HIPA: Hierarchical Patch Transformer for Single Image Super Resolution

Qing Cai, Member, IEEE, Yiming Qian, Jinxing Li, Member, IEEE, Jun Lyu, Yee-Hong Yang, Life Senior Member, IEEE, Feng Wu, Fellow, IEEE, and David Zhang, Life Fellow, IEEE

Abstract—Transformer-based architectures start to emerge in single image super resolution (SISR) and have achieved promising performance. However, most existing vision Transformer-based SISR methods still have two shortcomings: (1) they divide images into the same number of patches with a fixed size, which may not be optimal for restoring patches with different levels of texture richness; and (2) their position encodings treat all input tokens equally and hence, neglect the dependencies among them. This paper presents a HIPA, which stands for a novel Transformer architecture that progressively recovers the high resolution image using a hierarchical patch partition. Specifically, we build a cascaded model that processes an input image in multiple stages, where we start with tokens with small patch sizes and gradually merge them to form the full resolution. Such a hierarchical patch mechanism not only explicitly enables feature aggregation at multiple resolutions but also adaptively learns patch-aware features for different image regions, e.g., using a smaller patch for areas with fine details and a larger patch for textureless regions. Meanwhile, a new attention-based position encoding scheme for Transformer is proposed to let the network focus on which tokens should be paid more attention by assigning different weights to different tokens, which is the first time to our best knowledge.

Furthermore, we also propose a multi-receptive field attention module to enlarge the convolution receptive field from different branches. The experimental results on several public datasets demonstrate the superior performance of the proposed HIPA over previous methods quantitatively and qualitatively. We will share our code and models when the paper is accepted.

Index Terms—Image restoration, single image super-resolution, hierarchical patch transformer, attention-based position embedding.

I. INTRODUCTION

SINGLE Image Super-Resolution (SISR), aiming to recover a high-resolution (HR) image from its corresponding degraded low-resolution (LR) version, plays an important and fundamental role in computer vision and image processing due to its wide range of real-world applications, such as medical imaging [1], surveillance [2] and remote sensing [3], amongst others. SISR is a very challenging and ill-posed problem because there is no unique solution for any given LR input [4], [5].

Deep convolutional neural networks (CNNs) have achieved remarkable success in SISR and various architectures have been presented so far, for example, residual learning [6], [7], [8], dense connections [9], [10], UNet-like architectures with skip connections [11], [12], dilated convolutions [13], [14], generative models [15], [16], [17], [18] and other kinds of CNNs [19], [20], [21]. However, the convolution in CNN uses a sliding window to extract local features and hence, is weak in capturing long-range or non-local dependencies, which are important for SISR. In particular, for some regions with fine textures, faithful reconstruction depends not only on local relationships but also on long-range dependencies [22], [23]. To alleviate this issue, many attention mechanisms have been proposed and introduced into SISR, such as global attention mechanism [24], [25], [26], [27], [28] and non-local attention mechanism [29], [30], [31], [32]. As shown in Fig. 1, although some state-of-the-art (SOTA) methods such as NLSN [31] could recover some amounts of high-frequency details, the reconstructed slanted line structures exhibit fuzzy and blurry boundaries, which are faithfully restored using our hierarchical patch Transformer.

Inspired by the significant success of Transformer in natural language processing [33] for its advantages in modeling long-range context, vision Transformer is also introduced into the...
field of SISR [34], [35], [36], [37], [38] and has obtained superior results than many SOTA CNN-based methods due to the multi-head self-attention mechanism that is capable of modeling long-distance dependencies [39]. Very recently, hybrid architectures combining CNN and Transformer start to emerge in the community [37] to fully utilize the advantage of CNN in extracting local features and the advantage of Transformer in establishing long-range dependencies. Although existing Transformer-based SISR models have achieved superior results, the recovered results still exhibit blurry boundaries as shown in the result of SwinIR [37] in Fig. 1. The main reasons may lie in two shortcomings of existing vision Transformer-based SISR methods. First, almost all of them partition all input images into the same number of fixed-size patches, which may not be ideal considering different images on image regions have their own characteristics [40]. Second, the position encoding of most vision Transformer-based SISR methods treats all input tokens equally. However, the low-resolution input tokens contain abundant information for SISR, which are treated equally across tokens and hence, the representation ability of Transformer is limited.

In order to compensate the above two shortcomings, in this paper, we propose a Hierarchical Patch (HIPA) Transformer by partitioning an input image into a hierarchy of patches with different sizes. In particular, a multi-stage architecture is first developed by alternatingly stacking CNN and Transformer to boost their benefits in feature extraction. Then, to achieve different size patch input for the Transformer and to let the Transformer establish global dependencies from different numbers of tokens, the LR image is first partitioned into a hierarchy of subblocks, which are used as inputs to the Transformer by starting from the small-size blocks and gradually merging them in the next stage. In addition, we design a novel attention-based position encoding scheme for the Transformer based on dilated channel attention to model the position information with a continuous dynamical model. Besides, a multi-receptive field attention module is proposed based on dilated convolution with different dilation factors to enlarge the convolution receptive field from different branches.

