IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 January 2024, accepted 13 February 2024, date of publication 22 February 2024, date of current version 1 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368436

RESEARCH ARTICLE

Image Super-Resolution With Unified-Window Attention

GUNHEE CHO AND YONG SUK CHOI
Department of Computer Science, Hanyang University, Seoul 04763, South Korea

Corresponding author: Yong Suk Choi (cys@hanyang.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant 2018R1A5A7059549 and Grant 2020R1A2C1014037; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP), funded by the Korean Government [Ministry of Science and Information and Communication Technology (MSIT)] under Grant 2020-0-01373.

ABSTRACT Recent studies on image super-resolution (SR) have focused on just expanding the receptive field. However, the insight from local attribution maps has enlightened the need for an improved exploitation of information within the receptive field. To address this, we propose a novel image super-resolution model, named Uniwin, that balances local and global interactions through unifying two types of window-based local attention mechanisms: shifted-window attention and sliding-window attention. Uniwin combines swift global context access with comprehensive local context capture. Our approach involves the initial sliding-window attention to collect comprehensive local pattern information, followed by the non-overlapping shifted-window attention to expand the receptive field for global interactions. Empirical evaluations demonstrate Uniwin's superiority over state-of-the-art models across five benchmark datasets. Specifically, in SR×2 task, our model achieved 0.25dB higher PSNR on the Set14 dataset, and 0.14dB higher PSNR on the Urban100 dataset than existing state-of-the-art model using a similar number of parameters. Additionally, our model achieved comparable performance to existing large-scale models with only 58% of the parameters. Ablation studies on sliding-window and shifted-window attention mechanisms reveal the critical importance of the context harmonization for image super-resolution. In conclusion, Uniwin represents a reasonable solution that effectively integrates global and local attention mechanisms, enhancing image super-resolution performance.

INDEX TERMS Image super-resolution, window-based local attention, shifted-window attention, sliding-window attention.

I. INTRODUCTION

In recent years, the creation and sharing of digital images have become increasingly common activities worldwide. However, the resolution of the images often can be limited due to various environmental factors, such as constraints imposed by the capturing device or limitations in data transmission. These low-resolution images lack fine visual details and compromise the overall sharpness and fidelity of the image. Image super-resolution (SR) aims to enhance the visual quality of such low-resolution images by restoring missing information to produce high-resolution images.

Over the past few years, convolutional neural networks (CNNs) have played a significant role in SR research, leveraging their ability to capture local patterns such as textures, borders, lines, and color variations in images [1], [2], [3], [4]. Nevertheless, the scalability of CNN's modeling capabilities for long-range dependencies is limited due to the inherent constraints of the parameter-dependent receptive field scaling and the content-independent local interactions. To address these limitations, the self-attention mechanism of Transformers [5] has emerged as a promising alternative to CNNs. Extensive studies have demonstrated the efficacy of Transformers' global interactions not only in the field of natural language processing but also in the field of computer vision [5], [6], [7], [8], [9], [10], [11], [12]. Within the

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

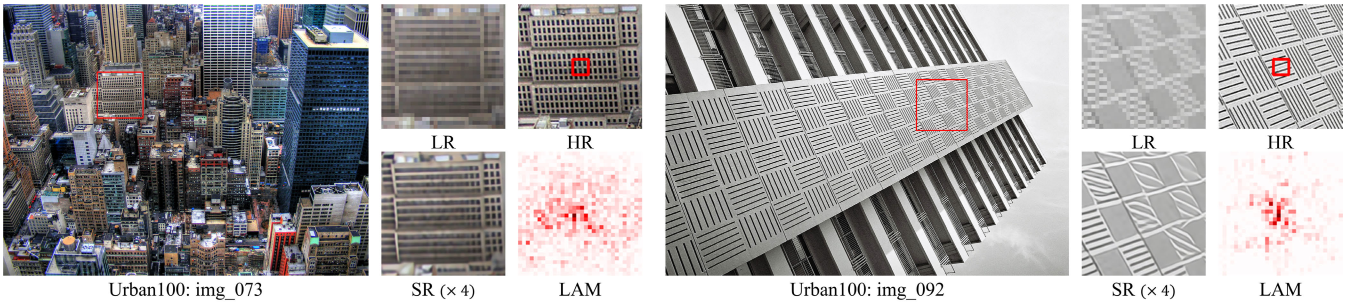


FIGURE 1. SR (×4) results from SwinIR and local attribution maps.

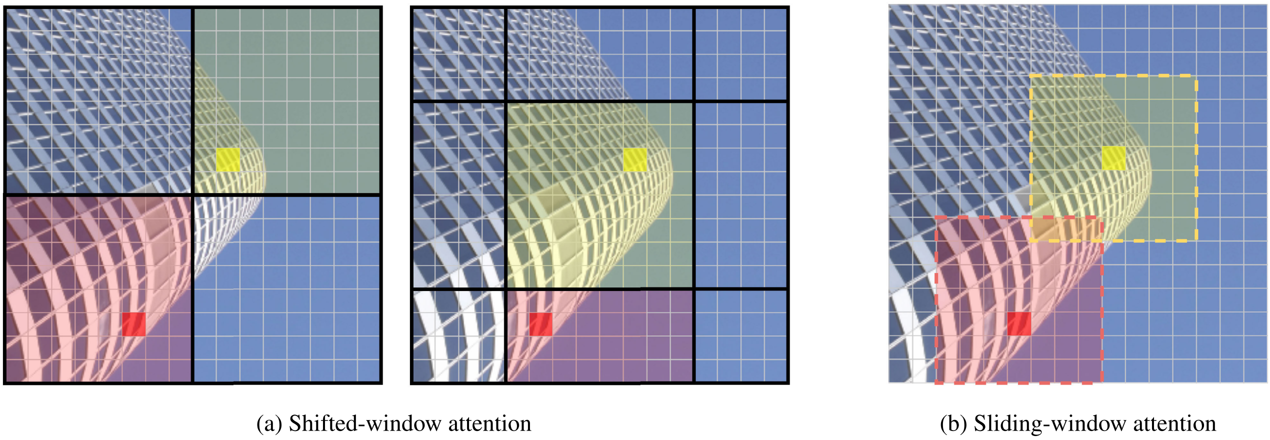


FIGURE 2. Comparison of attention span between shifted-window attention and sliding-window attention.

field of SR, several methods have successfully employed self-attention mechanisms, yielding impressive results [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. Notably, SwinIR [18] has achieved remarkable advancements by leveraging window-based local self-attention following Swin transformer [7].

The window-based self-attention provides quick access to a broad receptive field with modest computational costs. However, we noticed a wide receptive field does not always guarantee the improved SR performance by examining the local attribution maps (LAM) of SR model. LAM in Figure 1 represents pixels that participated in the reconstruction of the region marked with a small red box in high-resolution (HR) image. As shown in Figure 1, even though a SR model obtain information from a vast area to restore the boxed region, the resulting SR images exhibit blurriness or inaccurate textures. Based on these observations, we noticed that merely expanding the receptive field yield limited efficacy in learning features for SR.

These inaccurate restorations might be due to the non-overlapping window partitioning method which prevents pixels located at the edges of the window from gathering comprehensive local context. Consider Figure 2, which offers a simplified depiction of the non-overlapping window partitioning and the cyclic shifts of the shifted-window attention. Four local attention windows with bold borders are displayed in the left subplot, each containing 8 × 8 visual patches. The subsequent right subplot illustrates nine local attention windows resulting from the cyclic shifts. There are two highlighted patches in yellow and red. In the left subplot, the yellow patch contains a visual context of ‘building,’ yet most other patches in the same window (shaded in yellow) contain a visual context of ‘sky.’ This prevalence of ‘sky’ within the same window limits the yellow patch’s ability to obtain an appropriate local context through self-attention. Conversely, in the case of the red patch, self-attention produces more comprehensive information than the yellow one, as most patches in the same window (shaded in red) share a visual context of ‘building.’ Even with the shift in attention windows, the bias towards the local context within each attention window persists. The right subplot of Figure 2 illustrates that while the yellow patch can efficiently attend to a rich local context, the red patch can only access a weaker local context. The above observation emphasizes that SR model requires not only fast access to global information but also accurate collection of local information.

