [58]	DN	27.74	0.8026	26.13	0.6924	25.64	0.6495	24.28	0.7092	26.72	0.8590
RDN [31]	DN	28.46	0.8151	26.60	0.7101	25.96	0.6573	24.92	0.7362	28.00	0.8590
OverNet [16]	DN	28.49	0.8200	26.62	0.7116	25.95	0.6602	24.93	0.7365	–	–
SRFormer (Ours)	DN	28.62	0.8225	26.78	0.7129	26.12	0.6621	25.11	0.7384	28.17	0.8616
SRFormer+ (Ours)	DN	**28.66**	**0.8233**	**26.84**	**0.7137**	**26.20**	**0.6632**	**25.18**	**0.7391**	**28.23**	**0.8621**

As can be seen, the texture direction of the reconstructed images from all of the compared approaches is utterly incorrect while the text is blurred in all the cases at different levels. However, the results obtained by SRFormer are similar to ground truth texture.

C. RESULTS ON BD AND DN DEGRADATION MODELS

We also provide the performance of SRFormer and SRFormer+ on the BD (Blurry) and DN (Noisy) benchmark datasets in Table 2 and Table 3 to illustrate the strengths of the proposed model when it comes to a challenging situation with SOTA models. Due to degradation mismatch the following methods SRCNN and VDSR are re-trained for both BD and DN. As can be seen, SRFormer outperforms all other lightweight SOTA models on challenging benchmark datasets, and it is particularly impressive when compared to other lightweight SOTA models. A high-capability model, RDN [31] is also listed, which is used to demonstrate the superior performance of our SRFormer in comparison to a deep and costly model in these challenging datasets. SRFormer performs better in both datasets notwithstanding, RDN is a significantly expensive network compared to the low-cost SRFormer. RDN is near ×20 more expensive in terms of computational complexity. Furthermore, a visual representation of both challenging BD and DN benchmark datasets is shown in Fig. 4 and Fig. 5 respectively. As can be seen, our proposed method performs better in comparison with other SOTA methods in removing the noises and fuzzy regions from the input image, which results in generating sharper with fine details SR images.

V. ABLATION STUDY

The performance of the proposed model is further investigated through an extensive ablation study that includes in-depth examinations of the impact of each module. The ablation study is designed to provide additional insight into the performance of the proposed model.

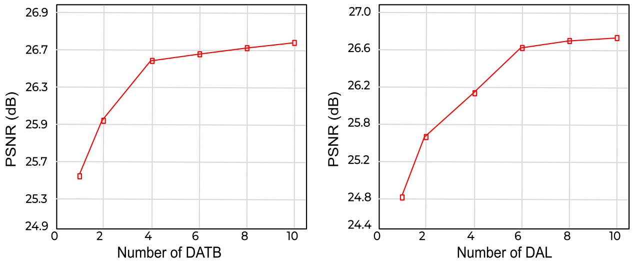


FIGURE 6. Performance investigation on different settings of SRFormer on Urban100 for scale factor ×4.

A. RELATION BETWEEN NUMBER OF TRANSFORMER BLOCKS AND LAYERS VS. PERFORMANCE AND NETWORK PARAMETERS

We investigate deeply the relation between the number of Transformer blocks (DATB) and Transformer layer (DAL) on the performance and model size of our proposed model architecture in Fig 6. We discovered that the performance (PSNR) of the network has a positive relationship with the aforementioned hyperparameters however performance gains by increasing the number of blocks and layers will not come for free. By increasing the number of blocks or layers while performance continuously improves, the overall number of network parameters and FLOPs increases, which makes the network computationally inefficient. Also, we can see that by increasing these hyperparameters, the performance benefit gets more and more limited until it is start to saturate progressively. Thus, we design our network by choosing four Transformer blocks and six Transformer layers inside of each block to still have a lightweight yet powerful feature extraction module.