As shown in Fig. 1, our HIPA obtains better visual quality compared with that of other state-of-the-art SISR methods.

Briefly, the contributions of this paper mainly include:

- A novel hierarchical patch Transformer has been designed to achieve multi-size patches for Transformers. This approach is more effective than treating all samples with the same number of fixed-size patches because the hierarchical patch Transformer allows patches with different texture richness to adopt different sizes, rather a single size patch;
- A new attention-based position encoding scheme is proposed for Transformer that allows the network to focus on which tokens should be paid more attention, which is the first time to our best knowledge;
- A multi-receptive field dilated attention module is designed to enlarge the convolution receptive field from different branches, which achieves relatively smaller increase of the computational complexity compared to the one by increasing the depth and the filter size of a CNN to enlarge the receptive field.

The rest of the paper is organized as follows: Section II briefly overviews related works. Section III presents the proposed HIPA Transformer and discusses its advantages and differences with existing methods. Section IV presents the experimental results and analysis of the proposed method by comparing it with state-of-the-art models. Finally, the paper concludes in Section V.

II. RELATED WORK

A. CNN-Based Models

The SRCNN model proposed by Dong et al. [41] is a pioneering work to apply CNN to single image super-resolution, which has achieved superior performance than traditional methods [42], [43], [44] by using only a three-layer CNN to represent the mapping between LF and HR images. Based on the SRCNN, many deeper and wider CNN based SISR models have been proposed to achieve better restoration performance. However, blindly increasing the depth of a network does not necessarily improve the performance but may introduce many new issues for training, for example, the vanishing or exploding gradient [45]. Later, residual learning is introduced into SISR to ease the training difficulty of deeper networks. For example, by introducing residual learning into a deeper network, Kim et al. can stack more convolutional layers and propose the VDSR [46]. However, all of the above models need to first pre-process the LR input to obtain the desired image size using interpolation, which is not only time consuming but also often introduces noise and blurriness in the input image. To address the above issues, Dong et al. [47] introduce a deconvolution layer as the last layer and achieve end-to-end training for SISR. Such a deconvolution layer is then substituted by a more efficient sub-pixel convolution layer [48] proposed by Shi et al., which is also adopted by our method similar to the EDSR [45] and the RCAN [24]. However, all of these models treat the LR features equally across channels, which inevitably limits the restoration capability of CNNs. Even worse, the convolution kernel usually has a limited
receptive field and cannot sufficiently extract long-range or non-local features. As a result, for some regions with fine details, these methods yield poor performance.

B. Attention-Based Models

To address the above issues, attention mechanism [24], [26], [28], [30] is introduced into SISR to guide the deep neural network to selectively pay more attention on features where there is more information. For example, by integrating channel attention and residual blocks, Zhang et al. propose the RCAN [24], which markedly improves the representational performance of the CNN. Dai et al. propose the SAN [28] using a novel trainable second-order channel attention. However, the channel attention treats different convolution layers independently and neglects the correlation among them. To alleviate this issue, Niu et al. propose the HAN [26] by integrating a layer attention module and a channel-spatial attention module into the residual blocks. More recently, non-local attention modules [29], [30], [31], [32] are proposed to address the inherent issue of CNNs in establishing long range or non-local dependencies among exacted features. For example, Zhang et al. [29], propose the RNAN by mixing a local masked branch and a non-local attention mechanism, which are, respectively, in charge of concentrating on extracting more local structures and considering more long-range dependencies in the extracted features. Mei et al. [30] propose the CSNLN by integrating a Cross-Scale Non-Local prior with local and in-scale non-local priors using a recurrent neural network, which can efficiently explore the existing cross-scale feature similarities in images. Xia et al. [32] propose an efficient non-local attention module by using the kernel function of approximation and the associative law of matrix multiplication, which successfully achieves comparable performance compared to that of the previous non-local attention module while requires only linear computation and space complexity with respect to the LR size. However, these models are still incapable of adequately and comprehensively compensate for the shortcomings of CNNs in establishing long-range dependencies.