To address this issue, we considered adopting the overlapping window partitioning approach, in which the attention

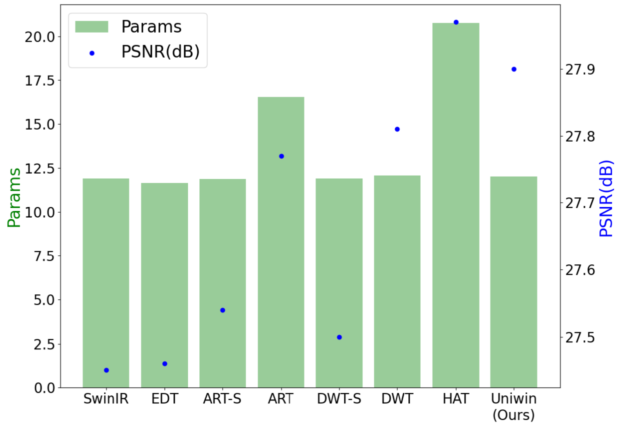


FIGURE 3. Comparison of parameters and PSNR scores of SR models. Green bars represent for the parameter count of each model, while blue dots indicate their corresponding PSNR of Urban100 (×4). Notably, Uniwin model demonstrates an impressive PSNR performance considering its number of parameters.

window is determined to include the nearest region of each patch rather than rigidly adhering to a fixed grid alignment. The attention method using the overlapping window partitioning is called a sliding-window attention [8], [10], [23]. It is illustrated in Figure 2b, where the placement of attention windows enables both the yellow and red patches to engage more effectively with their relevant local visual contexts. Compared to the shifted-window attention, the sliding-window attention is slower in accessing the global context, but more suitable for learning local context. Therefore, we combined these two types of window-based attention to train SR model.

In this study, we introduce a novel image super-resolution model named Uniwin, which effectively addresses the balance between global context and local context by integrating two distinct types of window-based local attention, namely shifted-window attention and sliding-window attention mechanisms. This integration allows for swift access to the broader context while ensuring the thorough capture of accurate local patterns. To describe the methodology, our Uniwin model is outlined as follows: Initially, all pixels capture precise local pattern information by implementing the sliding-window attention mechanism. Subsequently, the receptive field is rapidly expanded via the non-overlapping shifted-window attention strategy. This strategy empowers Uniwin to discern analogous patterns over extended distances, thereby enhancing its performance in image super-resolution tasks.

It is noteworthy that Uniwin incurs minimal additional computational overhead compared to the state-of-the-art models. Figure 3 presents a chart that displays the model parameters against PSNR performance. Uniwin not only achieves a superior Peak Signal-to-Noise Ratio (PSNR) with a moderate number of parameters but also surpasses the performance of models with a similar parameter count. Significantly, it approaches the performance of HAT [15], which operates with a much larger set of parameters. This remarkable efficiency is attributed to Uniwin's capability to capture both local and global contexts effectively.

The primary contributions of this work are summarized as follows:

- We introduce Uniwin, a new SR model designed to allow two types of window attention operations to work complementary to each other.
- Through empirical evaluations, we validate Uniwin's superior performance over state-of-the-art models with comparable parameters across five benchmark datasets.
- By conducting extensive ablation studies on sliding- and shifted-window attention, we illuminated the influence of each window attention mechanism on image super-resolution process.

The subsequent sections are structured as follows: Section II provides a concise overview of the related research. In Section III, our proposed Uniwin method is described. Section IV shows the experimental results and corresponding analyses. Conclusive insights are presented in Section V, summarizing the essential findings and implications.

II. RELATED WORK

A. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) are the foundational architecture for numerous image super-resolution (SR) models, capitalizing on their intrinsic ability to capture local patterns. Notably, SRCNN [1] was one of the pioneering works incorporating deep CNNs for image super-resolution. Subsequently, CNN-based advancements, including notable contributions such as VDSR [2], EDSR [3], RCAN [4], NLRN [24], and NLSA [25], have sought to enhance model representation. In models like VDSR and EDSR [2], [3], the residual learning was integrated to foster deeper network structures. Meanwhile, attention mechanisms within the CNN paradigm, such as channel attention [4], [26] and non-local attention [24], [25], have been improved in capturing intricate patterns. Noteworthy is the proposal in RCAN [4], introducing a deep network architecture by combining residual-in-residual design with channel attention.

Despite these advancements, the scalability of CNNs in modeling long-range dependencies remains limited by the complication of parameter-dependent receptive field expansion and contextually independent local interactions. Recognizing these limitations of CNNs, the emergence of Transformer-based self-attention mechanisms [5] has provided a compelling alternative. This paradigm shift has led to the integration of self-attention into the domain of image SR [14], [15], [17], [18], [19], [20], [21], [22], [27].

B. VISION TRANSFORMERS

Approaches such as Uformer [20] and Restormer [21] have proposed U-shaped transformer models infused with depthwise convolutions. IPT [14] has proposed a visual backbone model based on the standard Transformer [5], pre-trained across various restoration tasks. SwinIR [18] has proposed a deep feature extraction architecture through an iterative assembly of residual Swin Transform blocks, which has become a robust baseline for image restoration tasks. EDT [17] has offered an exhaustive exploration of the effects of pre-training across multi-related tasks. Recent contributions, including ART [22] and DWT [19], have proposed rapid receptive field expansion strategies by harnessing sparse attention through dilation.

1) SWINIR

What makes SwinIR different from previous transformer-based SR models is that it eliminates the downsampling process and significantly reduces the patch size, allowing the attention mechanism to effectively handle low-level vision tasks such as SR. Although a smaller patch size allows for more detailed local feature representation, it comes at the cost of increased computational complexity. To mitigate this, SwinIR partitions the input image into non-overlapping local windows and independently applies self-attention within these windows, following Swin Transformer's shifted-window attention mechanism [7]. Since the partitioned windows do not overlap, after applying self-attention to each window, SwinIR repeatedly performs cyclic shifts of the windows and reapplies self-attention to enable cross-window interactions. These shifted-window attentions are adequate for modeling long-range dependencies, as the receptive field grows each time the attention window shifts, allowing quick access to global information.

Despite the benefits of shifted-window attention, SwinIR frequently encounters challenges in accurately restoring intricate local patterns, even those that reveal repetition, as demonstrated in Figure 1. Prior research has asserted that the primary factor contributing to this inaccurate restoration of local patterns is the absence of comprehensive global information [15], [19], [22]. Consequently, numerous investigations have concentrated on refining approaches to rapidly expand the receptive field of SR models, aiming to enhance access to global information. Nevertheless, the insight from Local Attribution Maps (LAM) [28] suggests that the receptive fields of existing SR models are already sufficiently extensive. Hence, a mere expansion of the receptive field and faster access to global information yield limited efficacy in addressing the issue. Instead, the paramount importance lies in effectively exploiting information residing within the receptive field. To illustrate this point, we present a visual representation of the pixels within the input image contributing to the restoration of the red box in the high-resolution (HR) image (as indicated by the LAM in Figure 1). Notably, in the case of SwinIR, the interaction already occupies an area wide enough to contain all essential information necessary for the accurate restoration of local patterns.

Although these Transformer-based models have shown noteworthy performance, it has become evident that rapid access to the global context alone is not sufficient for acquiring the intricate feature representations essential for SR tasks. Notably, the adoption of non-overlapping partitioning in contemporary models [15], [17], [18], [19], [22], [27] that leverage window-based local attention mechanisms has revealed a limitation—difficulties in gathering accurate local contexts.

C. WINDOW-BASED LOCAL ATTENTION

Accordingly, we have investigated previous works to uncover a more suitable paradigm for a window-based local attention mechanism beneficial for super-resolution tasks. Approaches such as SASA [8] and HaloNet [10] have proposed substituting conventional spatial convolutions with localized self-attentions, mimicking the behavior of convolution kernels with zero padding. VOLO [29] has introduced progressive tokenization, enriching token representation with the contextual information of surrounding tokens. NAT [23] has restricted the scope of dot-product attention within localized regions, thus constraining the receptive field of each query token within a fixed-size neighborhood. This approach ensures all pixels maintain a uniform attention span, thereby mitigating the reduction in attention reach that corner pixels often face in zero-padded alternatives like SASA [8]. These previous studies have shown that sliding-window attention with overlapping windows improves the local context representation of vision models. Therefore, our Uniwin model integrated two types of window-based attention: shifted-window attention and sliding-window attention.

III. METHOD

A. NETWORK ARCHITECTURE

As illustrated in Figure 4, the overall architecture of Uniwin consists of three stages: shallow feature extraction, deep feature extraction, and image upscale. The role of each stage is as follows. The shallow feature extraction stage employs simple operations to preserve the low-frequency components inherent in the input image, which are essential for restoring the rough contours of high-resolution images. The deep feature extraction stage extracts features for the sophisticated restoration of high-frequency components by deeply stacked layers within a residual-in-residual framework. The image upscale stage synthesizes a high-resolution image leveraging the features extracted from the previous stages.