B. VISUALIZATION ON INFLUENCE OF CONV LAYER IN TRANSFORMER BLOCK

Figure 7 shows the average feature maps of each stage of our Dense Feature Extraction module to investigate the impact of the conv layer when it stacks up with Transformer layers. Each average feature map is the mean of F_{out} in channel dimension, which represents the output of the Transformer block at each stage. The average feature maps without a conv layer are shown on the top row, and with a conv layer within Transformer blocks are illustrated on the bottom row. By visualizing the feature maps, we can first see that, using a conv layer within a Transformer, helps the Transformer to learn sharper representations compared to without a conv layer. Second, as the network focuses more on high-level information, feature maps tend to include more negative values at each stage, indicating a stronger impact of suppressing the smooth area of the input image, which further leads to a more accurate residual image.

C. IMPACT OF DUAL ATTENTION

We further study the impact of both proposed SpAM and SeAM to illustrate the effectiveness of the proposed Dual Attention. We investigate the performance of SRFormer with the standard self-attention layer [20] and each sub-branch of our Dual Attention layer. As can be seen in Table 4, the SRFormer with Dual Attention boosts the performance of the network while using less computational cost compared to when the standard self-attention layer replaces in the network. In contrast to other self-attention layers, Dual Attention is built up with two parallel branches, which are able to encode the spatial information more efficiently and enables Dual Attention to preserve a rich representation while shrinking its depth to make further computation lightweight. Also, it helps the network train faster compare to other transformer-based networks.

TABLE 4. Influence of different settings of the dual attention layer on Urban100 for scale factor ×4.

	SeAM	SpAM	Parameters(K)	PSNR(dB)↑
Meta-Former	–	–	953K	26.47
Dual Attention	–	✓	955K	26.38
	✓	–	942K	26.54
	✓	✓	**958K**	**26.61**

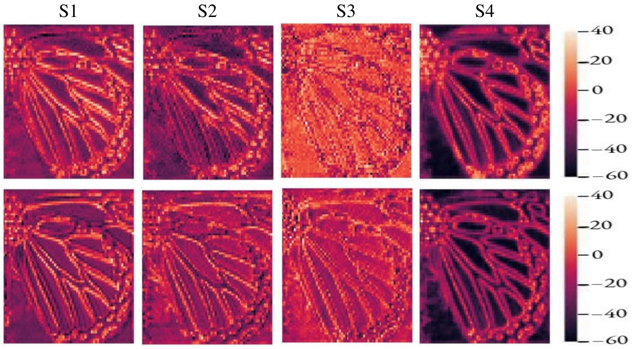


FIGURE 7. Average feature maps of Transformer blocks (DATB). Top: DATB without Conv layer. Bottom: DATB with Conv layer.

D. INFLUENCE OF GATED MLP FEATURE FUSION

Table 5 shows the impact of our proposed lightweight Gated MLP Feature Fusion compared to without and with baseline MLP on the performance of the proposed network. In addition, we investigate the impact of the usage of depthwise, pointwise conv layer, and gated mechanism in our Gated MLP Feature Fusion. As can be seen, SRFormer obtains performance gain compared to when the network does not contain any MLP module or even when it is compared to the baseline MLP with a less computation cost. The intuition behind that is, GMFF uses a gated mechanism to allow gradients to backpropagate more easily through depth, and a Dw-Conv layer between the MLP layers to leak the location information, which leads the network to pay attention to positional information, unlike the baseline MLP that uses positional encoding [22] to introduce the location information, which is not suitable when the test resolution is different from training resolution. Furthermore, we illustrate the performance gain of our Gated MLP Feature Fusion with pointwise, depthwise convolution layers, Gated Mechanism, and without GMFF. As shown in Table 6, the performance of our SRFormer boosts when a depthwise convolution layer with a gated mechanism is used compared to other settings.

TABLE 5. Gated MLP feature fusion performance investigation on Urban100 for ×4.