C. Transformer-Based Models

Inspired by the significant performance of the vision Transformer [33], [49], [50], it has also been applied to the SISR field [34], [35], [36], [37], [38]. For example, Chen et al. propose the image processing Transformer (IPT) [35] model for various image restoration tasks based on a pre-trained standard Transformer [33]. Recently, to capture local relationships, researchers begin to introduce convolutions to Transformers by integrating the vision Transformer module with convolution [37], [51], [52], [53], [54]. For example, Liang et al. [37] propose the Swin Transformer-based image resolution model (SwinIR) by combining CNN and Transformer and achieves superior performance while maintaining computational efficiency. Huang et al. propose DGS-Swin [52] by introducing a learned Gaussian Scale Mixture (GSM) prior into the Swin Transformer. In addition to classic performance-oriented SISRs, hybrid architectures have also emerged in the field of light-weigh SISR [53], [54], [55]. For example, Lu et al. propose a novel Efficient Super-Resolution Transformer (ESRT) [53], which integrates a lightweight CNN backbone and a lightweight Transformer backbone to achieve a small GPU memory footprint using an efficient multi-head attention. Fang et al. proposed a Hybrid Network of CNN and Transformer (HNCT) [54] for lightweight image SISR, which can exploit both local and non-local priors by integrating CNN and Transformer. Although Transformer-based SISRs have achieved impressive results, most existing Transformers divide images into the same number of fixed-size patches, which may not be ideal for restoring patches with different levels of texture richness. Besides, the position encodings used in most existing Transformer are predefined and treat the positional information of different tokens equally.

III. METHODOLOGY

A. Issues and Motivations

As discussed in the contribution part of Section I and the Transformer-based models part of Section II, using fixed-size patches with varying level of texture richness is suboptimal and can limit the recovery performance of many existing Transformer-based SISR models. To further explain this, we provide an example shown in Fig. 2 that demonstrates the impact of patch size on the “monarch” image from the Set14 dataset. Fig. 2(a) shows the fixed-size split of vision Transformer-based methods (left column), in which the input image is split into the same number of fixed-size tokens, a selected textureless background region (middle column) and a butterfly head region with fine detail (right column). Fig. 2(b) shows the multi-size split of our HIPA Transformer (left column), in which the input image is partitioned into multiple stages, where tokens with small patch sizes are used first and gradually merged with larger patches to form the full resolution, two selected regions (last two columns) the same as that in Fig. 2(a). From the visual and quantitative comparison of the selected background region (middle column) between using the existing fixed-size patch and our multi-size patch.
it can be observed that their visual quality and PSNR values are very similar, which suggests that using a large patch size for textureless regions is enough for the network to finish the final restoration. However, from the recovery performance comparison of the selected butterfly head region (right column) between using the existing fixed-size patch and using our multi-size patch, it can be found that their visual quality and PSNR values have a certain gap, which suggests that using a large patch size for regions with fine detail is not optimal. In contrast, in this case, using a smaller patch size is more helpful to recover fine details, which is demonstrated by the improved PSNR value using our multi-size patch. From the above discussion, we can summarize that: (1) using fixed-size patches with different texture richness in the whole restoration process is inappropriate, which is the reason that many existing Transformer-based models still exhibit blurry boundaries; (2) using multi-size patches with different richness is helpful to improve the restoration results, which motivates the proposed HIPA Transformer.

### B. Hierarchical Patch Transformer

As shown in Fig. 3, the proposed HIPA consists of three stages to progressively recover the high-resolution (HR) image from its low-resolution (LR) input. The Transformer of the first two stages mainly learn broad contextual information, while the last stage focuses more on learning the desired details. Finally, the training loss is defined based on the summation over all outputs $\ell_1(\cdot), \ell_2(\cdot), \ell_3(\cdot)$ of different stages to optimize HIPA.

Fig. 3. Overall framework of our HIPA method for progressive SISR. Each stage in the proposed model is constructed based on the proposed HIPA block, which consists of two main modules. The first module is a Transformer designed to learn global dependencies between contexts (See Section III-B for details). The second module is a multi-receptive field attention group used to exhaustively mine local features contained in the original LR image (See Section III-D for details).

HIPA, respectively. $I_{LR}^{j}$ denotes the $j$-th patch at Stage $i$. For example, $I_{LR}^{1}$ denotes the 1-st patch at Stage 1, i.e., the upper left corner patch of Stage 1 input shown in Fig. 3.

Following [45] and [24], we also use one convolution layer to extract the shallow feature (SF) $F_{0}^{j}$ from the original LR image. For Stage 1:

$$F_{0}^{j} = H_{SF}(I_{LR}^{j}) \quad (j = 1, 2, 3, 4),$$

where $H_{SF}$ denotes the convolution operation. Then, the extracted shallow feature is input to the proposed HIPA block to further extract deep features:

$$F_{HIPA}^{j} = H_{HIPA}(F_{0}^{j}) \quad (j = 1, 2, 3, 4),$$

where $H_{HIPA}$ denotes the proposed HIPA block. After stitching $F_{HIPA}^{1}$ with $F_{HIPA}^{2}$ and stitching $F_{HIPA}^{3}$ with $F_{HIPA}^{4}$ using concatenate operation, dubbed vertical stitching, we obtain the output features of Stage 1, which are then concatenated with the shallow features of Stage 2 as shown in Fig. 3:

$$F_{Sti}^{1} = H_{Sti}(F_{HIPA}^{1}, F_{HIPA}^{2}) + H_{Sti}(F_{0}^{1}, F_{0}^{2}),$$