We will explain the specific operations within each stage to provide a more comprehensive understanding. First, in the shallow feature extraction stage, an input image is fed into a single convolutional layer, which projects the input image into a high-dimensional feature space. It can be described as follows. For a given low-resolution input $I_{\text{LR}} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$, where H, W, and C_{in} are the width, the height, and the channel of an input image, respectively, the shallow feature

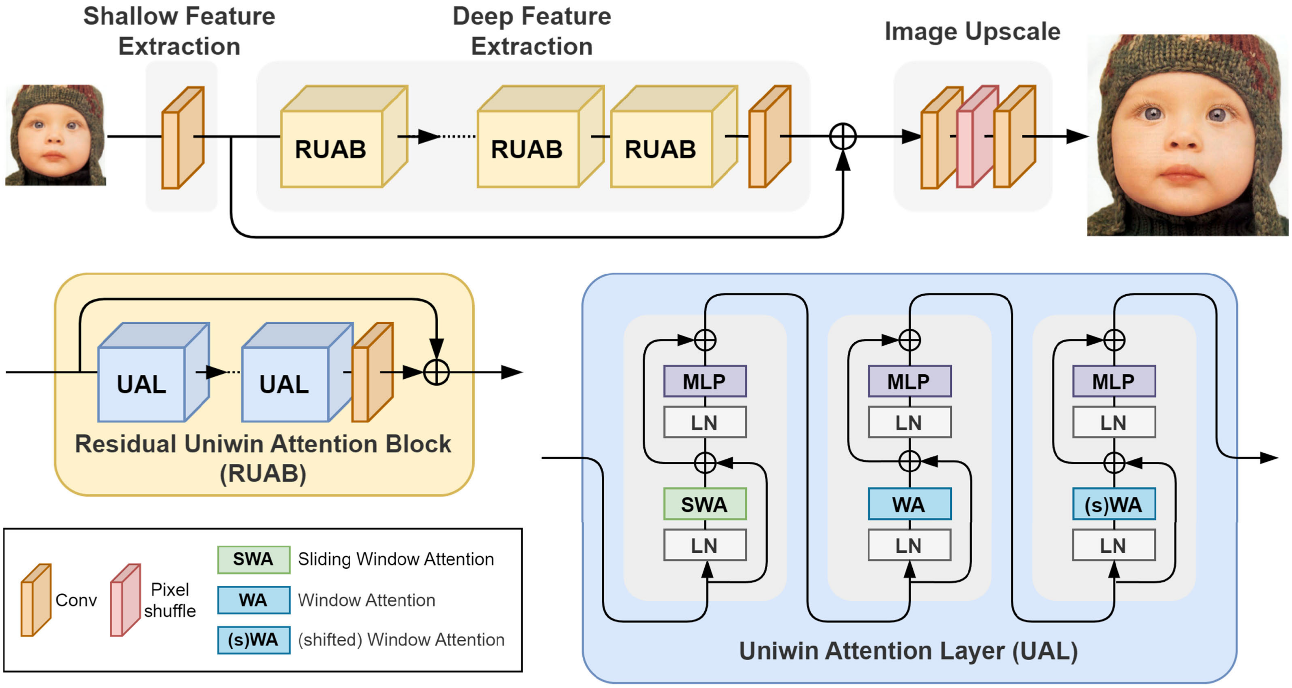


FIGURE 4. An illustration of the overall architecture of Uniwin.

$F_0 \in \mathbb{R}^{H\times W\times C}$, is obtained as,

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

where C is the feature dimension and $H_{\mathrm{SF}}(\cdot)$ is a single 3×3 convolution layer. A convolutional layer is a simple operation yet excellent at capturing local information. It contributes to preserving the low-frequency components of an input image, which guides a model to stable optimization.

Next, the shallow feature F_0 is then passed through the deep feature extraction stage to uncover the richer feature representation, obtaining the deep feature $F_{\mathrm{DF}} \in \mathbb{R}^{H\times W\times C}$ as,

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

where $H_{\mathrm{DF}}(\cdot)$ denotes the deep feature extraction process consisting of N residual Uniwin attention blocks(RUAB), which performs the unified window-based self-attention and a convolution. It can be expressed as,

$$\begin{aligned} F_i &= H_{\mathrm{RUAB}_i}(F_{i-1}), \quad i = 1, 2, \ldots, N, \\ F_{\mathrm{DF}} &= H_{\mathrm{Conv}}(F_N), \end{aligned} \tag{3}$$

where $H_{\mathrm{RUAB}_i}(\cdot)$ is the i-th RUAB and H_{Conv} denotes a single 3×3 convolution layer. For integrating the shallow and deep features, inserting a convolutional layer at the end of the feature extraction process proves beneficial [18]. This insertion introduces the inductive bias inherent in convolutional operations to self-attention operation.

Finally, in the image upscale stage, the deep feature F_{DF} and the shallow feature F_0 are fused by a global residual connection to reconstruct the high-resolution image I_{SR} as,

$$I_{\mathrm{SR}} = H_{\mathrm{IU}}(F_0 + F_{\mathrm{DF}}), \tag{4}$$

where H_{IU} denotes the image upscale process by the sub-pixel convolution layers [30].

The parameters of our Uniwin model are optimized through the minimization of L_1 pixel loss. The loss function $\mathcal{L}$ is defined as the sum of absolute differences between the reconstructed image I_{SR} and the corresponding ground-truth high-resolution image I_{HR}, which can be written as,

$$\mathcal{L} = \|I_{\mathrm{SR}} - I_{\mathrm{HR}}\|_1, \tag{5}$$

where $\|\cdot\|_1$ denotes the L_1 norm.

B. RESIDUAL UNIWIN ATTENTION BLOCK

The residual Uniwin attention block (RUAB) consists of M Uniwin attention layers (UAL) and a convolutional layer. The operation of RUAB is formulated as follows. For a given input feature F_{i-1} of the i-th RUAB, the intermediate features $F_{i-1,0}, F_{i-1,1}, \ldots, F_{i-1,M}$ and the output feature F_i can be obtained as,

$$\begin{aligned} F_{i-1,0} &= F_{i-1}, \\ F_{i-1,j} &= H_{\mathrm{UAL}_{i,j}}(F_{i-1,j-1}), \quad j = 1, 2, \ldots, M, \\ F_i &= H_{\mathrm{Conv}_i}(F_{i-1,M}) + F_{i-1,0}, \end{aligned} \tag{6}$$

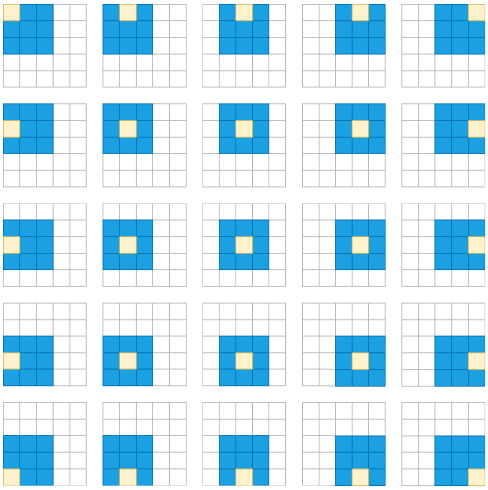


FIGURE 5. Attention patterns of the sliding-window attention.

where F_{i-1} and F_i denotes the input and the output feature of the i-th RUAB, and $H_{\mathrm{UAL}_{i,j}}(\cdot)$ indicates the j-th UAL in the i-th RUAB.

Within the j-th UAL, the feature flows sequentially through sliding-window attention (SWA), non-overlapping window attention (WA), and corresponding shifted-window attention ((s)WA). Recognizing that using SWA with overlapping windows enhances how vision models understand the local context, the features first pass through the SWA layer to capture local details effectively. After this initial step of engaging with the local context, the features then move through WA and (s)WA layers. This sequence helps to quickly expand the area of attention, allowing the model to interact with the global context. It can be written as,

$$H_{\mathrm{UAL}_{i,j}}(F_{i-1,j-1}) = H_{\mathrm{(s)WA}_{i,j}}(H_{\mathrm{WA}_{i,j}}(H_{\mathrm{SWA}_{i,j}}(F_{i-1,j-1}))), \tag{7}$$

where $H_{\mathrm{(s)WA}_{i,j}}(\cdot)$, $H_{\mathrm{WA}_{i,j}}(\cdot)$, and $H_{\mathrm{SWA}_{i,j}}(\cdot)$ denote (s)WA layer, WA layer and SWA layer, respectively, in the j-th UAL in the i-th RUAB. For the actual implementation, we made a minor modification to Neighborhood Attention [23] for the SWA in our Uniwin. For the design of WA and (s)WA pairs, we followed the approach of SwinIR [18].

We will provide a more detailed description of the three window-based attention operations that constitute UAL. Initially, a sliding-window attention is employed to collect sufficient local information. Subsequently, the interaction with a broader context is facilitated by applying a non-overlapping window attention layer and a corresponding shifted-window attention layer inspired by Swin Transformers [7]. These paired attention layers effectively model long-range dependency by quickly enhancing the receptive field.

In the sliding-window attention layer, for a given input feature $X \in \mathbb{R}^{n\times d}$, which represents n token vectors of dimension d, the query, key, and value tokens are obtained through linear projections, which is denoted by Q, K, and V, respectively. Regarding the sliding-window attention, the window partitioning does not adhere to a rigid grid structure. Instead, every query token is associated with a unique $\rho \times \rho$ attention window, with the token positioned at the center of the window. Inspired by the window partitioning method introduced by NAT [23], the window placement strategy is adopted. Precisely, when the query token resides at the image's border, the window is adjusted to contain as many neighboring tokens as possible, as shown in Figure 5.

