IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 May 2023, accepted 1 June 2023, date of publication 9 June 2023, date of current version 20 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3284539

RESEARCH ARTICLE

Image Super-Resolution Using Dilated Window Transformer

SOOBIN PARK[1] AND YONG SUK CHOI[2]
[1]Department of Artificial Intelligence, Hanyang University, Seoul 04763, South Korea
[2]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) Grant funded by the Korean Government [Ministry of Science and Information and Communication Technology (MSIT)] under Grant 2018R1A5A7059549 and Grant 2020R1A2C1014037; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government (MSIT), Artificial Intelligence Graduate School Program, Hanyang University, under Grant 2020-0-01373.

ABSTRACT Transformer-based networks using attention mechanisms have shown promising results in low-level vision tasks, such as image super-resolution (SR). Specifically, recent studies that utilize window-based self-attention mechanisms have exhibited notable advancements in image SR. However, window-based self-attention, results in a slower expansion of the receptive field, thereby restricting the modeling of long-range dependencies. To address this issue, we introduce a novel dilated window transformer, namely DWT, which utilizes a dilation strategy. We employ a simple yet efficient dilation strategy that enlarges the window by inserting intervals between the tokens of each window to enable rapid and effective expansion of the receptive field. In particular, we adjust the interval between the tokens to become wider as the layers go deeper. This strategy enables the extraction of local features by allowing interaction between neighboring tokens in the shallow layers while also facilitating efficient extraction of global features by enabling interaction between not only adjacent tokens but also distant tokens in the deep layers. We conduct extensive experiments on five benchmark datasets to demonstrate the superior performance of our proposed method. Our DWT surpasses the state-of-the-art network of similar sizes by a PSNR margin of 0.11dB to 0.27dB on the Urban100 dataset. Moreover, even when compared to state-of-the-art network with about 1.4 times more parameters, DWT achieves competitive results for both quantitative and visual comparisons.

INDEX TERMS Image super-resolution, self-attention mechanism, transformer, window-based self-attention.

I. INTRODUCTION

Image super-resolution (SR) is a well-known problem in low-level vision tasks, which aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart. Recent advancements in deep learning have enabled image SR using deep convolutional neural networks such as residual learning [1], [2], dense blocks [3], attention mechanisms [4], [5], [6], and adversarial learning [7], [8], [9]. Due to improved results, convolutional neural networks have become the de facto standard for this field.

In recent years, inspired by the remarkable success of Transformer [10] in the field of natural language processing, several researchers have attempted to adopt transformer-based networks for high-level vision tasks with promising results across high-level vision tasks, such as image classification [11], [12], [13], object detection [14], [15], [16], and dense prediction [17], [18]. Following the success of this approach, researchers have also introduced transformer-based networks for low-level vision tasks, including SR [19], [20], [21]. In particular, SwinIR [21], which adopts window-based self-attention of Swin Transformer [15], has achieved breakthrough performance in the SR task.

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

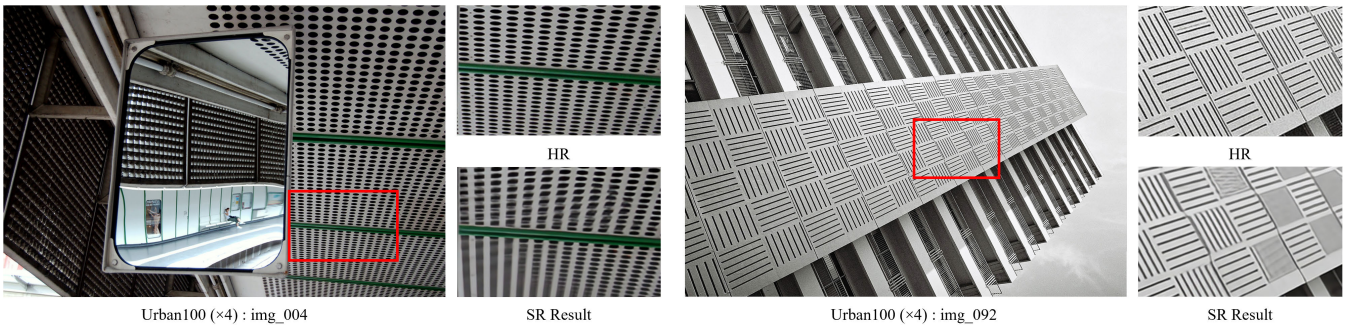


FIGURE 1. SR results of SwinIR. SwinIR fails to accurately restore textures even for images with self-repeating patterns. These findings indicate that SwinIR cannot utilize the complete global information of the image.

However, window-based self-attention leads to slower growth of receptive field, which limits the potential of modeling long-range dependencies [16]. Window-based self-attention is effective at capturing local context, but it falls short in capturing global context. This limitation affects the image quality reconstructed by SR networks using window-based self-attention. As shown in Fig. 1, SwinIR restores unclear textures even for images with repetitive patterns. Due to the lack of global information, SwinIR cannot utilize information of similar patterns located at a farther distance. To solve this problem, some recent studies have achieved outstanding performance compared to SwinIR by proposing pre-training methods on large-scale datasets [22] such as ImageNet [23] or by combining CNN-based channel attention [4] and window-based self-attention to use their complementary advantages [24]. However, these gains have come at the cost of additional large-scale training data and a greater number of parameters compared to SwinIR.

In this work, we propose a dilated window transformer (DWT) that complements the limitations of SwinIR without introducing additional training data and parameters. Our DWT introduces two types of window-based multi-head self-attention blocks, named the window attention block (WAB) and the dilated window attention block (DWAB). We adopt a structure that alternates between WAB and DWAB. The WAB is responsible for extracting local features using standard window attention, while the DWAB uses a dilation strategy to extract global features. Similar to dilated convolution [25], DWAB places intervals between the tokens in each window to expand the receptive field. In contrast to SwinIR's window-based self-attention, which only interacts with adjacent tokens at all layers, we use the dilation strategy in an effective way to allow each token to interact with tokens that are farther away as the layers become deeper. As a result, our DWT effectively utilizes both local and global context when restoring images. Experimental results show that our DWT achieves better performance than the state-of-the-art models of similar sizes on five benchmark datasets.