$$F_{Sti}^{2} = H_{Sti}(F_{HIPA}^{2}, F_{HIPA}^{3}) + H_{Sti}(F_{0}^{2}, F_{0}^{4}),$$

where $H_{Sti}$ denotes the stitch using the concatenate operation. We utilize vertical stitching for sub-patches rather than horizontal stitching, i.e., stitching $F_{HIPA}^{1}$ with $F_{HIPA}^{2}$ and stitching $F_{HIPA}^{3}$ with $F_{HIPA}^{4}$ using the concatenate operation. Although we also investigated horizontal stitching, it did not yield significant differences. Finally, the recovered HR image of Stage 1: $I_{HR}^{1}$, is obtained by further stitching $F_{Sti}^{1}$ and $F_{Sti}^{2}$ using concatenate operation, and then successively input the stitched result into an upscale module and a reconstruction module (i.e., one convolution layer) as follows:

$$I_{HR}^{1} = H_{Rec}(H_{U}(H_{Sti}(F_{HIPA}^{1}, F_{HIPA}^{2}))),$$

where $H_{U}$ and $H_{Rec}$ denote the upscale and reconstruction module, respectively.
For Stage 2 and Stage 3, the extracted shallow features $F_{0}^{2,j}$ ($j = 1, 2$) and $F_{0}^{3,1}$ need to be first concatenated with the output features of the upper stage, which is then input into the next operation similar to Stage 1. Finally, the recovered HR images of Stage 2: $I_{HR}^{2}$ and Stage 3: $I_{HR}^{3}$ can be obtained. As shown in Fig. 3, the predictions of the three stages are gradually improved. For example, the prediction of Stage 2 is the refinement of Stage 1. With the multi-stage refinement, image regions with high spatial frequency are gradually recovered.

Finally, the proposed HIPA is trained using a training loss, which is the sum over all the outputs of $I_{HR}^{1}$ (Stage 1), $I_{HR}^{2}$ (Stage 2) and $I_{HR}^{3}$ (Stage 3):

$$L(\Theta) = \ell^1(I_{HR}^{1}, I_{GT}) + \ell^2(I_{HR}^{2}, I_{GT,T}) + \ell^3(I_{HR}^{3}, I_{GT}),$$  \hspace{1cm} (5)

where $\Theta$ denotes the parameter set of the proposed network. $\ell^1(\cdot)$, $\ell^2(\cdot)$ and $\ell^3(\cdot)$, respectively, stand for the loss of Stage 1, Stage 2 and Stage3. This work also uses the $L_1$ loss following previous work for the sake of fairness. $I_{GT}$ denotes the ground-truth HR image.

### C. Attention-Based Position Encoding

To incorporate the order of the token sequence, position encodings are usually adopted in Transformers [56]. However, the original position embedding of ViT is pre-defined and independent of input tokens. When an input LR image with a new size is input, the number of patches will be different from that before and the learned position embedding will be mismatched with the new size. So, the input image with a new size has to be first interpolated to the desired size, although increasing the depth and the filter size of the CNN introduces more parameters but also increases the computational complexity [14]. Thus, we propose the dilated convolution based channel attention to enlarge the receptive field of the networks, which is the most significant difference between ours and the Squeeze-and-Excitation network (SE) [57].

Specifically, for each dilated convolution based channel attention shown in Fig. 5, denote $X_i = [x_{i,1}, \ldots, x_{i,c}, \ldots, x_{i,C}]$ to be the input, which contains $C$ 2D feature map $x_{i,c} \in R^{H \times W}$, where $H$ and $W$, respectively, are the height and width of the feature map. Firstly, by shrinking the extracted features using max pooling, the output feature $Z_i = [z_{i,1}, \ldots, z_{i,c}, \ldots, z_{i,C}]$ of each branch can be obtained, where $z_{i,c} \in R^{H/W \times W}$ denotes the output feature. Then, two dilated convolution layers and an activation function are applied to fully exploit feature dependencies from the aggregated information. Finally, the sigmoid function is adopted as the activation function:

$$s_{i,c} = f(H_{GPL}(W_{D} \delta(W_{G}z_{i,c}))),$$  \hspace{1cm} (6)

where $f(\cdot)$, $H_{GPL}(\cdot)$ and $\delta(\cdot)$, respectively, stand for the sigmoid function, the global average pooling and the ReLU function.
function. \(W_D\) is the weight set of the first dilated convolution layer in the channel attention set shown in Fig. 5, which plays the role of downsampling with a reduction ratio \(\gamma\) (we set \(\gamma = 16\)). After the ReLU function, the low-dimension feature is then upsampled with ratio \(\gamma\) by the second dilated convolution layer. \(W_U\) denotes its weight set. The channel statistics \(s\) can be obtained to rescale the input \(x_{i,c}\):

\[
\hat{x}_{i,c} = s_{i,c} \cdot x_{i,c},
\]

where \(s_{i,c}\) and \(x_{i,c}\) denote, respectively, the scaling factor and feature maps of the \(c\)-th channel.