Once the attention window is decided, the attention weights are calculated by the dot product similarities between the query token and key tokens within the window. For a formal description, given the k-th input query token Q_k and key tokens $K_{1:\rho\times\rho}$ that exists within a $\rho \times \rho$ window, the attention weight A_k^ρ can be written as,

$$A_k^\rho = \begin{bmatrix} Q_k {K_1}^T + B_1 \\ Q_k {K_2}^T + B_2 \\ \vdots \\ Q_k {K_{\rho\times\rho}}^T + B_{\rho\times\rho} \end{bmatrix}, \quad k = 1, 2, \ldots, n, \tag{8}$$

where $B_{1:\rho\times\rho}$ indicates the relative position bias within the $\rho\times\rho$ window. Then, the value, V_k^ρ, can be defined as a matrix whose rows are the k-th token's $\rho \times \rho$ nearest neighboring value:

$$V_k^\rho = \left[{V_1}^T\ {V_2}^T\ \cdots\ {V_{\rho\times\rho}}^T\right]^T. \tag{9}$$

Finally, the result of the sliding-window attention for the k-th token with window size of $\rho \times \rho$ is formulated as,

$$\mathrm{SWA}_k^\rho = \mathrm{softmax}(\frac{A_k^\rho}{\sqrt{d}})V_k^\rho, \tag{10}$$

where d indicates the scaling parameter.

Although the SWA layer is similar to Neighborhood attention (NA) from NAT [23], there are important differences in detailed design considerations. First of all, since NA was designed with the intention of being a general visual backbone, it performs hierarchical down-sampling through convolutional layers during the feature extraction process to reduce the spatial size. It is a reasonable design choice for performing high-level vision tasks, but in low-level vision tasks such as SR, it causes information loss and reduces performance. Second, NAT uses 4 × 4 patch embedding, but since SWA considers the SR task, the spatial size of a visual patch is extremely reduced to 1 × 1 and the patch dimension is increased.

We followed the pipeline of a standard transformer at each layer of attention [5]. So, after the sliding-window attention operation, the features pass through an MLP layer with GELU activation. The residual connections are adopted, and the LayerNorm (LN) layers are inserted before every SWA, WA, (s)WA, and MLP layer.

TABLE 1. Quantitative comparison with the state-of-the-art models with comparable parameters for image super-resolution on 5 benchmark datasets. Best and Second-best results are in red and blue colors, respectively.

Method	Scale	Training Dataset	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR[3]	×2	DIV2K	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
RCAN[4]	×2	DIV2K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
NLSA[25]	×2	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
SwinIR[18]	×2	DF2K	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
EDT[17]	×2	DF2K	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
ART-S[22]	×2	DF2K	38.48	0.9625	34.50	0.9258	32.53	0.9043	34.02	0.9437	40.11	0.9804
ART[22]	×2	DF2K	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
DWT-S[19]	×2	DF2K	38.40	0.9628	34.44	0.9254	32.53	0.9048	33.77	0.9419	39.88	0.9798
DWT[19]	×2	DF2K	38.49	0.9632	34.56	0.9265	32.56	0.9054	34.14	0.9444	40.01	0.9802
Uniwin	×2	DF2K	38.57	0.9635	34.84	0.9274	32.63	0.9060	34.44	0.9469	40.28	0.9810
EDSR[3]	×3	DIV2K	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
RCAN[4]	×3	DIV2K	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
NLSA[25]	×3	DIV2K	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
SwinIR[18]	×3	DF2K	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
EDT[17]	×3	DF2K	34.97	0.9316	30.89	0.8527	29.44	0.8142	29.72	0.8814	35.13	0.9534
ART-S[22]	×3	DF2K	34.98	0.9318	30.94	0.8530	29.45	0.8146	29.86	0.8830	35.22	0.9539
ART[22]	×3	DF2K	35.07	0.9325	30.99	0.8540	29.51	0.8159	30.10	0.8871	35.39	0.9548
DWT-S[19]	×3	DF2K	34.94	0.9320	30.91	0.8530	29.45	0.8159	29.73	0.8806	35.10	0.9533
DWT[19]	×3	DF2K	35.00	0.9327	30.97	0.8541	29.49	0.8170	30.07	0.8860	35.27	0.9542
Uniwin	×3	DF2K	35.01	0.9327	31.04	0.8555	29.51	0.8175	30.17	0.8891	35.31	0.9546
EDSR[3]	×4	DIV2K	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
RCAN[4]	×4	DIV2K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
NLSA[25]	×4	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
SwinIR[18]	×4	DF2K	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
EDT[17]	×4	DF2K	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
ART-S[22]	×4	DF2K	32.86	0.9029	29.09	0.7942	27.91	0.7489	27.54	0.8261	32.13	0.9263
ART[22]	×4	DF2K	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
DWT-S[19]	×4	DF2K	32.88	0.9046	29.06	0.7947	27.91	0.7507	27.50	0.8253	32.03	0.9253
DWT[19]	×4	DF2K	32.92	0.9055	29.12	0.7961	27.94	0.7519	27.81	0.8324	32.20	0.9274
Uniwin	×4	DF2K	32.76	0.9051	29.18	0.7973	27.94	0.7524	27.90	0.8362	32.22	0.9278

To summarize, the entire process of UAL can be written as,

$$
\begin{aligned}
X_{\mathrm{SWA}} &= \mathrm{SWA}(\mathrm{LN}(X_{\mathrm{in}})) + X_{\mathrm{in}}\\
X_{\mathrm{MLP_{SWA}}} &= \mathrm{MLP}(\mathrm{LN}(X_{\mathrm{SWA}}) + X_{\mathrm{SWA}}\\
X_{\mathrm{WA}} &= \mathrm{WA}(\mathrm{LN}(X_{\mathrm{MLP_{SWA}}})) + X_{\mathrm{MLP_{SWA}}}\\
X_{\mathrm{MLP_{WA}}} &= \mathrm{MLP}(\mathrm{LN}(X_{\mathrm{WA}}) + X_{\mathrm{WA}}\\
X_{\mathrm{(s)WA}} &= \mathrm{WA}(\mathrm{LN}(X_{\mathrm{MLP_{WA}}})) + X_{\mathrm{MLP_{WA}}}\\
X_{\mathrm{out}} &= \mathrm{MLP}(\mathrm{LN}(X_{\mathrm{(s)WA}}) + X_{\mathrm{(s)WA}}. \qquad (11)
\end{aligned}
$$

We followed the implementation of SwinIR's Swin transformer layer (STL) [18] for the paired WA and (s)WA layer. However, using only shifted-window types for feature extraction limits its direct interaction with the local context, despite the strengths of Swin transformer—structural robustness and long-range dependency modeling. To address this limitation, Uniwin first ensures the capture of sufficient local information by integrating the SWA layer. Subsequently, by introducing WA and (s)WA layer to facilitate the global interaction, Uniwin harmonizes the advantages of both local and global contexts.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

We set our model parameters to change minimally to ensure a fair comparison with other SR models. Following SwinIR, the number of RUAB, channels, and attention heads are 6, 180, and 6, respectively. The number of UAR is two, and the window size of SWA and WA is 9 and 16, respectively.

1) DATASETS AND EVALUATION METRICS

Following previous works [15], [17], [18], [19], [22], we used DF2K [3], [31] as training data and evaluated our model's performance on five benchmark datasets: Set5 [32], Set14 [33], BSD100 [34], Urban100 [35], and Manga109 [36]. For each of the five datasets, we measured the average PSNR and SSIM scores for the Y channel of the YCbCr space.

2) PREPROCESSING

To minimize the IO bottlenecks during training, we preprocessed all the high-resolution (HR) images of the training data by cropping them into 480 × 480-sized sub-images with half-overlapping. Also, all low-resolution (LR) images were cropped according to each scale factor (×2, ×3, ×4) to match the size of the HR sub-images.