To summarize, the main contributions of our DWT are as follows:

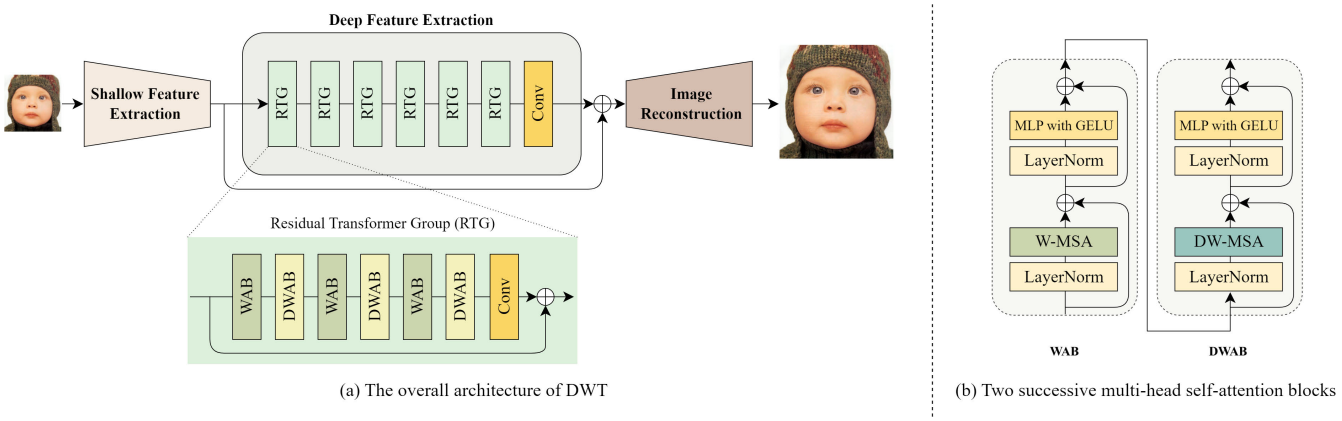
- We introduce a DWT, which leverages a dilation strategy to effectively extract both local and global features, addressing the limitations of window-based self-attention.
- We propose a dilation strategy that is adopted in the DWAB of our DWT. This strategy efficiently widens the receptive field, allowing for improved modeling of long-range dependencies.
- By conducting extensive experiments, we demonstrate that our DWT achieves promising results compared to other state-of-the-art methods, while the number of parameters and computational cost are competitive in comparison with conventional methods. The superior performance of our model highlights the effectiveness of our proposed dilation strategy for image SR.

The rest of the paper is organized as follows. In Section II, we summarize the related work. Then, we describe our proposed method in Section III. Section IV presents the experimental results and analysis. Finally, conclusions are drawn in Section V.

II. RELATED WORK

A. IMAGE SUPER-RESOLUTION

Since the introduction of SRCNN [26], [27], which applied the deep convolutional neural network to image SR for the first time, various deep neural networks with different designs have been proposed. Kim et al. [2] utilized residual learning to build a deeper network and speed up convergence. Ledig et al. [7] introduced adversarial learning to improve the texture details of the reconstructed images. Several methods using attention mechanisms, such as channel attention [5], [28] and non-local attention [6], [29] have significantly improved the performance of the SR task. In addition, networks using recurrent neural networks [6], [30] and graph neural networks [31] have also been proposed. Recently, transformer-based networks [19], [32] have been applied to the SR task and have shown remarkable performance. These networks take advantage of their ability to model long-range dependencies to improve the quality of reconstructed images.



(a) The overall architecture of DWT

(b) Two successive multi-head self-attention blocks

FIGURE 2. **(a) Overall architecture of DWT. The input image goes through a shallow feature extraction module, and the shallow features are fed into deep feature extraction module to extract deep features. Finally, the shallow features and deep features are fused by a skip connection, and an image reconstruction module is utilized to generate the SR result. (b) The inner structure of two successive multi-head self-attention blocks. The WAB and the DWAB are alternately applied in a pair of multi-head self-attention blocks.**

B. VISION TRANSFORMER

Inspired by the success of Transformer [10] in machine translation tasks, transformer-based networks have been introduced in computer vision community. Dosovitskiy et al. [11] introduced VIT, which was the first application of a transformer architecture in computer vision to non-overlapping medium-sized image patches. The ability of transformer-based networks to model long-range dependencies through self-attention has shown impressive performance in high-level vision tasks, such as image classification [11], [12], [13], [33], object detection [14], [15], [34], [35], [36], and dense prediction [17], [18], [37], [38]. Transformer-based architectures have also been applied to low-level vision tasks. Chen et al. [19] proposed a standard transformer architecture called IPT for various low-level vision tasks, including image SR, denoising, and deraining. Liang et al. [21] proposed SwinIR, which introduced window-based self-attention of the Swin Transformer [15] instead of standard self-attention, leading to tremendous growth in the SR task. However, SwinIR has a structural limitation that it cannot model long-range dependencies in input images, despite its advantage of efficient local feature extraction using window-based self-attention. Zhang et al. [39] proposed an attention retractable transformer named ART for low-level vision tasks to compensate for the limitations of dense attention used in window-based self-attention. ART introduced a sparse attention strategy similar to the dilation strategy we proposed. In this paper, we propose a DWT that uses a dilation strategy more efficiently than ART to successfully extract both local and global features for improved performance in image SR.

III. PROPOSED METHOD

A. MOTIVATION

SwinIR [21], which applied the Swin Transformer architecture [15], has demonstrated the significant potential of transformer-based networks in image SR. SwinIR extracts deep features using window-based self-attention and shifted window-based self-attention, which demonstrates robust capabilities in local feature extraction.

However, SwinIR's window-based self-attention has a structural limitation that falls short of capturing global context. This is due to the use of a smaller and slowly growing receptive field in comparison to the full-sized receptive field used in standard self-attention. This limitation leads to serious defects that cannot produce high-quality output images in image SR. For instance, SwinIR restores incorrect textures, even for images with self-repeating patterns, as shown in Fig. 1. This phenomenon indicates that SwinIR does not fully leverage the global information of the image, and extracting both local and global information is important to achieve better performance in image SR. Thus, effectively utilizing global information to reconstruct HR images may overcome the limitations of SwinIR. Based on this motivation, we propose a DWT, which can effectively enlarge the receptive field.