	Parameters(K)	Memory(M)	PSNR(dB)↑
w/o MLP	955K	2,631	26.52
Baseline MLP	962K	2,875	26.58
GMFF(Ours)	958K	2,739	**26.61**

TABLE 6. Impact of different gated MLP feature fusion setting on Urban100 for scale factor ×4.

	PwConv	DwConv	Gated Mech.	Parameters (K)	PSNR (dB) ↑
GMFF	–	–	–	957K	26.55
	✓	–	–	959K	26.57
	–	✓	–	960K	26.59
	–	✓	✓	**958K**	**26.61**

TABLE 7. Perceptual index comparison between proposed method and recent lightweight state-of-the-art methods on benchmark datasets for scale factor ×4. The lower is better.

Methods	Parameters	Set5	Set14	B100	Urban100	Manga109
CARN [32]	1.5M	6.297	5.775	5.700	5.540	5.132
SRFBN-S [28]	0.6M	6.451	5.775	5.702	5.549	5.010
SRDenseNet [66]	2M	6.128	5.615	5.653	5.526	4.762
RFDN_L [34]	0.6M	6.124	5.644	5.659	5.531	4.810
A^{2}F_L [65]	1.3M	6.084	5.499	5.532	5.179	4.771
SRFormer(Ours)	0.9M	**4.931**	**4.821**	**4.474**	**4.649**	**3.880**

TABLE 8. Average running time (s) and memory consumption (MB) comparison on Urban100 for scale factor ×4.

Methods	Parameters	Memory	Running Time(s)↓	PSNR(dB)↑
CARN [32]	1.5M	1,116	0.032	26.07
SRFBN-S [73]	0.5M	2,154	0.031	25.71
SRDenseNet [66]	2M	5,531	0.221	26.05
RFDN-L [34]	0.6M	3,215	0.033	26.22
A^{2}F-L [65]	1.3M	3,015	0.032	26.32
RCAN [74]	16M	1,531	0.297	26.82
EDSR [30]	43M	2,731	0.085	26.64
RDN [31]	23M	5,015	0.172	26.61
SwinIR [19]	0.9M	3,340	0.216	26.47
SRFormer(Ours)	0.9M	2,739	0.102	**26.61**

E. PERCEPTUAL INDEX METRIC

To assess the quality of the generated super resolution images, the Perceptual Index (PI) is used, which is more accurate in reflecting human perceptions of image quality compared to other metrics (PSNR and SSIM). Table 7 illustrates the PI metric between SRFormer and SOTA methods with the same order of magnitude in terms of network model size. It can be seen that the proposed model achieves lower results (lower is better) compared to other models. This demonstrates the ability of the proposed SRFormer for generating more realistic images.

F. MODEL COMPLEXITY AND INFERENCE TIME ANALYSIS

Table 8 illustrates the advantages of the proposed SRFormer architecture in terms of Network Parameters (M) Inference Time (s) and Memory Consumption (MB) compared to existing light- and heavy-weight SOTA CNN and Transformer base architectures on Urban100. In order to make a fair comparison, all the models are measured with the same configuration with their published source code and default hyper-parameters on a single NVIDIA RTX3090 GPU. As shown, our model has the shortest inference time and less memory hunger per image compared to Transformer models. This comparison illustrates that our model successfully strikes a balance between performance and running time requirements.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present a novel and efficient Transformer architecture-based network called SRFormer. The proposed model is designed by using the strength of both Convolutional and Transformer layers to extract and preserve the fine details of the features while remaining memory efficient. To do so, we introduce a Dual Attention layer, a Transformer layer, which generates the global attention map from two different branches (SpAM and SeAM) in order to capture both local context information and global dependency between sequences. Also, we introduce a lightweight Gated MLP Feature Fusion to aggregate the multi-stage feature representation by focusing on inner spatial information before upsampling module. We demonstrate the efficiency of the proposed method through a series of ablation investigations. We have empirically demonstrated that our approach outperforms previous lightweight state-of-the-art methods on all benchmark datasets, despite having similar or fewer network parameters. In the future, we will expand our proposal for blind super resolution when there is no ground truth during training and inference. To do so, we will attempt to change the methodology of our proposed architecture to use Generative Adversarial Network.