3) TRAINING

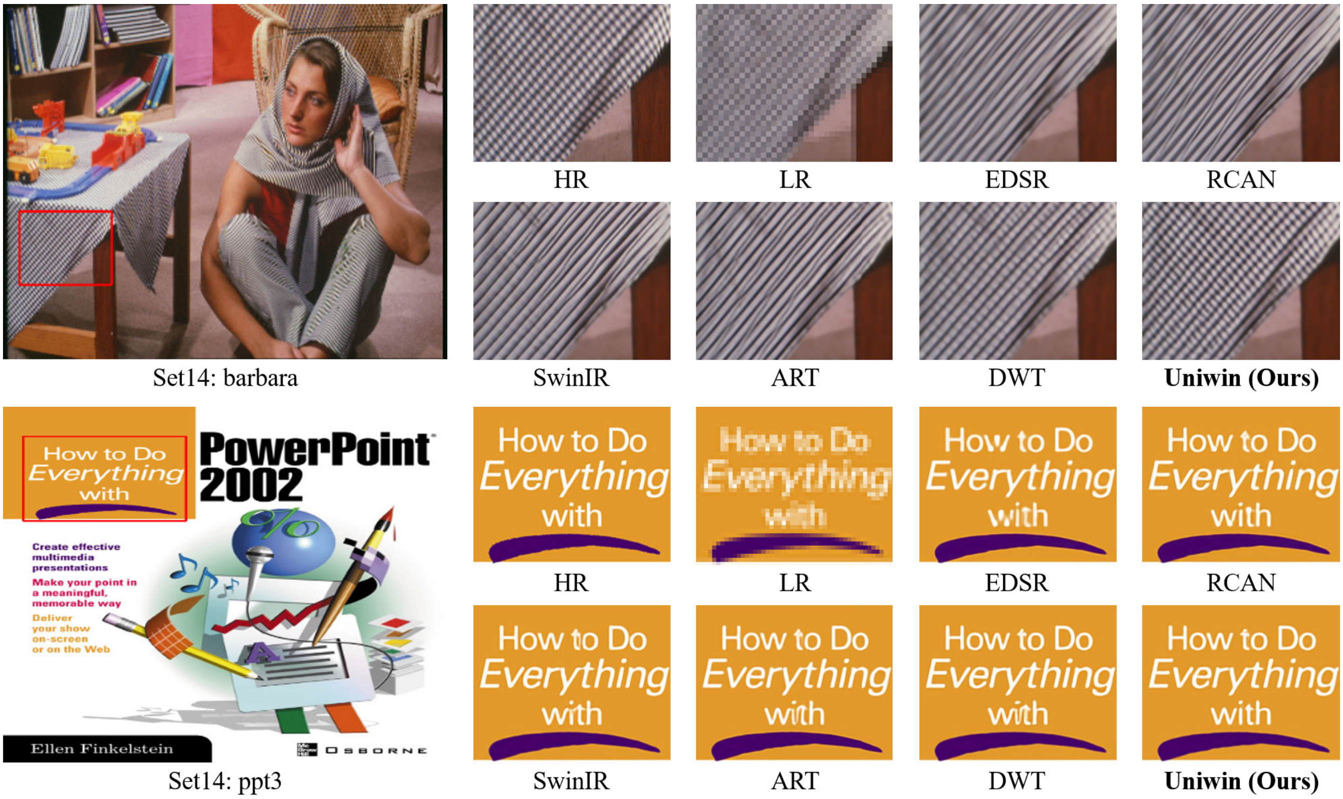
During training, all images are randomly cropped into 64 × 64 sized patches, and then data augmentation is applied through random rotation of 90°, 180°, and 270° and random horizontal flip. Following SwinIR [18], the training batch is

TABLE 2. Quantitative (PSNR(dB)/SSIM) comparison (×4 SR) with the state-of-the-art models with similar parameter counts. Best and Second-best results are in red and blue colors, respectively.

Method	#Params(M)	Mult-Adds(G)	Set5	Set14	BSD100	Urban100	Manga109
SwinIR[18]	11.90	316	32.92/0.9044	29.09/0.7950	27.92/0.7489	27.45/0.8254	32.03/0.9260
EDT[17]	11.63	354	32.82/0.9031	29.09/0.7939	27.91/0.7483	27.46/0.8246	32.05/0.9254
ART-S[22]	11.87	319	32.86/0.9029	29.09/0.7942	27.91/0.7489	27.54/0.8261	32.13/0.9263
ART[22]	16.55	448	33.04/0.9051	29.16/0.7958	27.97/0.7510	27.77/0.8321	32.31/0.9283
DWT-S[19]	11.90	316	32.88/0.9046	29.06/0.7947	27.91/0.7507	27.50/0.8253	32.03/0.9253
DWT[19]	12.06	319	32.92/0.9055	29.12/0.7961	27.94/0.7519	27.81/0.8324	32.20/0.9274
Uniwin	12.02	381	32.76/0.9051	29.18/0.7973	27.94/0.7524	27.90/0.8362	32.22/0.9278

TABLE 3. PSNR(dB)/SSIM comparison with the large scale state-of-the-art method. Best results are in red color.

Method	#Params(M)	Scale	Set5	Set14	BSD100	Urban100	Manga109
HAT[15]	20.62	×2	38.63/0.9630	34.86/0.9274	32.62/0.9053	34.45/0.9466	40.26/0.9809
Uniwin	11.87	×2	38.57/0.9635	34.84/0.9274	32.63/0.9060	34.44/0.9469	40.28/0.9810
HAT[15]	20.81	×3	35.07/0.9329	31.08/0.8555	29.54/0.8167	30.23/0.8896	35.53/0.9552
Uniwin	12.05	×3	35.01/0.9327	31.04/0.8555	29.51/0.8175	30.17/0.8891	35.31/0.9546
HAT[15]	20.77	×4	33.04/0.9056	29.23/0.7973	28.00/0.7517	27.97/0.8368	32.48/0.9292
Uniwin	12.02	×4	32.76/0.9051	29.18/0.7973	27.94/0.7524	27.90/0.8362	32.22/0.9278


Set14: barbara

HR LR EDSR RCAN

SwinIR ART DWT **Uniwin (Ours)**

Set14: ppt3

HR LR EDSR RCAN

SwinIR ART DWT **Uniwin (Ours)**

PowerPoint 2002
How to Do Everything with
Create effective multimedia presentations
Make your point in a meaningful, memorable way
Deliver your show on-screen or on the Web
Ellen Finkelstein
OSBORNE

FIGURE 6. Visual comparison (×4) with state-of-the-art SR methods on Set14 datasets.

set to 32, and the total training iteration is set to 500k. Also, the learning rate is initialized as 2×10^{-4} and scheduled to be halved when the training iteration reaches 250K, 400K, 450K, and 475K. For ×3 and ×4 SR, we initialize the model with pre-trained ×2 SR model weights and reduce both the iterations for each learning rate decay and total iterations by half. We use ADAM [37] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ to optimize our model. Uniwin is implemented on PyTorch and trained using 4 NVIDIA RTX A5000 GPUs.

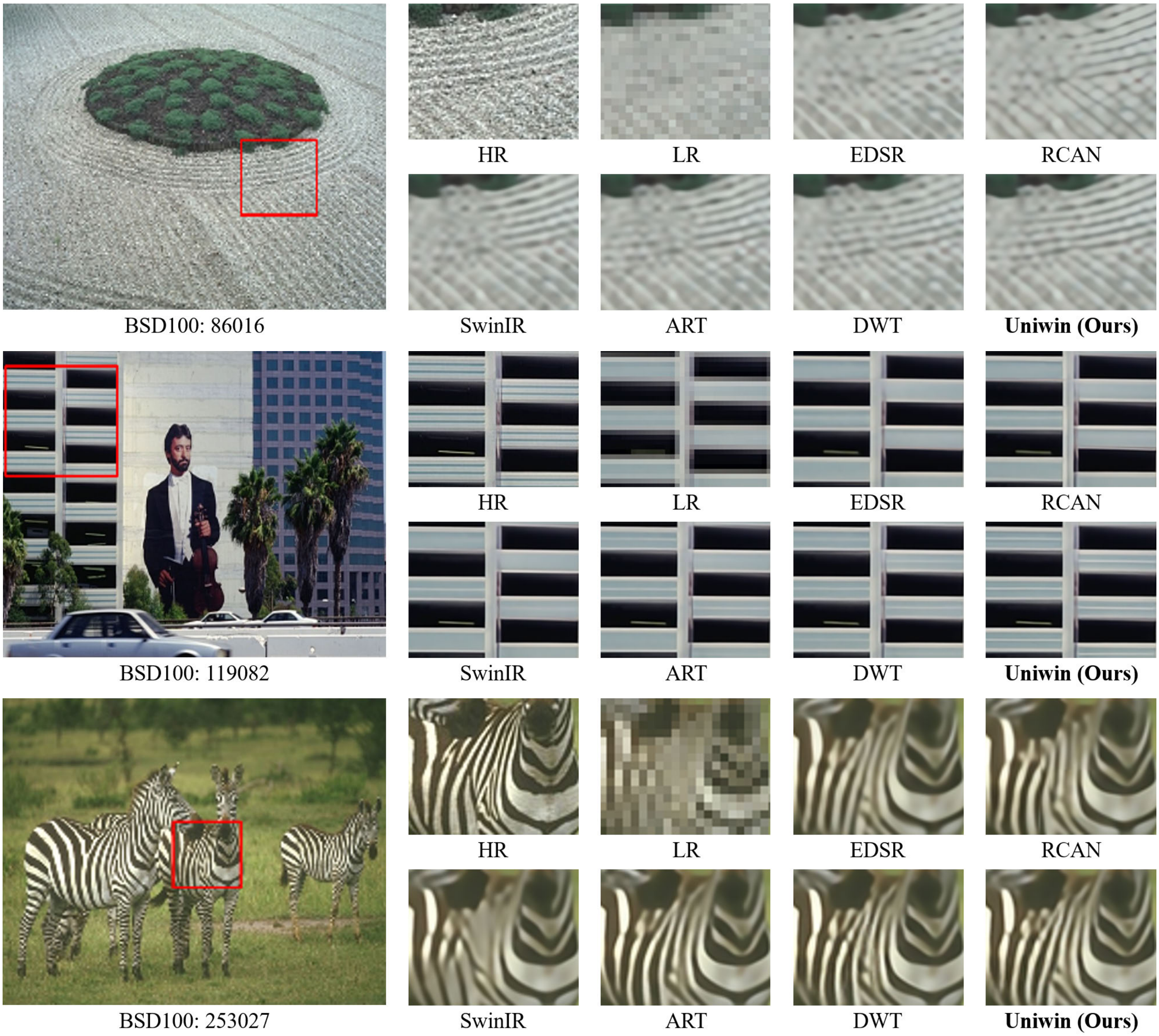


FIGURE 7. Visual comparison (×4) with state-of-the-art SR methods on BSD100 dataset.