B. THE OVERALL ARCHITECTURE

As shown in Fig. 2(a), our DWT consists of three modules, including shallow feature extraction, deep feature extraction, and image reconstruction. Given a LR image $I_{LR} \in \mathbb{R}^{H\times W\times C_{in}}$ (H, W, and C_{in} are the image height, width, and the input channel number, respectively), we apply a single 3×3 convolutional layer $H_{SF}(\cdot)$ to obtain shallow features $F_{SF} \in \mathbb{R}^{H\times W\times C}$ as:

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

where C is the channel number of the feature. According to [40], a convolutional stem can result in more reliable optimization. Additionally, it can effectively map an input image from a low-dimensional space to a high-dimensional space. Then, deep features $F_{DF} \in \mathbb{R}^{H\times W\times C}$ are extracted as:

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

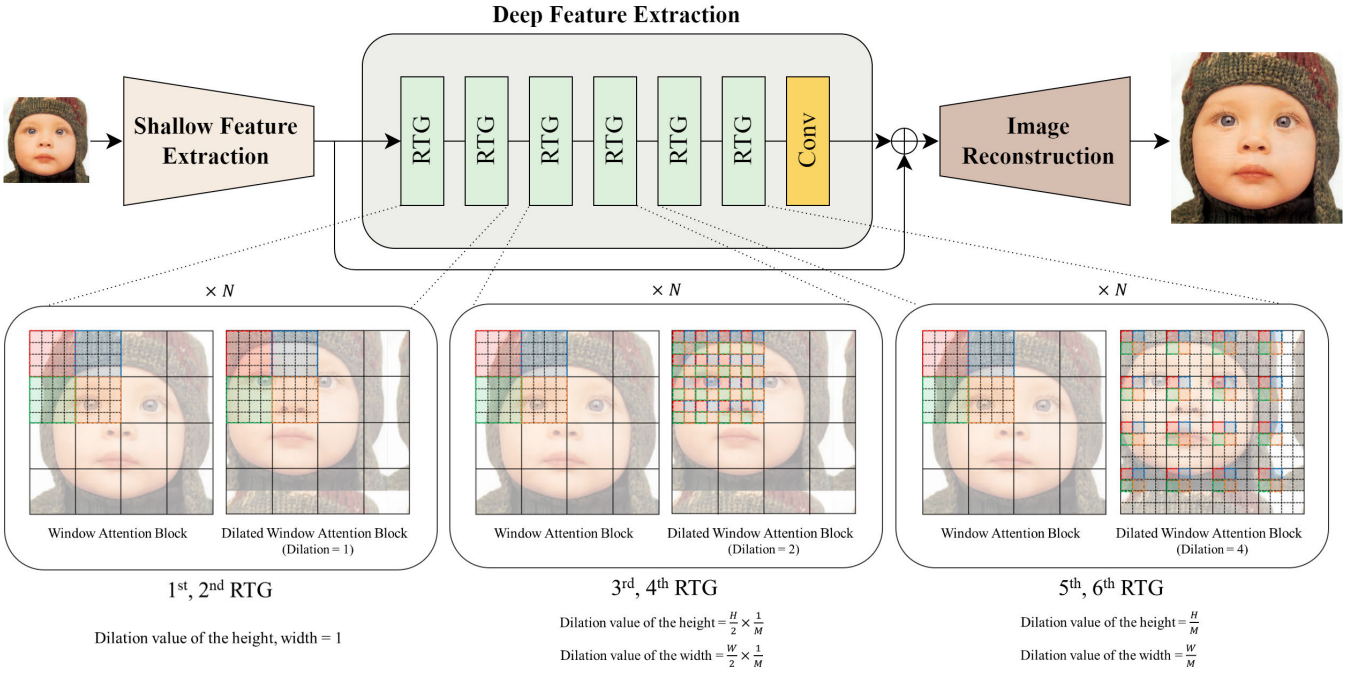


FIGURE 3. Illustration of two types of attention blocks in RTG. The sizes of the baby picture and window are 16 × 16 and 4 × 4, respectively. A small square represents one pixel, and pixels of the same color belong to the same window. The dilation value depends on the depth of the network: smaller values for shallow layers, and larger values for deep layers. A larger dilation value encourages the model to capture long-range dependencies.

where $H_{DF}(\cdot)$ is a deep feature extraction module consisting of M residual transformer groups (RTG) and one 3×3 convolutional layer. The intermediate features of the deep feature extraction module are sequentially extracted as:

$$\begin{aligned} F_i &= H_{RTG_i}(F_{i-1}), i = 1, 2, \ldots, M, \\ F_{DF} &= H_{Conv}(F_M), \end{aligned} \tag{3}$$

where $H_{RTG_i}(\cdot)$ denotes the i-th RTG and $H_{Conv}(\cdot)$ denotes the last convolutional layer at the end of the deep feature extraction module. This last convolutional layer can bring the inductive biases into the transformer-based network and lead to better aggregation of deep features [21]. Shallow features F_{SF} and deep features F_{DF} are fused by a long skip connection and passed through the image reconstruction module $H_{Rec}(\cdot)$ to generate a HR image I_{SR} as:

$$I_{SR} = H_{Rec}(F_{SF} + F_{DF}). \tag{4}$$

Specifically, we use the sub-pixel convolutional layer [41] to upscale the feature. We optimize our model parameters with L_1 pixel loss, which is known to be effective in image SR.

C. RESIDUAL TRANSFORMER GROUP (RTG)

As shown in Fig. 2(a), the RTG is a residual group consisting of N pairs of multi-head self-attention blocks and one 3×3 convolutional layer. For the i-th RTG, it is formulated as:

$$\begin{aligned} F_{i,j} &= H_{MHSAB_{ij}}(F_{i,j-1}), j = 1, 2, \ldots, N, \\ F_{i,out} &= H_{Conv}(F_{i,N}) + F_{i,0}, \end{aligned} \tag{5}$$

where $H_{MHSAB_{ij}}(\cdot)$ is the j-th pair of multi-head self-attention blocks in the i-th RTG. Following [21], at the end of the RTG, we employ a single 3×3 convolutional layer $H_{Conv}(\cdot)$ and the residual connection is also added.

D. SUCCESSIVE MULTI-HEAD SELF-ATTENTION BLOCKS

We introduce two types of window-based multi-head self-attention blocks: WAB and DWAB. Commonly, window-based self-attention proceeds as follows. Given an input feature of size $H \times W \times C$, it is first partitioned into $\frac{H}{M} \times \frac{W}{M}$ non-overlapping windows of size $M \times M$. Note that we treat each pixel as a token so that our DWT can learn pixel-level information. Then, self-attention is calculated separately for each window. For a local window feature $X \in \mathbb{R}^{M^2 \times C}$, the $query$, key, and $value$ metrics Q, K, and V are computed by linear projection as:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \tag{6}$$

where W_Q, W_K, and W_V denote the weight metrics for linear projection. Then, the attention matrix is computed by the window-based self-attention as:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V, \tag{7}$$

where d is the dimension of the $query/key$ and B is the learnable relative positional encoding.