Besides, we introduce the LFS connection to ensure stability in training the network and to bypass redundant features in the low-quality image. The final output of MRFAG is obtained as

\[
F_{\text{MRFAG}} = F_{0}^{i,j} + \omega_{\text{LFS}}(F_{\text{MRFAM}_G}),
\]

where \(\omega_{\text{LFS}}\) denotes the weight of the convolution layer at the tail of MRFAG. \(F_{\text{MRFAG}}\) and \(F_{\text{MRFAM}_G}\), respectively, denote the output of MRFAG and the \(G\)-th output of MRFAM.

Fig. 6 shows a comparison of the class active map (CAM) before and after using the MRFAG. It can be found that the CAM after using MRFAG becomes sharper than before, validating that the MRFAG has a better capability of recovering high-frequency signals. This also enables the network to focus more on recovering textures and details.

### E. Discussions

Below, we discuss the significant differences between our HIPA Transformer and two most closely related Transformers: IPT [35] and SwinIR [37].

1) **Differences to the IPT Model:** Based on a pre-trained standard Transformer [33], Chen et al. propose the image processing Transformer (IPT) [35] model for image restoration tasks, which achieves superior performance than most CNN-based SISR methods. Although it is a Transformer-based SISR method similar to our method, there are three significant differences between the IPT and our HIPA Transformer: (i) the IPT model uses a pre-trained Transformer, which means that when we apply it for image super-resolution, we need to first pre-train the Transformer using large labeled datasets and then fine-tune the whole network. As a result, its performance is limited by the lack of sufficient labeled samples for fine-tuning. In contrast, our HIPA Transformer is an end-to-end network, which successfully avoids the tedious pre-training and fine-tuning process; (ii) IPT uses fixed-size patches for all input tokens with different richness, which is not optimal and limits its performance as discussed in Section III-A, while our HIPA Transformer uses multi-size patches for tokens with different richness, e.g., using a smaller patch for areas with fine details and a large patch for textureless regions; (iii) The IPT uses Transformer to extract features and to construct long-range dependencies, while our HIPA Transformer is a hybrid architecture combining CNN and Transformer, which can fully utilize the advantage of CNN in local feature extraction and the advantage of Transformer in establishing long-range dependencies. More comparisons of experimental results are shown in Section IV-B.

2) **Differences to the SwinIR Model:** Liang et al. propose the SwinIR [37] by combining CNN with the Swin Transformer [64] into one network, which is also a hybrid architecture similar to our HIPA Transformer. Our key distinctions from it are summarized as follows: (i) The SwinIR uses a plain concatenation of CNN and Transformer with a fixed patch size, while we use a multi-stage model that divides the input into different blocks and aggregates them from small to large by alternating CNN and Transformer, which not only explicitly enables feature aggregation at multiple resolutions but also adaptively learns patch-aware features for different image regions; (ii) The local-resolution input tokens contain abundant information for SISR, however, the SwinIR treats all the input tokens equally and hence, limits its representation ability. In contrast, we design a novel attention-based encoding method to focus on the important tokens and to improve its performance for regions with fine details; (iii) The CNN used in the SwinIR model detects local image features using the same scale, treats all LR image features equally, and neglects the dependencies among them. In contrast, our HIPA Transformer uses multi-size patches for tokens with different richness, e.g., using a smaller patch for areas with fine details and a large patch for textureless regions; (iv) The IPT uses Transformer to extract features and to construct long-range dependencies, while our HIPA Transformer is a hybrid architecture combining CNN and Transformer, which can fully utilize the advantage of CNN in local feature extraction and the advantage of Transformer in establishing long-range dependencies. More comparisons of experimental results are shown in Section IV-B.

### IV. Experiments

#### A. Settings

1) **Datasets:** Following previous works [24], [26], [28], [30], we also choose DIV2K [65] as our training dataset, which contains 800 training images and 100 validation images. For testing, we select the standard public datasets: Set5 [66], Set14 [67], B100 [68], Urban100 [69], and Manga109 [70] as our test datasets. All degraded datasets are obtained by the bicubic interpolation model.

2) **Evaluation Metrics:** To quantitatively compare the recovered HR results of the proposed model with that of the state-of-the-art models, PSNR and SSIM are used, which are calculated based on the luminance channel of the YCbCr space of the recovered RGB results.