B. RESULTS

1) QUANTITATIVE COMPARISON

Table 1 provides a comprehensive quantitative comparison between Uniwin and several state-of-the-art models, including EDSR [3], RCAN [4], NLSA [25], SwinIR [18], EDT [17], ART [22], and DWT [19]. We have marked the best-performing models in red and the second-best in blue for each dataset. Across the spectrum of benchmark datasets, Uniwin consistently demonstrates superior performance, as evident from the data in Table 1. This distinction is particularly evident in the Urban100 dataset. The impressive performance can be attributed to the inherent characteristics of Urban100 dataset, which include numerous images with repetitive local patterns. Uniwin leverages the advantage to capture local context effectively.

We further compared parameters and Mult-Adds for transformer-based networks, as detailed in Table 2. The Mult-Adds calculations assume a 3 × 160 × 160 input size for ×4 SR. As highlighted in Table 2, Uniwin exhibits superior performance compared to models with similar parameter counts. Remarkably, despite possessing 73% of the parameters of ART, Uniwin delivers comparable or slightly superior performance across multiple benchmarks.

Our performance evaluation also includes HAT [15], a state-of-the-art model with considerably more parameters than Uniwin. This information is summarized in Table 3,

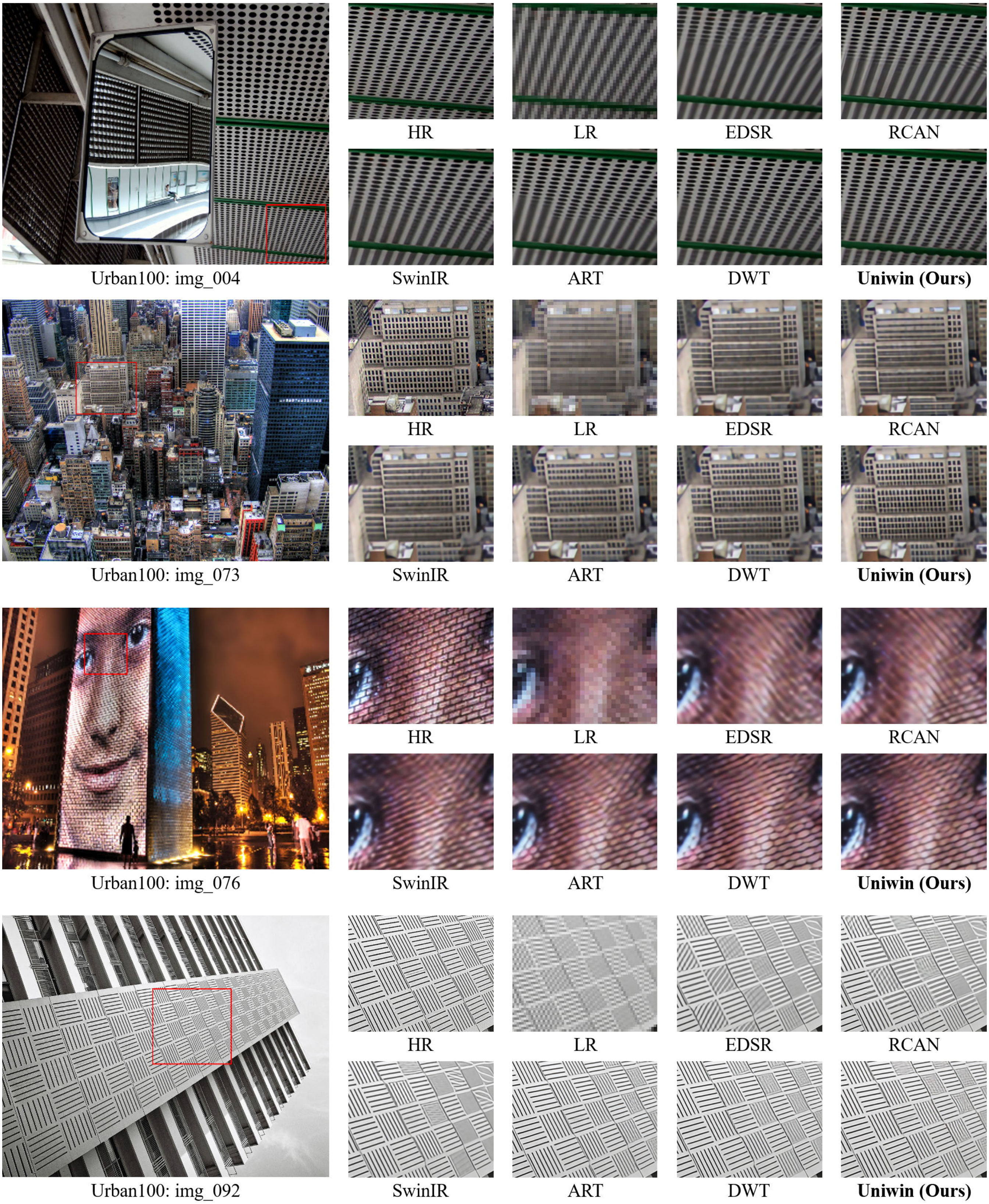


FIGURE 8. Visual comparison (×4) with state-of-the-art SR methods on Urban100 dataset.

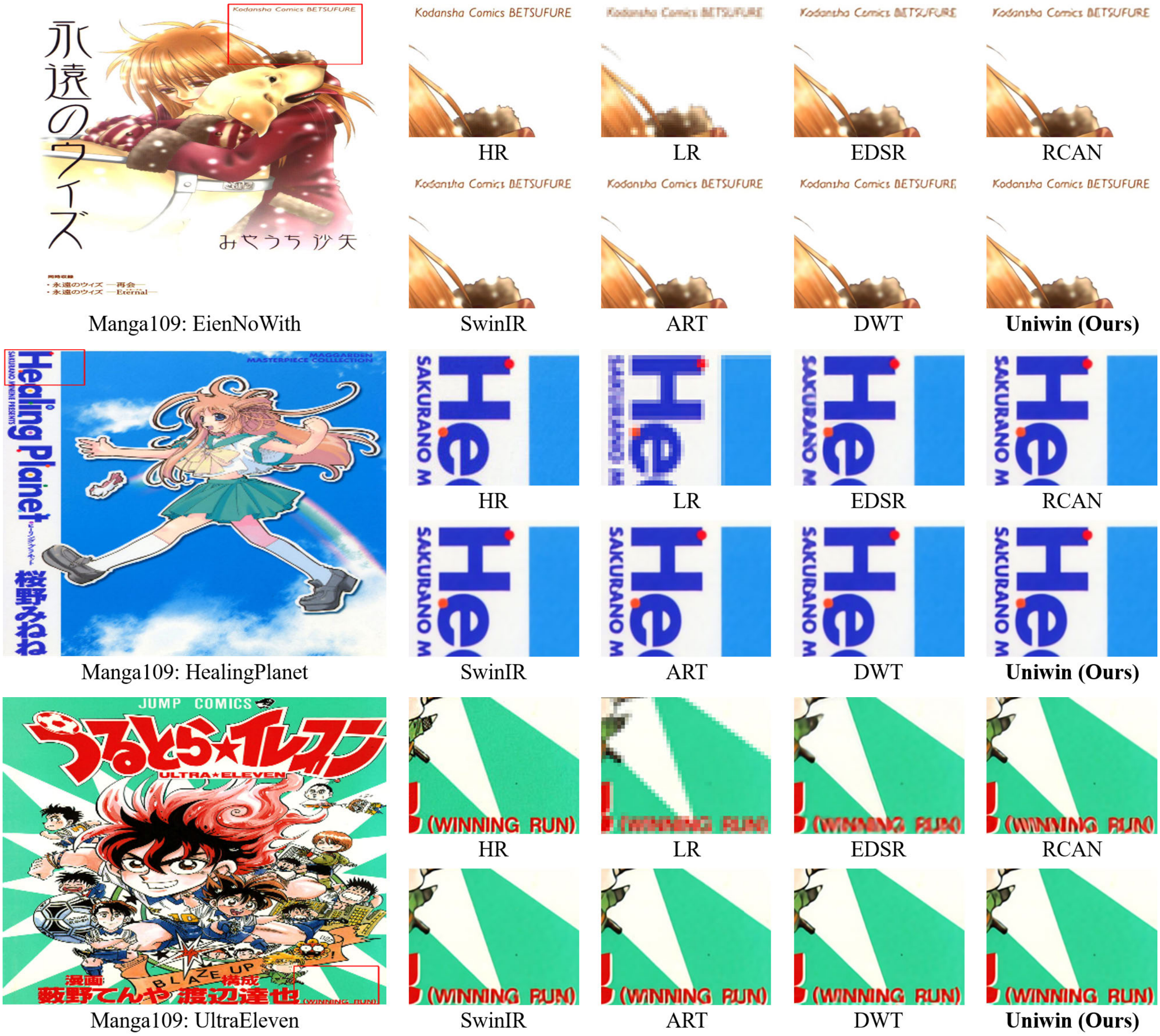


FIGURE 9. Visual comparison (×4) with state-of-the-art SR methods on Manga109 dataset.

where the parameters of Uniwin amount to just 58% of that of HAT. Surprisingly, despite this substantial parameter difference, the performance gap is not substantial. Uniwin outperforms HAT in specific scenarios, such as the BSD100 and Manga109 datasets at ×2 SR. This outcome underscores the remarkable efficiency of Uniwin with fewer parameters, mainly when the scale factor is small, such as ×2 SR.