As shown in Fig. 2(b), WAB and DWAB are alternately applied in a pair of successive multi-head self-attention blocks. Both WAB and DWAB are based on the window-based self-attention of the Swin Transformer [15]. The key difference between the two lies in the dilation strategy

employed in DWAB. Different from SwinIR, we use the dilation strategy with the shifted window mechanism in DWAB to obtain a wider receptive field. As shown in Fig. 3, for WAB, every $M \times M$ token of each window is adjacent. In contrast, for DWAB, $M \times M$ tokens of each window are sampled with a dynamic interval size. Therefore, while WAB can extract local features through interactions with adjacent tokens, DWAB can extract global features through interactions with tokens that are further away. With the dilation strategy, consecutive multi-head self-attention blocks are computed as:

$$\begin{aligned} \hat{x}^{l} &= \text{W-MSA}(\text{LN}(x^{l-1})) + x^{l-1}, \\ x^{l} &= \text{MLP}(\text{LN}(\hat{x}^{l})) + \hat{x}^{l}, \\ \hat{x}^{l+1} &= \text{DW-MSA}(\text{LN}(x^{l})) + x^{l}, \\ x^{l+1} &= \text{MLP}(\text{LN}(\hat{x}^{l+1})) + \hat{x}^{l+1}, \end{aligned} \tag{8}$$

where $\hat{x}^{l}$ and x^{l} denote the output feature of the (D)W-MSA and the MLP for l-th attention block, respectively. MLP denotes a multi-layer perceptron that has two fully-connected layers with GELU activation function between them, and LN denotes the layer normalization. W-MSA and DW-MSA denote window-based multi-head self-attention and dilated window-based multi-head self-attention, respectively.

1) WINDOW ATTENTION BLOCK (WAB)

As shown in Fig. 3, in a WAB, multi-head self-attention is computed within non-overlapping windows. Each token can interact with neighboring $M \times M$ tokens, including itself.

2) DILATED WINDOW ATTENTION BLOCK (DWAB)

In this section, we elaborate on the key design element of DWT, the DWAB.

Inspired by dilated convolution [25], we introduce a dilation strategy. Dilated convolution is a type of convolution that enlarges the kernel by inserting holes between the kernel elements. In a similar method, we employ a dilation strategy with the shifted window mechanism to even-numbered attention blocks in every RTG. Similar to SwinIR [21], DWAB utilizes a shifted window mechanism for cross-window connections. As illustrated in Fig. 3, the dilation value depends on the depth of the network and the input image size. The dilation value indicates the interval between the tokens of each window. Thus, in DWAB, as the dilation value increases, the tokens within each window interact with tokens located at a greater distance. In the first two RTGs, we set the dilation value as 1 to sufficiently extract local features in the early stages of deep feature extraction. Therefore, the DWAB of the first two RTGs is exactly the same as the shifted window-based self-attention of SwinIR [21]. In the third and fourth RTGs, DWT extracts $M \times M$ tokens for a window from the $\frac{1}{4}$ area of the shifted input image, while in the last two RTGs, DWT extracts $M \times M$ tokens for a window from the entire area of the shifted input image. The dilation value increases with depth of the network, allowing each token to interact with a wider area as the layers become deeper.

In Fig. 3, we illustrate an example where the height and width of the input image are both 16 and the window size is set to 4×4. For this example, the dilation values of the height and width are computed as $\frac{16}{2} \times \frac{1}{4} = 2$ in the third and fourth RTGs, while in the last two RTGs, the dilation values of the height and width are computed as $\frac{16}{4} = 4$. The proposed DWAB enables the receptive field to be widened faster and more efficiently than the standard shifted window-based self-attention. Specifically, local features are extracted in the shallow layers through interactions between adjacent tokens, while global features are extracted in the deep layers by interacting with neighboring and distant tokens.

In summary, our DWAB provides an effective means of widening the receptive field and improving local and global feature extraction.

E. DIFFERENCES FROM RELATED WORK

In this section, we provide a detailed comparison between our proposed DWT and the ART [39] introduced in Section II. ART is an attention retractable transformer that uses sparse attention, which is similar to our dilation strategy. However, there are several differences between these methods. We compare the differences in two aspects.

1) WINDOW SIZE IN THE ATTENTION BLOCK

In our proposed DWT, the window size in DWAB is always fixed to 16×16, irrespective of the input image size. However, in ART's sparse attention block (SAB), the window size varies depending on the input image size. ART uses a fixed interval size of 4, meaning that as the input image size increases, the window size also increases, leading to a higher computational cost. For example, if the height and width of the input image are both 160, the window size in SAB of ART is 40×40, while the window size in DWAB of our DWT is 16×16.

2) DESIGN OF THE ATTENTION BLOCK

As explained in Section III-D2, our proposed DWT employs a simple yet efficient dilation strategy to gradually expand the receptive field. In contrast to ART, which has a fixed interval size for all layers, our DWT increases the dilation value as the layers become deeper, allowing for a wider area where dilation is applied. As a result, our DWT can extract both local and global features efficiently in the shallow and deep layers, respectively. We provide a detailed analysis of the effectiveness of our proposed dilation strategy in Section IV-B.

IV. EXPERIMENT AND ANALYSIS

A. EXPERIMENTAL SETUP

1) IMPLEMENTATION DETAILS

For DWT implementation, the RTG number is set to 6. Both the WAB and DWAB number of each RTG are set to 3 and attention head is set to 6. Therefore, each RTG is composed of three pairs of multi-head self-attention blocks. Since the