3) **Training Settings:** We set the number of MRFAMs as \(G = 5, 5, 20\) in the MRFAG structure for Stage 1, Stage 2...
and Stage 3, respectively. In each MRFAM, we set the number of residual blocks as $M = 5$. All the convolution layers have $C = 64$ filters except for those in the dilated convolution layer as shown in Fig. 5, where the convolution layer has $C = 4$ filters. We use $3 \times 3$ as the filter size for all convolution layers except for those in the dilated convolution based channel attention where the kernel sizes are $1 \times 1$, $3 \times 3$ and $5 \times 5$, which are shown in Fig. 5. Following previous works [24], [28], [45], we adopt the sub-pixel convolution [48] to upsample the LR features to HR. During training, we also augment the training dataset by randomly rotating by $90^\circ$, $180^\circ$, and $270^\circ$ and flipping horizontally [24], [28], [45]. In each training batch, LR images with patch size $48 \times 48$ are cropped as inputs. The proposed model is trained by the ADAM optimizer with a fixed initial learning rate of $10^{-4}$. The whole process is implemented in the PyTorch platform with 4 Nvidia TITAN TRX GPUs, each with 24GB of memory.

### B. Comparisons With State-of-the-Arts

In this section, we compare our HIPA with 17 state-of-the-art SISR methods: SRMD [58], DBPN [59], RDN [9], MRRN [60], RCAN [24], SRFBN [61], SAN [28], CSNLN [30], HAN [26], RDN [9], SRFDN [62], SAN [28], CSNLN [30], HAN [26], NSR [62], IGNN [63], RFANet [27], NLSN [31], SwinIR [37], TDPP [5], ELAN [51] and DGSMSwin [52]. Following previous works [24], [28], [37], we also perform self-ensemble on our HIPA to further improve its performance and dub it HIPA++.  

#### 1) Quantitative Comparison

Table I reports the quantitative comparisons between our method and 17 state-of-the-art SISR methods on five benchmark datasets for scale factor $2 \times$, $3 \times$ and $4 \times$. The best results are highlighted in red and the second best in blue. All the reported methods are proposed in recent 5 years and have achieved competitive results. Compared with these methods, our HIPA++ achieves the best results on multiple benchmarks for all scaling factors and surpasses most state-of-the-art methods in terms of PSNR and SSIM. Without using self-ensemble our network HIPA still achieves the best results on multiple benchmarks for all scale factors. It is noteworthy that our proposed HIPA is superior to SwinIR [37] and DGSMSwin [52], both of which are all hybrid architecture similar to HIPA. Specifically, the values of PSNR on the Urban100 dataset for scale factor $4 \times$ are

| Methods | Year | SRMD [58] | DBPN [59] | RDN [9] | MRRN [60] | RCAN [24] | SRFBN [61] | SAN [28] | CSNLN [30] | HAN [26] | RDN [9] | SRFDN [62] | SAN [28] | CSNLN [30] | HAN [26] | NSR [62] | IGNN [63] | RFANet [27] | NLSN [31] | SwinIR [37] | TDPP [5] | ELAN [51] | DGSMSwin [52] |
|---------|------|-----------|-----------|--------|----------|----------|-----------|---------|-----------|---------|--------|-----------|---------|-----------|---------|-------|---------|-----------|---------|-----------|--------|--------|---------|
| HIPA++  | 2023 | 38.38     | 34.35     | 32.35  | 30.93    | 34.20    | 32.16     | 34.07   | 32.03     | 34.27   | 32.50  | 34.20     | 33.50   | 32.00     | 34.35   | 33.40 | 33.90   | 32.00     | 33.45   | 32.00     | 34.35   | 33.40 | 33.90   |
| HIPA    | 2023 | 38.61     | 34.30     | 32.93  | 31.60    | 34.10    | 32.02     | 33.97   | 32.51     | 34.24   | 32.60  | 34.10     | 33.70   | 32.70     | 34.15   | 33.80 | 34.00   | 32.70     | 34.15   | 32.70     | 34.15   | 33.80 | 34.00   |
Fig. 7. Visual comparisons with state-of-the-art SISR methods for 4× SR on the B100, the Urban100 and the Manga109 datasets. Best viewed on screen.

Fig. 7. Visual comparisons with state-of-the-art SISR methods for 4× SR on the B100, the Urban100 and the Manga109 datasets. Best viewed on screen.

improved by 0.2 dB and 0.54 dB, respectively, compared to SwinIR and DGSM-Swin. The main reasons may lie in that i) the designed multi-stage progressive model not only can exploit features from different size patches but also can gradually recover the HR image from coarse to fine; and ii) the proposed MRFAG can let the network exhaustively mine local features contained in the original LR image from different receptive fields based on dilated convolution with different dilation factors.

2) Qualitative Comparison: In Fig. 7, we also visually illustrate the zoomed in comparison results with some state-of-the-art methods on several images from the test datasets. From the results, we find that our proposed HIPA can always obtain sharper results and recover more high-frequency textures and details, while most competing SISR models suffer from some unpleasant blurring artifacts. Take “img_109” in Manga109 shown in Fig. 7 as an example, existing methods obtain heavy blurring artifacts. The early proposed Bicubic fails to generate the clear structures. Although more recent methods, e.g. RCAN [24], SAN [28], CSNLN [30], RFAFNet [27], IGNN [63], NSLN [31] and SwinIR [37] can recover the main outlines, they fail to recover textures and details, and even generate some distorted and deformed textures. In contrast, our method effectively recovers textures through using the proposed HIPA and MRFAG.