2) QUALITATIVE COMPARISON

To facilitate a qualitative comparison between Uniwin and other state-of-the-art models, we present challenging examples of ×4 scale models in Figures 6, 7, 8, and 9. These figures respectively showcase the SR results from various models on the Set14 dataset [33], BSD100 dataset [34], Urban100 dataset [35], and Manga109 dataset [36].

In Figure 6, we analyze 'barbara' for texture restoration precision and 'ppt3' for edge information restoration accuracy. The results demonstrate that Uniwin excels in restoring textures and edges with higher clarity than its counterparts. In Figure 7, the restoration of '86016' relies significantly on the use of global information, such as the repeated textures found throughout the image, whereas the restoration of the other two images primarily depends on precise local information. This illustrates that our model is capable of utilizing both global and local information for image super-resolution. The images in Figure 8 have repetitive patterns at varying scales, illuminating how each SR model interacts

TABLE 4. PSNR(dB) results of ×2 SR with the addition of a SWA layer to the front and back of SwinIR residual blocks. Best results are in red color.

Method	#Params(M)	Set5	Set14	BSD100	Urban100	Manga109
SwinIR	11.75	38.42/0.9623	34.46/0.9250	32.53/0.9041	33.81/0.9427	39.92/0.9797
SwinIR + SWA (front)	13.49	38.51/0.9631	34.67/0.9264	32.59/0.9056	34.29/0.9457	40.02/0.9802
SwinIR + SWA (back)	13.49	38.48/0.9631	34.47/0.9245	32.57/0.9055	34.04/0.9443	39.93/0.9801

TABLE 5. PSNR(dB)/SSIM results of ×2 SR with gradual replacement the two shifted-window attention layers with two sliding-window attention layers. Best and Second-best results are in red and blue colors, respectively.

Method	#Params(M)	Set5	Set14	BSD100	Urban100	Manga109
WA×6 (**SwinIR**)	11.75	38.42/0.9623	34.46/0.9250	32.53/0.9041	33.81/0.9427	39.92/0.9797
SWA×2+WA×4	11.87	38.49/0.9630	34.59/0.9259	32.57/0.9054	34.16/0.9450	39.93/0.9800
SWA×4+WA×2	11.82	38.46/0.9629	34.58/0.9262	32.55/0.9052	34.02/0.9442	39.90/0.9799
SWA×6	11.78	38.30/0.9622	34.22/0.9242	32.47/0.9041	33.45/0.9398	39.68/0.9794
(SWA×1+WA×2)×2 (**Uniwin**)	11.87	38.57/0.9635	34.84/0.9274	32.63/0.9060	34.44/0.9469	40.28/0.9810

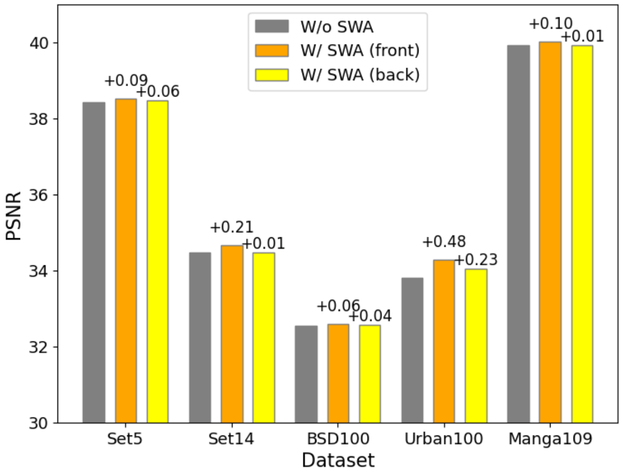


FIGURE 10. Chart of PSNR improvement for ×2 SR when integrating Sliding Window Attention (SWA) layer. The value above each bar represents the performance gain compared to the model without SWA layer. It shows better improvement when the SWA layer is positioned in the front of the model.

within the global region. Finally, Figure 9 displays images containing small letters, highlighting SR models' capability to preserve fine-grained details. Collectively, these examples illustrate Uniwin's remarkable performance across diverse scenarios, underscoring its superior SR capabilities.

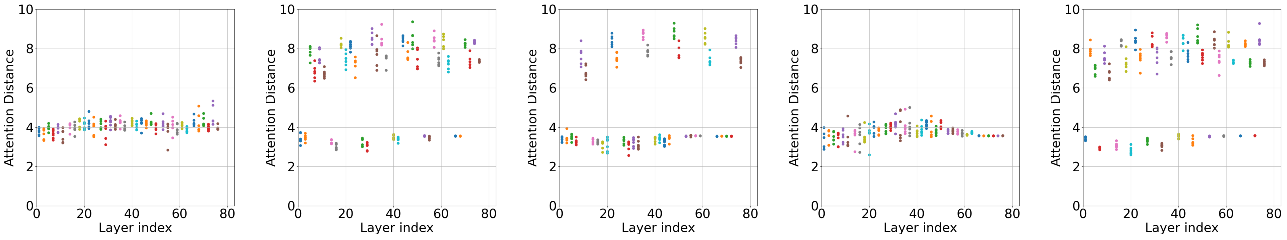
C. ABLATION STUDY

To examine the distinct contributions of two types of window-based attention layers, we manipulated the number and the order of sliding- and shifted-window attention layers integrated in RUAB and observed their effects.

Starting with SwinIR as the baseline, we evaluated the benchmark performance for ×2 SR after adding a single SWA layer to both the front and back of SwinIR's residual attention blocks. The results are detailed in Table 4, while Figure 10 visually represents this data through a bar chart, providing a clear visual representation of SWA layer's contribution to performance gain. Analysis of both Table 4 and Figure 10 clearly reveals that the incorporation of the SWA layer at the front of the model brings a more significant improvement. However, this causes a non-negligible parameter increase, so we replaced two shifted-window attention layers with two sliding-window attention layers to minimize the increase in model parameters.

Previous experiments have shown that applying SWA to the front side of the residual blocks is more effective than the back side, so we gradually replaced the front two WA layers with two SWA layers. The reason for replacing the two WA layers as a unit is that WA requires two attention operations to be performed in pairs due to the window shift. The results are shown in Table 5. Replacing some layers with SWA layers increases performance, but replacing all layers with SWA degrades performance. It means that not only SWA's local pattern interaction but also WA's fast access to global information has a significant impact on SR performance. Therefore, we have considered using only two SWA layers in Uniwin's design. We also measured the performance of model designs with different orders of SWA and WA layers. Instead of replacing the first two WA layers with SWA, we designed SWA and WA to be used repeatedly by replacing the first and fourth layers with SWA. This result is shown in the last row of Table 5, and we finally adopted it as the design of our model.

We measured the mean attention distance across various model variants to investigate whether SWA interacts more effectively with the local context than WA. Inspired by [38], we computed the mean attention distances using 1000 randomly selected images from the ImageNet [39] dataset, and the results are presented in Figure 11. In Figure 11, each plot corresponds to different models: SwinIR (WA×6), a model with SWA×2+WA×4, a model with SWA×4+WA×2, a model with SWA×6, and Uniwin, respectively from left to right. The x-axis in each plot represents the attention layers of each model, while the y-axis represents the mean attention distance.



(a) SwinIR (b) SWA×2+WA×4 (c) SWA×4+WA×2 (d) SWA×6 (e) **Uniwin (Ours)**

FIGURE 11. Mean attention distance of the model variants. The x-axis represents the model layers and the y-axis represents the mean attention distance.