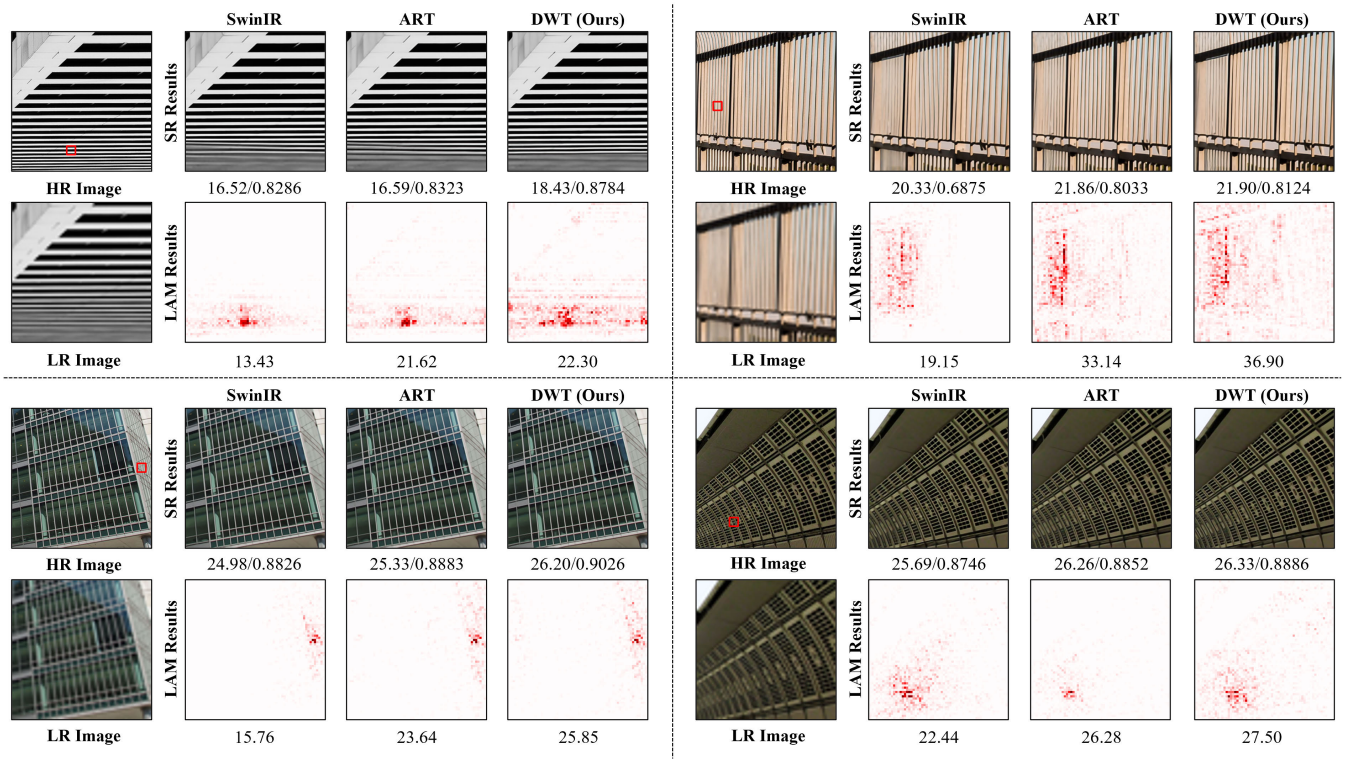


FIGURE 4. LAM comparison results. The LAM results represent the importance of each pixel in the input LR image with respect to the SR results of the patch marked with a red box [42]. The higher DI provided below the LAM results indicates a wider range of pixels used.

TABLE 1. Ablation study on the design of DWAB (×2 SR).

	Set5 [43]	Set14 [44]	BSD100 [45]	Urban100 [46]	Manga109 [47]
w/ dynamic dilation value (used in DWT)	38.49/0.9632	34.56/0.9265	32.56/0.9054	34.14/0.9444	40.01/0.9802
w/ fixed dilation value	38.45/0.9630	34.42/0.9259	32.54/0.9051	33.94/0.9429	39.92/0.9799

DWAB uses shifted window mechanism, we adopt a masking strategy in DWAB to restrict self-attention between the non-adjacent areas, similar to the Swin Transformer [15] and SwinIR [21]. The channel number for all the modules except the image reconstruction module is set to 180, while in the image reconstruction module, it is set to 64. All convolutional layers in DWT have 3 × 3 kernel, stride of length 1, and padding of length 1, so the height and width of the feature map remain the same as the input size before upsampling. In [22] and [24], the authors showed the effectiveness of using a large window size. Therefore, we set the window size to 16 × 16. To ensure fair comparison, we also provide a smaller version of DWT, which we refer to as DWT-S. In DWT-S, we set the window size to 8 × 8, while keeping the other settings the same as DWT.

2) DATASETS

Following previous work [21], [39], we use DF2K as the training dataset, which consists of 800 images from DIV2K [1] and 2560 images from Flicker2K [48]. We evaluate our model on Set5 [43], Set14 [44], BSD100 [45], Urban100 [46], and Manga109 [47] datasets.

3) EVALUATION METRICS

For evaluation of the SR result, we use PSNR and SSIM [49] computed on the Y channel of the YCbCr color space. The PSNR and SSIM are commonly used full-reference image quality assessment (FR-IQA) metrics for image SR. These metrics evaluate the fidelity of the reconstructed image, with higher values indicating better image fidelity. In addition, for the comparison of the perceptual quality, we also utilize no-reference image quality assessment (NR-IQA) metrics, NIQE [50] and BRISQUE [51]. A lower value of both NIQE and BRISQUE indicates higher perceptual quality.

4) TRAINING DETAILS

We generate LR images by downsampling the ground truth images using the "bicubic" method in MATLAB. During the training phase, we randomly crop the LR images into input patches of size 64 × 64 and apply data augmentation

TABLE 2. FR-IQA results comparison with numerous state-of-the-art SR methods. The best and the second-best values are highlighted with red and blue, respectively.

Method	Scale	Training Dataset	Set5 [43]		Set14 [44]		BSD100 [45]		Urban100 [46]		Manga109 [47]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [1]	×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 [29]	×2	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
SwinIR [21]	×2	DF2K	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
EDT [22]	×2	DF2K	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
ART-S [39]	×2	DF2K	38.48	0.9625	34.50	0.9258	32.53	0.9043	34.02	0.9437	40.11	0.9804
ART [39]	×2	D2FK	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
DWT-S (Ours)	×2	DF2K	38.40	0.9628	34.44	0.9254	32.53	0.9048	33.77	0.9419	39.88	0.9798
DWT (Ours)	×2	DF2K	38.49	0.9632	34.56	0.9265	32.56	0.9054	34.14	0.9444	40.01	0.9802
EDSR [1]	×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 [29]	×3	DIV2K	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
SwinIR [21]	×3	DF2K	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
EDT [22]	×3	DF2K	34.97	0.9316	30.89	0.8527	29.44	0.8142	29.72	0.8814	35.13	0.9534
ART-S [39]	×3	DF2K	34.98	0.9318	30.94	0.8530	29.45	0.8146	29.86	0.8830	35.22	0.9539
ART [39]	×3	D2FK	35.07	0.9325	30.99	0.8540	29.51	0.8159	30.10	0.8871	35.39	0.9548
DWT-S (Ours)	×3	DF2K	34.94	0.9320	30.91	0.8530	29.45	0.8159	29.73	0.8806	35.10	0.9533
DWT (Ours)	×3	DF2K	35.00	0.9327	30.97	0.8541	29.49	0.8170	30.07	0.8860	35.27	0.9542
EDSR [1]	×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 [29]	×4	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
SwinIR [21]	×4	DF2K	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
EDT [22]	×4	DF2K	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.05	0.9254
ART-S [39]	×4	DF2K	32.86	0.9029	29.09	0.7942	27.91	0.7489	27.54	0.8261	32.13	0.9263
ART [39]	×4	D2FK	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
DWT-S (Ours)	×4	DF2K	32.88	0.9046	29.06	0.7947	27.91	0.7507	27.50	0.8253	32.03	0.9253
DWT (Ours)	×4	DF2K	32.92	0.9055	29.12	0.7961	27.94	0.7519	27.81	0.8324	32.20	0.9274

TABLE 3. Model resource comparison with numerous transformer-based SR methods (×4 SR). Input size is 3 × 160 × 160 for Mult-Adds calculation.