3) Further Comparison: Table II compares the number of parameters, computational complexity and average running time comparisons for various SISR methods. Except for ESRT [53], HNCT [54] and Swin2SR-s [55], which are light weight SISRs, the other methods (including our HIPA) are classic performance-oriented SISRs. EDSR [45], RDN [9] and RCAN [24] are CNN-based SISR methods, while IPT [35], SwinIR [37], ESRT [53], HNCT [54] and Swin2SR-s [55] are state-of-the-art Transformer-based SISRs.
Compared to EDSR [45], RDN [9] and RCAN [24], HIPA not only has fewer parameters but also achieves a better PSNR value. Compared to two similar performance-oriented methods: IPT [35] and SwinIR [37], our method is more efficient in terms of computational complexity than light weight SISRs, they have more parameters and high computational complexity than light weight SISRs, they have better PSNR values because they focus more on performance.

C. Ablation Study

1) Ablation Study of HIPA Transformer: In Table III, we report the quantitative comparisons between the proposed HIPA Transformer with fixed-size patches and with multi-size patches by letting the network with and without partitioning the input LR image into a hierarchy of subblocks for scale factor \( \times 2, \times 3 \) and \( \times 4 \) on the Set14 and Urban100 datasets. From the PSNR results, we find that the HIPA Transformer using patches of different sizes outperforms that using fixed-size patches by a maximum of 0.12dB. The main reason is that the hierarchy of subblocks let the network learn one LR image from different sizes and improves the overall performance of the final results. Our result not only validates the effectiveness of the proposed multi-size patch but also further validates the effectiveness of the proposed hierarchical multi-stage structure.

Besides, in Table IV, we show the effects of the HIPA Transformer size on model performance. It can be found that the PSNR is positively correlated with the HIPA Transformer size. Even though the performance keeps increasing, the total number of parameters of the proposed HIPA Transformer grows also. To balance the performance and model size, we choose HIPA_M (PatS = 8, HeadNv = 8 and LayerN = 8) in the rest of the experiments.

2) Ablation Study of the Proposed APE: To validate the effectiveness of the proposed attention position encoding (APE), a comparison experiment between the proposed method using the previous position embedding (PE) [71], the condition position encoding (CPE) [56] and the proposed APE for scale \( \times 2, \times 3 \) and \( \times 4 \) on the MANGA109 dataset, we conduct a comparison experiment between the proposed MRFAG using SE attention and using the RCAB, which validates the effectiveness of the proposed MRFAG, as shown in Table VI, another comparison between the proposed MRFAG using SE attention and using the RCAB, which validates the effectiveness of the proposed MRFAG.

In addition, as shown in Table VII, another comparison between the proposed MRFAG using SE attention and using the RCAB, which validates the effectiveness of the proposed MRFAG.
the proposed dilated convolution based attention to validate the effectiveness of the proposed dilated convolution based attention. It can be found that the proposed dilated convolution based attention can improve PSNR value by a mean 0.043 dB on the B100, Urban100 and Manga109 datasets for scale factor \( \times 4 \) than that of the standard SE attention.

V. CONCLUSION

In this paper, we propose the Hierarchical Patch Transformer (HIPA) for accurate single image super resolution, which progressively recovers the high resolution image by partitioning the input into a hierarchy of patches. Specifically, a multi-stage progressive model is employed where the earlier stages use smaller patches as tokens and the final stage operates at full resolution. Our architecture is a cascade CNNs and Transformers for feature aggregation across multiple stages. In addition, we develop a novel attention-based position encoding scheme that allows the Transformer focus on the important tokens and easily process an input low resolution images with varying sizes. Besides, the proposed multi-receptive field attention module can enlarge the convolution receptive field from different branches. The quantitative and qualitative evaluations on different benchmark datasets demonstrate the effectiveness of the hierarchical patch partition over using fixed-size patches, as well as the superior performance of the proposed HIPA over most state-of-the-art methods in PSNR, SSIM and visual quality.