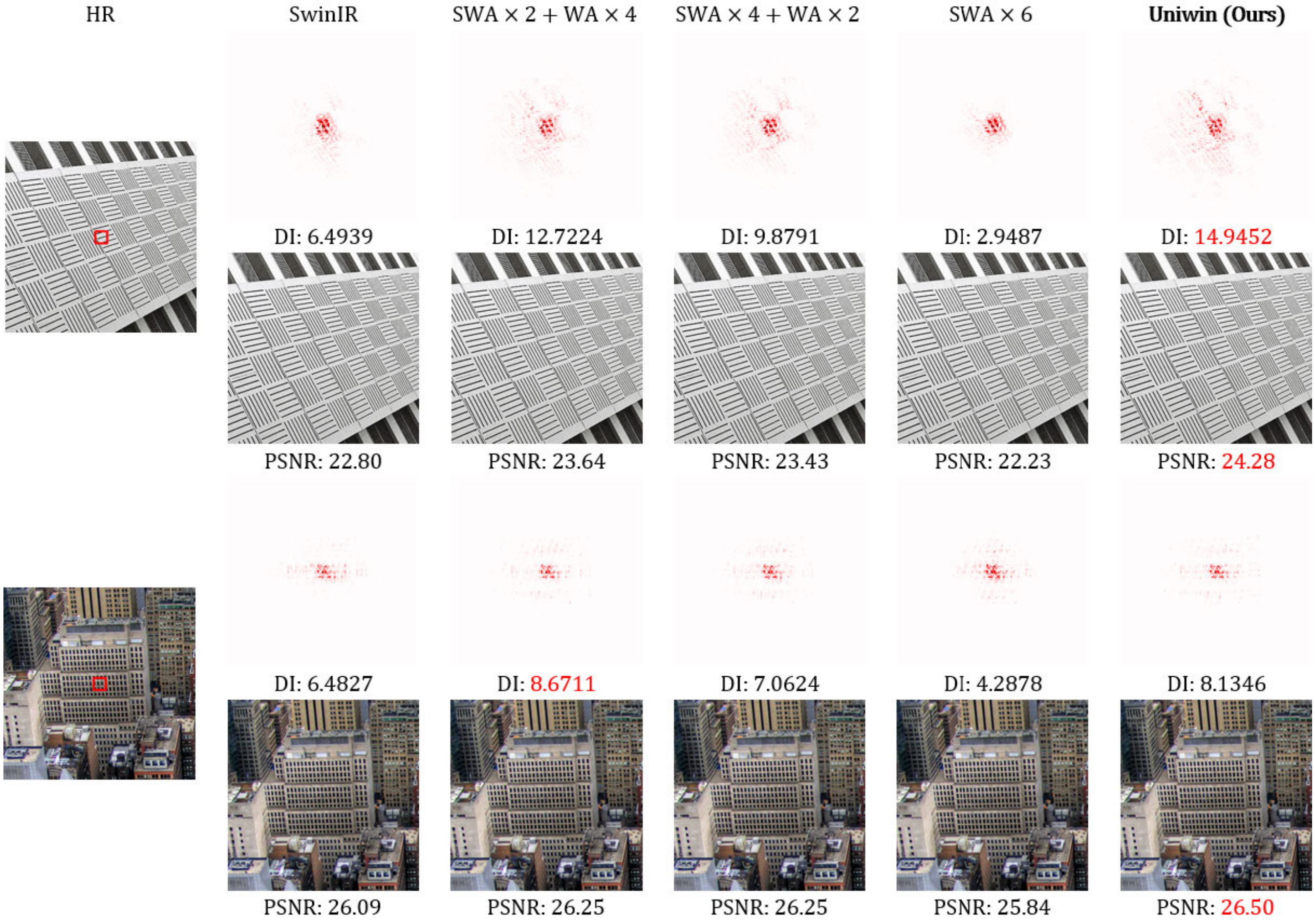


FIGURE 12. LAM results of the model variants. DI refers to Diffusion Index, and the larger the value, the more distributed it is.

In the case of SwinIR, shown in the leftmost plot, the mean attention distance remains around 4. A clear distinction can be observed in the mean attention distance between SWA layers and WA layers in the second, third, and fifth plots. Notably, SWA layers exhibit mean attention distances of less than 4, indicating their effectiveness in collecting local context by interacting with a smaller area than SwinIR's WA. On the other hand, the WA layers interact with a larger area than SwinIR, which seems to be influenced by the increased window size of the WA layer in our models.

The fourth plot illustrates the mean attention distances of the model where all layers are SWA. It shows that the mean attention distance eventually converges to less than 4, which means the limited access to the global context of this model. The relatively poorer performance of this model can be attributed to its limited interaction with the global context.

We also generated visualizations of the Local Attribution Maps (LAM) [28] for these model variants, presented in Figure 12. Each LAM image illustrates the pixels contributing to the reconstruction of the red-boxed area within the leftmost high-resolution (HR) image. Additionally, the Diffusion Index (DI) is provided below each LAM image, where a higher DI value indicates a broader scattering of pixels involved in the reconstruction process. As we can see, it becomes evident that the extent of an interacting region narrows as the proportion of SWA layers increases within a model.

In conclusion, these findings emphasize the importance of achieving a balanced integration between collecting accurate local context and interacting with global regions in SR. We believe that the performance of our model, Uniwin, results from a harmonious balance between local and global interactions achieved by applying alternating SWA and WA layers.

V. CONCLUSION

In this study, we introduced an innovative unified-window attention-based image super-resolution (SR) model named Uniwin. Uniwin capitalizes on the strengths of both sliding-window attention and shifted-window attention. Due to its overlapping window partition approach, sliding-window attention excels in capturing local context. Meanwhile, shifted-window attention, utilizing a non-overlapping window partition, swiftly accesses global information. By alternately applying these two window-based attentions in the Uniwin Attention Layer (UAL), we construct a potent feature extractor tailored for SR tasks.

Our experiments, including benchmark evaluations, showcased the efficacy of the Uniwin model. Furthermore, the conducted ablation studies, focusing on sliding window attention and shifted window attention, yielded crucial insights that it is importance to balance between gathering local context and engaging with the global region—an essential aspect for achieving successful SR outcomes.

In summary, our work presents Uniwin as a robust and effective solution that harmonizes local and global attention mechanisms to enhance image super-resolution performance.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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*.

[6] 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.

[7] 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.

[8] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 68–80.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[10] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, “Scaling local self-attention for parameter efficient visual backbones,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12894–12904.

[11] 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.

[12] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, “Vision transformer with deformable attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4784–4793.

[13] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, “Transformer for single image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 456–465.

[14] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, “Pre-trained image processing transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12294–12305.

[15] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, “Activating more pixels in image super-resolution transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22367–22377.

[16] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, “Swin2SR: SwinV2 transformer for compressed image super-resolution and restoration,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 669–687.

[17] W. Li, X. Lu, S. Qian, J. Lu, X. Zhang, and J. Jia, “On efficient transformer-based image pre-training for low-level vision,” 2021, *arXiv:2112.10175*.

[18] 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.

[19] S. Park and Y. S. Choi, “Image super-resolution using dilated window transformer,” *IEEE Access*, vol. 11, pp. 60028–60039, 2023.

[20] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, “Uformer: A general U-shaped transformer for image restoration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.

[21] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, “Restormer: Efficient transformer for high-resolution image restoration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[22] J. Zhang, Y. Zhang, J. Gu, Y. Zhang, L. Kong, and X. Yuan, “Accurate image restoration with attention retractable transformer,” 2022, *arXiv:2210.01427*.

[23] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, “Neighborhood attention transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6185–6194.

[24] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, “Non-local recurrent network for image restoration,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1673–1682.

[25] Y. Mei, Y. Fan, and Y. Zhou, “Image super-resolution with non-local sparse attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3516–3525.

[26] 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.

[27] D. Zhang, F. Huang, S. Liu, X. Wang, and Z. Jin, “SwinFIR: Revisiting the SwinIR with fast Fourier convolution and improved training for image super-resolution,” 2022, *arXiv:2208.11247*.

[28] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9195–9204.

[29] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision outlooker for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6575–6586, May 2023.

[30] 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.

[31] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1110–1121.

[32] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. 23rd Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2012, pp. 135.1–135.10.

[33] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.*, Avignon, France. Cham, Switzerland: Springer, 2010, pp. 711–730.

[34] 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. Vision. ICCV*, vol. 2, 2001, pp. 416–423.

[35] 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.

[36] 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.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[38] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12116–12128.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

GUNHEE CHO was born in Republic of Korea, in 1992. He received the B.S. degree in automotive engineering from Hanyang University, Seoul, South Korea, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. His research interests include ontology, knowledge-based systems, and computer vision.

YONG SUK CHOI was born in Republic of Korea, in 1969. He received the B.S., M.S., and Ph.D. degrees in computer science from Seoul National University, Seoul, South Korea, in 1993, 1995, and 2000, respectively. He joined Hanyang University, Seoul, in 2001, after working with the Telecommunication Research Laboratory, Samsung Electronics Company, from 1997 to 2001. He is currently a Professor with the Department of Computer Science and Engineering, Hanyang University. His research interests include deep learning algorithms, text understanding and summarization, large language model, image translation, visual question and answering, and multi-modal AI.