REFERENCES

[1] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu, "A comprehensive review of deep learning-based single image super-resolution," *PeerJ Comput. Sci.*, vol. 7, p. e621, Jul. 2021.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[3] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2021, pp. 387–395.

[4] H. Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, Jan. 2009.

[5] H. Li, M. Trocan, D. Galayko, and M. Sawan, "CMISR: Circular medical image super-resolution," 2023, *arXiv:2308.08567*.

[6] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.

[7] J. Gao, Z. Yue, Y. Liu, S. Xie, X. Fan, and R. Liu, ‘‘Diving into darkness: A dual-modulated framework for high-fidelity super-resolution in ultra-dark environments,’’ 2023, *arXiv:2309.05267*.

[8] S. R. Malakshan, M. S. E. Saadabadi, M. Mostofa, S. Soleymani, and N. M. Nasrabadi, ‘‘Joint super-resolution and head pose estimation for extreme low-resolution faces,’’ *IEEE Access*, vol. 11, pp. 11238–11253, 2023.

[9] J.-S. Kim, K. Ko, H. Kim, and C.-S. Kim, ‘‘RPF: Reference-based progressive face super-resolution without losing details and identity,’’ *IEEE Access*, vol. 11, pp. 46707–46718, 2023.

[10] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, ‘‘Learning a deep convolutional network for light-field image super-resolution,’’ in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Springer, Dec. 2015, pp. 57–65.

[11] A. Mehri, P. B. Ardakani, and A. D. Sappa, ‘‘MPRNet: Multi-path residual network for lightweight image super resolution,’’ in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2703–2712.

[12] P. Behjati, P. Rodriguez, C. Fernández, I. Hupont, A. Mehri, and J. Gonzàlez, ‘‘Single image super-resolution based on directional variance attention network,’’ *Pattern Recognit.*, vol. 133, Jan. 2023, Art. no. 108997.

[13] A. Mehri, P. Behjati, and A. D. Sappa, ‘‘TnTViT-G: Transformer in transformer network for guidance super resolution,’’ *IEEE Access*, vol. 11, pp. 11529–11540, 2023.

[14] P. Behjati, P. Rodriguez, C. F. Tena, A. Mehri, F. X. Roca, S. Ozawa, and J. Gonzàlez, ‘‘Frequency-based enhancement network for efficient super-resolution,’’ *IEEE Access*, vol. 10, pp. 57383–57397, 2022.

[15] A. Mehri, P. B. Ardakani, and A. D. Sappa, ‘‘LiNet: A lightweight network for image super resolution,’’ in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7196–7202.

[16] P. Behjati, P. Rodríguez, A. Mehri, I. Hupont, C. F. Tena, and J. Gonzàlez, ‘‘OverNet: Lightweight multi-scale super-resolution with overscaling network,’’ in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2693–2702.

[17] Z. Hou and S.-Y. Kung, ‘‘Efficient image super resolution via channel discriminative deep neural network pruning,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3647–3651.

[18] Y. Xu, C. Xu, X. Chen, W. Zhang, C. Xu, and Y. Wang, ‘‘Kernel based progressive distillation for adder neural networks,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12322–12333.

[19] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, ‘‘SwinIR: Image restoration using Swin transformer,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ‘‘Swin transformer: Hierarchical vision transformer using shifted windows,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ‘‘An image is worth 16×16 words: Transformers for image recognition at scale,’’ 2020, *arXiv:2010.11929*.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, ‘‘End-to-end object detection with transformers,’’ in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[24] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, ‘‘Visual attention network,’’ 2022, *arXiv:2202.09741*.