Methods	SwinIR [21]	EDT [22]	ART-S [39]	ART [39]	DWT-S (Ours)	DWT (Ours)
Params(M)	11.90	11.63	11.87	16.55	11.90	12.06
Mult-Adds (G)	316	354	319	448	316	319

techniques such as horizontal flip and random rotation. However, during the evaluation phase, the input image size is not fixed. Therefore, we employ a reflection padding strategy on the input image to ensure that the number of windows is always an integer. We use a mini-batch size of 32 and train for a total of 500K iterations, with the learning rate initialized at 2e-4 and reduced by half at [250K,400K,450K,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 adopt the Adam [52] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and zero weight decay to optimize our model. DWT is implemented on the PyTorch [53] framework with 4 NVIDIA RTXA5000 GPUs.

B. ABLATION STUDY

1) EFFECTIVENESS OF DWAB DESIGN

As described in Section III-D2, we use the dilation strategy to extract both local and global features effectively. In our DWAB, the dilation value increases as the layers become deeper, resulting in a wider receptive field.

To demonstrate the effectiveness of our proposed dilation strategy, we conduct an ablation study. We compare the dynamic dilation value strategy, which gradually increases the dilation value and expands the area where dilation is applied as the layers become deeper, with the strategy that uses a fixed dilation value for all layers. In the fixed dilation strategy, the dilation values of the height and width are set to $\frac{H}{M}$, and $\frac{W}{M}$, respectively for all DWAB, the same as the fifth and sixth RTGs in Fig. 3.

We evaluate the quantitative performance of both strategies on five benchmark datasets for ×2 SR. These results are shown in Table 1. The results demonstrate that the strategy of gradually increasing the dilation value yields better performance on all benchmark datasets. This implies that, in order to achieve good performance in SR, extracting both local and global features are important and the design of our DWT makes it possible to achieve this.

C. LAM RESULTS COMPARISON

We propose a dilation strategy to exploit global information in image reconstruction by gradually expanding the receptive field.

TABLE 4. NR-IQA results comparison with numerous transformer-based SR methods. The top three values are highlighted with red, blue and purple, respectively.

Method	Scale	Set5 [43]		Set14 [44]		BSD100 [45]		Urban100 [46]		Manga109 [47]	
		NIQE	BRISQUE	NIQE	BRISQUE	NIQE	BRISQUE	NIQE	BRISQUE	NIQE	BRISQUE
SwinIR [21]	×2	5.061	26.351	4.919	27.099	4.829	33.520	4.919	25.289	4.438	31.193
EDT [22]	×2	5.003	26.268	4.966	26.953	4.858	34.131	4.431	25.391	4.453	31.256
ART-S [39]	×2	5.048	26.050	5.006	26.552	4.856	33.277	4.463	25.445	4.380	31.582
ART [39]	×2	5.011	26.229	4.999	26.651	4.856	33.385	4.455	25.502	4.221	32.769
DWT-S (Ours)	×2	5.016	26.222	5.019	26.735	4.887	33.859	4.496	25.599	4.585	30.947
DWT (Ours)	×2	5.035	26.181	5.061	26.664	4.877	33.560	4.492	25.354	4.577	30.865
SwinIR [21]	×3	6.447	34.918	5.614	37.398	5.592	43.125	5.038	33.816	4.920	37.379
EDT [22]	×3	6.413	35.900	5.731	37.988	5.685	43.457	5.075	34.281	4.911	37.587
ART-S [39]	×3	6.396	34.870	5.771	38.028	5.641	42.835	5.082	33.965	4.896	37.378
ART [39]	×3	6.355	35.530	5.784	37.348	5.635	42.802	5.088	33.495	4.886	39.623
DWT-S (Ours)	×3	6.335	34.596	5.691	38.262	5.685	43.104	5.137	34.161	4.934	37.629
DWT (Ours)	×3	6.339	34.814	5.710	37.819	5.686	43.112	5.137	33.690	4.939	37.349
SwinIR [21]	×4	7.096	39.972	6.221	44.001	6.084	46.693	5.462	38.835	5.337	41.401
EDT [22]	×4	7.344	41.295	6.246	45.169	6.169	47.276	5.500	39.342	5.286	41.467
ART-S [39]	×4	6.944	40.858	6.280	44.638	6.147	46.814	5.513	38.558	5.271	40.990
ART [39]	×4	7.012	41.042	6.285	44.146	6.099	46.645	5.498	38.085	5.294	40.849
DWT-S (Ours)	×4	6.838	39.474	6.257	45.370	6.076	46.571	5.545	38.383	5.320	41.139
DWT (Ours)	×4	6.861	39.315	6.275	45.514	6.061	46.394	5.583	38.163	5.312	40.823

To analyze whether our dilation strategy works as intended, we use LAM [42]. LAM is a sophisticated attribution method for SR that identifies the input pixels that significantly affect the SR results and quantifies the results into a diffusion index (DI) that evaluates the extraction and utilization of information from the LR image. A LAM result with a higher DI means more pixels are involved in restoring images in a specific area. Fig. 4 shows the LAM results (DI) and SR results (PSNR/SSIM) for SwinIR, ART, and our DWT. In the LAM results, the contribution areas are illustrated in red. As we can see, among the three models, our DWT has the highest DI, PSNR, and SSIM values, and achieves better visual results. Furthermore, we can observe that the red area of the DWT's LAM result is more widely spread than that of other models (Fig. 4). This means that DWT can leverage a wider range of information than SwinIR and ART. Therefore, these results demonstrate the efficiency of our DWT, which can effectively utilize both local and global information to improve SR performance.

D. QUANTITATIVE COMPARISON

1) FR-IQA RESULTS

Table 2 presents the FR-IQA results comparison between our proposed DWT and other state-of-the-art methods, including EDSR [1], RCAN [4], NLSA [29], SwinIR [21], EDT [22], ART-S [39], and ART [39]. As illustrated in Table 2, our DWT achieves the best or second-best performance across all scale factors. Especially, DWT shows greater performance improvement in SSIM metric than in PSNR. Since SSIM is a metric that considers the human visual perception system, these results imply that our DWT generates higher quality images in terms of human perception.