REFERENCES

[1] Y. Li, Y. Iwamoto, L. Lin, R. Xu, R. Tong, and Y. Chen, “VolumeNet: A lightweight parallel network for super-resolution of MR and CT volumetric data,” IEEE Trans. Image Process., vol. 30, pp. 4840–4854, 2021.
[2] W. Wen, W. Ren, Y. Shi, Y. Nie, J. Zhang, and X. Cao, “Video super-resolution via a spatio-temporal alignment network,” IEEE Trans. Image Process., vol. 31, pp. 1761–1773, 2022.
[3] M. R. Arefin et al., “Multi-image super-resolution for remote sensing using deep recurrent networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2020, pp. 816–825.
[4] K. Zhang, L. Van Gool, and R. Timofte, “Deep unfolding network for image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3214–3223.
[5] Q. Cai, J. Li, H. Li, Y. Yang, F. Wu, and D. Zhang, “TDPN: Texture and detail-preserving network for single image super-resolution,” IEEE Trans. Image Process., vol. 31, pp. 2375–2389, 2022.
[6] S. Anwar and N. Barnes, “Real image denoising with feature attention,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3155–3164.
[7] S. Nie et al., “A dual residual network with channel attention for image restoration,” in Proc. Eur. Conf. Comput. Vis., Aug. 2020, pp. 352–363.
[8] Q. Wang, Q. Gao, L. Wu, G. Sun, and L. Jiao, “Adversarial multi-path residual network for image super-resolution,” IEEE Trans. Image Process., vol. 30, pp. 6648–6658, 2021.
[9] 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.
[10] D. Song, C. Xu, X. Jia, Y. Chen, C. Xu, and Y. Wang, “Efficient residual dense block search for image super-resolution,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 12007–12014.
[11] X. Hu, M. A. Niael, A. Wong, M. Lamm, and P. Fieguth, “RUnet: A robust UNet architecture for image super-resolution,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 505–507.
[12] K. Prajapati et al., “Channel split convolutional neural network (ChSNet) for thermal image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2021, pp. 4363–4372.
Jinxing Li (Member, IEEE) received the B.Sc. degree from the Department of Automation, Hangzhou Dianzi University, Hangzhou, China, in 2012, the M.Sc. degree from the Department of Automation, Chongqing University, Chongqing, China, in 2015, and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2018. He is currently an Associate Professor with the Harbin Institute of Technology Shenzhen. His research interests include pattern recognition, deep learning, medical biometrics, and machine learning.

Jun Lyu received the B.Sc. degree in intelligence science and technology from Xidian University in 2013 and the Ph.D. degree in biomechanics and medical engineering from Peking University in 2018. From October 2017 to April 2018, she was a Visiting Ph.D. Student with the David Geffen School of Medicine, University of California at Los Angeles, Los Angeles. She is currently a Postdoctoral Fellow with the School of Nursing, The Hong Kong Polytechnic University. Her research interests include deep learning, and medical image processing and analysis.

Yee-Hong Yang (Life Senior Member, IEEE) received the B.Sc. degree (Hons.) from The University of Hong Kong, the M.Sc. degree from Simon Fraser University, and the Ph.D. degree from the University of Pittsburgh. He was a Faculty Member of the Department of Computer Science, University of Saskatchewan, from 1983 to 2001, and the Graduate Chair from 1999 to 2001. While there, in addition to department level committees, he also served on many college and university level committees. Since July 2001, he has been a Professor with the Department of Computing Science, University of Alberta. He was the Associate Chair (Graduate Studies) of the Department of Computing Science from 2003 to 2005. His research interests include computer graphics to computer vision, which include the physically-based animation of Newtonian and non-Newtonian fluids, texture analysis and synthesis, human body motion analysis and synthesis, computational photography, stereo and multiple view computer vision, and underwater imaging. He has published over 100 papers in international journals and conference proceedings in the areas of computer vision and graphics. He serves on the editorial board of the journal Pattern Recognition. In addition to serving as a reviewer to numerous international journals, conferences, and funding agencies, he has served on the program committees for many national and international conferences. In 2007, he was invited to serve on the expert review panel to evaluate computer science research in Finland.

Feng Wu (Fellow, IEEE) received the B.Sc. degree in electrical engineering from Xidian University, Xi’an, China, in 1992, and the M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1996 and 1999, respectively. He was a Principle Researcher and a Research Manager with Microsoft Research Asia, Beijing, China. He is currently a Professor with the University of Science and Technology of China, Hefei, China, where he is also the Dean of the School of Information Science and Technology. His research interests include image and video compression, media communication, and media analysis and synthesis.

David Zhang (Life Fellow, IEEE) received the Graduate degree in computer science from Peking University, the M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT) in 1982 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 1994. From 1986 to 1988, he was a Postdoctoral Fellow with Tsinghua University and then an Associate Professor with the Academia Sinica, Beijing. He has been a Chair Professor with The Hong Kong Polytechnic University, where he has been the Founding Director of the Biometrics Research Centre (UGC/CRC) supported by the Hong Kong SAR Government since 1998. Currently, he is the Presidential Chair Professor with The Chinese University of Hong Kong (Shenzhen). He has published over 20 monographs, more than 500 international journal articles and more than 40 patents from USA/Japan/Hong Kong/China. He has been continuously eight years listed as a Global Highly Cited Researcher in Engineering by Clarivate Analytics (2014–2021). He is also ranked about 85 with H-Index 123 at top 1000 scientists for international computer science and electronics. He is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, a fellow of the Royal Society of Canada and the Canadian Academy of Engineering, and an IAPR/AAIA Fellow.