[25] A. Mehri, ‘‘Deep learning based architectures for cross-domain image processing,’’ Ph.D. dissertation, Dept. Comput. Sci., Auton. Univ. Barcelona, UAB, 2023.

[26] C. Dong, C. C. Loy, and X. Tang, ‘‘Accelerating the super-resolution convolutional neural network,’’ in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 391–407.

[27] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[28] J. Kim, J. K. Lee, and K. M. Lee, ‘‘Accurate image super-resolution using very deep convolutional networks,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[29] Y. Tai, J. Yang, X. Liu, and C. Xu, ‘‘MemNet: A persistent memory network for image restoration,’’ in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.

[30] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, ‘‘Enhanced deep residual networks for single image super-resolution,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[31] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, ‘‘Residual dense network for image super-resolution,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[32] N. Ahn, B. Kang, and K.-A. Sohn, ‘‘Fast, accurate, and lightweight super-resolution with cascading residual network,’’ in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 252–268.

[33] A. Muqeet, J. Hwang, S. Yang, J. H. Kang, Y. Kim, and S.-H. Bae, ‘‘Multi-attention based ultra lightweight image super-resolution,’’ 2020, *arXiv:2008.12912*.

[34] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, ‘‘Residual feature aggregation network for image super-resolution,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2356–2365.

[35] X. Chu, B. Zhang, R. Xu, and H. Ma, ‘‘Multi-objective reinforced evolution in mobile neural architecture search,’’ 2019, *arXiv:1901.01074*.

[36] X. Chu, B. Zhang, H. Ma, R. Xu, and Q. Li, ‘‘Fast, accurate and lightweight super-resolution with neural architecture search,’’ 2019, *arXiv:1901.07261*.

[37] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, ‘‘Tokens-to-token ViT: Training vision transformers from scratch on ImageNet,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.

[38] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, ‘‘Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[39] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, ‘‘SegFormer: Simple and efficient design for semantic segmentation with transformers,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–12.

[40] L. Shi, Y. Zhang, J. Cheng, and H. Lu, ‘‘Decoupled spatial–temporal attention network for skeleton-based action recognition,’’ 2020, *arXiv:2007.03263*.

[41] C. Plizzari, M. Cannici, and M. Matteucci, ‘‘Skeleton-based action recognition via spatial and temporal transformer networks,’’ *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[42] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, ‘‘Transformer for single image super-resolution,’’ 2021, *arXiv:2108.11084*.

[43] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, ‘‘SRFormer: Permuted self-attention for single image super-resolution,’’ 2023, *arXiv:2303.09735*.

[44] M. Kumar, D. Weissenborn, and N. Kalchbrenner, ‘‘Colorization transformer,’’ 2021, *arXiv:2102.04432*.

[45] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, ‘‘Uformer: A general U-shaped transformer for image restoration,’’ 2021, *arXiv:2106.03106*.

[46] S. Waqas Zamir, A. Arora, S. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, ‘‘Restormer: Efficient transformer for high-resolution image restoration,’’ 2021, *arXiv:2111.09881*.

[47] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick, ‘‘Early convolutions help transformers see better,’’ in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–9.

[48] D. Hendrycks and K. Gimpel, ‘‘Gaussian error linear units (GELUs),’’ 2016, *arXiv:1606.08415*.

[49] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, ‘‘MetaFormer is actually what you need for vision,’’ 2021, *arXiv:2111.11418*.

[50] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, *Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding*. BMVA Press, 2012, pp. 1–10.

[51] R. Zeyde, M. Elad, and M. Protter, ‘‘On single image scale-up using sparse-representations,’’ in *Proc. Int. Conf. curves Surf.* Cham, Switzerland: Springer, 2010, pp. 711–730.

[52] 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 *Proc. 8th IEEE Int. Conf. Comput. Vis.*, Oct. 2001, pp. 416–423.