We also provide a comparison of the parameter numbers and Mult-Adds for transformer-based networks in Table 3. The Mult-Adds are calculated assuming a $3 \times 160 \times 160$ input size for ×4 SR. As indicated in Table 3, ART has about 1.4 times more parameters and computational cost than our DWT. However, DWT exhibits superior or competitive performance than ART. The small version of DWT, DWT-S, also shows competitive results with less or similar computational cost compared to SwinIR and ART-S. When compared to state-of-the-art models of similar sizes, with the exception of ART, DWT outperforms all of them in terms of PSNR and SSIM. Specifically, while DWT and ART-S have similar computational cost, DWT surpasses ART-S by a PSNR margin of up to 0.11dB to 0.27dB on the Urban100 dataset. Furthermore, in the ×4 SR results of the Urban100, it can be observed that DWT achieves superior performance not only compared to ART-S but also to ART. The Urban100 dataset exhibits a higher disparity in performance compared to other datasets due to its abundance of repeated patterns. These characteristics of the Urban100 are advantageous for our dilation strategy that enable access distant features. Considering the fact that the resolution of the Urban100 is relatively higher than that of Set5 and Set14, it can be inferred that our model is much more efficient than ART.

All of these results collectively demonstrate the effectiveness of the DWT, which achieves superior performance with a competitive number of parameters and acceptable computational cost.

2) NR-IQA RESULTS

For a more comprehensive comparison between our DWT and other transformer-based networks, we also provide NR-IQA results comparison in Table 4. The goal of NR-IQA is to estimate the perceptual quality of the image rather than the image fidelity. However, it is debatable whether the NR-IQA metrics precisely reflect human perceptual quality [54], [55]. Hence, we use these two metrics solely

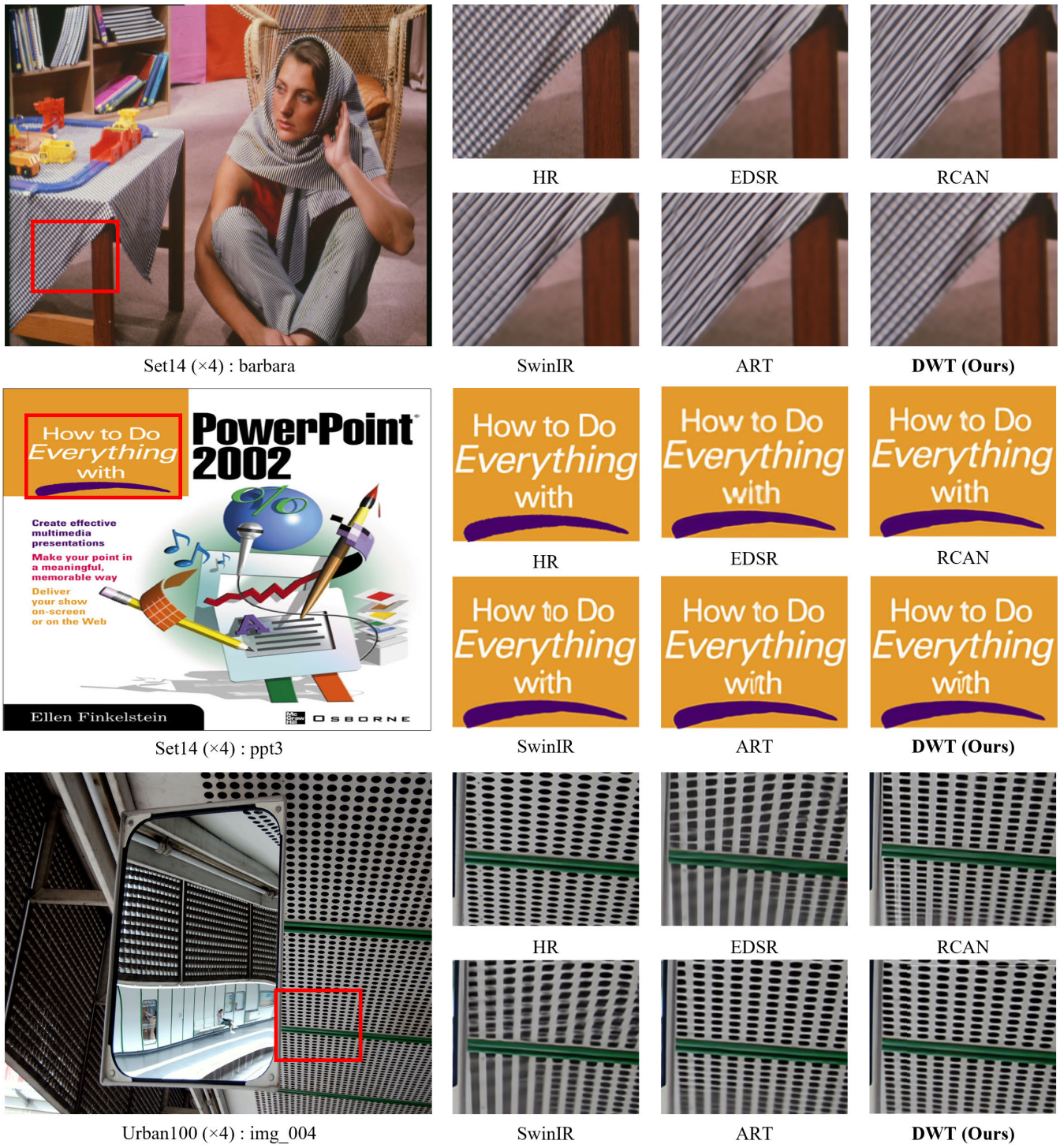


FIGURE 5. Visual comparison (x4) with numerous state-of-the-art SR methods on Set14 and Urban100 datasets. The red boxed areas have been cropped from the results and enlarged for better visibility.

as a rough reference. Unlike the FR-IQA results, the NR-IQA results show inconsistent performance across methods, datasets, and scales. However, our DWT shows competitive performance in ×4 SR results, particularly in terms of the BRISUQE. Especially, unlike the FR-IQA results, the BRISQUE of DWT on the Manga109 dataset is the best at all scales. Although DWT is not the best in both scores, as can be seen from the visual comparison results in the Fig. 5 and Fig. 6, our DWT shows impressive results in terms of human visual perception.