[53] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[54] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based Manga retrieval using MANGA109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.

[55] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[56] F. Zhu and Q. Zhao, "Efficient single image super-resolution via hybrid residual feature learning with compact back-projection network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2453–2460.

[57] X. Chu, B. Zhang, H. Ma, R. Xu, and Q. Li, "Fast, accurate and lightweight super-resolution with neural architecture search," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 59–64.

[58] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3262–3271.

[59] W. Li, K. Zhou, L. Qi, N. Jiang, J. Lu, and J. Jia, "LAPAR: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[60] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng, "ODE-inspired network design for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1732–1741.

[61] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu, "LatticeNet: Towards lightweight image super-resolution with lattice block," in *Proc. 16th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 272–289.

[62] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, Mar. 2021.

[63] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Hierarchical dense recursive network for image super-resolution," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107475.

[64] Y. Liang, R. Timofte, J. Wang, S. Zhou, Y. Gong, and N. Zheng, "Single-image super-resolution—When model adaptation matters," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107931.

[65] X. Wang, Q. Wang, Y. Zhao, J. Yan, L. Fan, and L. Chen, "Lightweight single-image super-resolution network with attentive auxiliary feature learning," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1-17.

[66] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4809–4817.

[67] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[68] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2808–2817.

[69] Y. Yang and Y. Qi, "Image super-resolution via channel attention and spatial graph convolutional network," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107798.

[70] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.

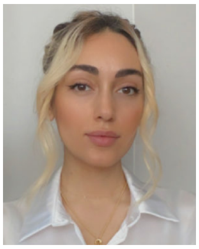
[71] B. Heo, S. Chun, S. Joon Oh, D. Han, S. Yun, G. Kim, Y. Uh, and J.-W. Ha, "AdamP: Slowing down the slowdown for momentum optimizers on scale-invariant weights," 2020, *arXiv:2006.08217*.

[72] R. Timofte, R. Rothe, and L. Van Gool, "Seven ways to improve example-based single image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1865–1873.

[73] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3862–3871.

[74] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

ARMIN MEHRI received the B.Sc. and M.Sc. degrees in computer science from Eastern Mediterranean University, in 2014 and 2017, respectively, and the Ph.D. degree (cum laude) in deep learning and computer vision from the Computer Vision Center, Autonomous University of Barcelona. He is currently the Computer Vision/Deep Learning Team Lead with Camaleonic Analytics, working on AI-based software for sponsorship in sports. His main research interests include computer vision and image processing under cross-modal frameworks.



PARICHEHR BEHJATI received the bachelor's and master's degrees in computer science from Eastern Mediterranean University and the Ph.D. degree (cum laude) in deep learning and computer vision from the Computer Vision Center, Autonomous University of Barcelona. From 2014 to 2016, she was a Research Assistant with the Department of Computer Science, Eastern Mediterranean University.

DARIO CARPIO received the degree in computer science engineering from ESPOL Polytechnic University, Guayaquil, Ecuador. He is currently a member of the CIDIS Research Center, ESPOL Polytechnic University. He is also working on thermal images and visible spectrum image super-resolution, where he is contributing to novel guided SR architectures.

ANGEL DOMINGO SAPPA (Senior Member, IEEE) received the degree in electromechanical engineering from the National University of La Pampa, General Pico, Argentina, in 1995, and the Ph.D. degree in industrial engineering from the Polytechnic University of Catalonia, Barcelona, Spain, in 1999. In 2003, after holding research positions in France, the U.K., and Greece, he joined the Computer Vision Center, Barcelona, where he currently holds a senior scientist position. Since 2016, he has been a Full Professor with ESPOL Polytechnic University, Guayaquil, Ecuador, where he leads the computer vision team with the CIDIS Research Center. He is also the Director of the Electrical Engineering Ph.D. Program.