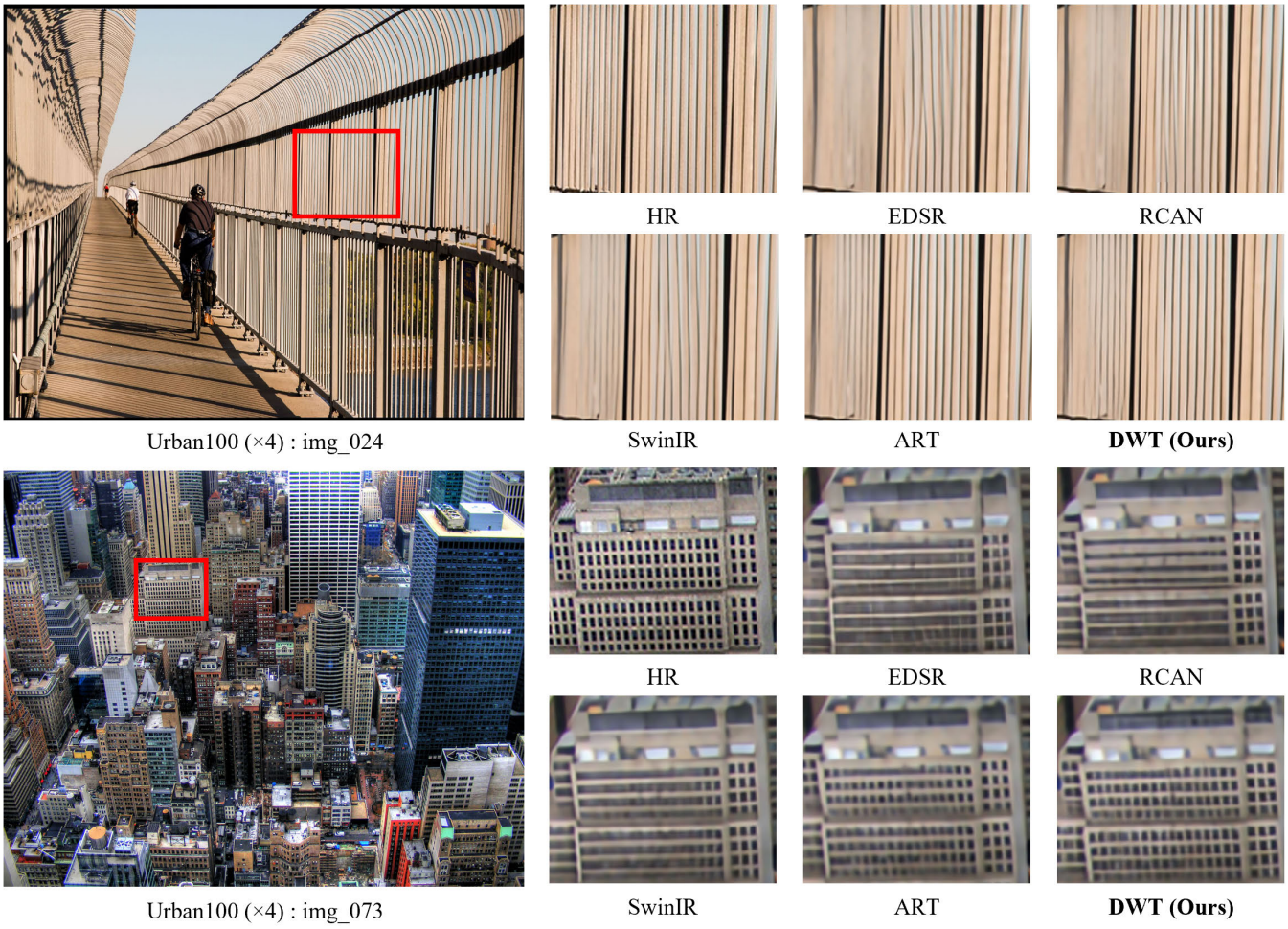


FIGURE 6. Visual comparison (×4) with numerous state-of-the-art SR methods on Urban100 dataset. The red boxed areas have been cropped from the results and enlarged for better visibility.

E. VISUAL COMPARISON

We also provide visual comparison with state-of-the-art methods in Fig. 5 and Fig. 6. These results demonstrate that DWT generates clearer textures and restores high-frequency details, leading to sharper edges when compared to other methods.

In particular, the Urban100 dataset's "img_004", "img_024", and "img_073" images serve as excellent examples that highlight the strengths of our DWT. We observe that these images contain comparable patterns that can be used as points of reference when restoring the areas marked with a red box. Despite containing repetitive patterns in the image, most other methods struggle to recover clear structures and tend to generate blurry outcomes. In comparison, our DWT recovers more details while reducing blurring artifacts. For "img_004", DWT utilizes information from both neighboring and distant regions through its dilation strategy to produce results almost identical to the original. We can find similar behavior on "img_024" in the Urban100 dataset. The red boxed area in "img_024" is composed of repeated vertical lines. However, it can be observed that SwinIR generates blurry results by failing to accurately restore most of the vertical lines. In contrast, DWT is able to restore the vertical lines relatively sharply. The result of "image_073" also highlights that DWT restores the building's windows more clearly than other models. Overall, our findings demonstrate the superiority of the proposed dilation strategy in producing high-quality SR results.

V. CONCLUSION

In this paper, we propose a novel dilated window transformer, DWT, for image SR that aims to address the limitations of window-based self-attention. Without introducing additional computational cost, we employ a dilation strategy to expand the receptive field more quickly and effectively. This simple yet efficient strategy enables our DWT to extract both local and global features, leading to improved performance in image SR. Extensive experiments under numerous benchmark datasets show the effectiveness of our proposed DWT. Notably, DWT records the state-of-the-art SR performance in

terms of both quantitative and qualitative evaluations with a competitive number of parameters and reasonable computational cost.

REFERENCES

[1] 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.

[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] 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.

[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] T. Dai, J. Cai, Y. Zhang, S. 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.

[6] 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. 1680–1689.

[7] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[8] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 63–79.

[9] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[11] 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*.

[12] 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. (CVPR)*, Jun. 2021, pp. 12889–12899.

[13] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "LocalViT: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 213–229.

[15] 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.

[16] 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.

[17] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12159–12168.

[18] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision—ECCV 2022 Workshops*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 205–218.

[19] 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.

[20] J. Cao, Y. Li, K. Zhang, J. Liang, and L. Van Gool, "Video super-resolution transformer," 2021, *arXiv:2106.06847*.

[21] 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.

[22] 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*.

[23] 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.

[24] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," 2022, *arXiv:2205.04437*.

[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[26] 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.

[27] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 184–199.

[28] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 191–207.

[29] 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.

[30] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.

[31] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 3499–3509.

[32] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim, "Enriched CNN-transformer feature aggregation networks for super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4945–4954.

[33] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*.

[34] 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.

[35] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[36] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.

[37] G. Huang, Y. Wang, K. Lv, H. Jiang, W. Huang, P. Qi, and S. Song, "Glance and focus networks for dynamic visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4605–4621, Apr. 2023.

[38] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

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

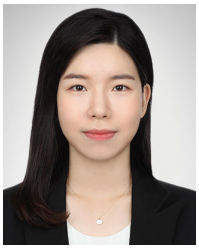
[40] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.

[41] 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.

[42] 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.

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

[44] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. 7th Int. Conf. Curves Surf.*, Avignon, France. Berlin, Germany: Springer, 2012, pp. 711–730.
[45] 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, Jul. 2001, pp. 416–423.
[46] J. 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.
[47] 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.
[48] 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.
[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[50] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.
[51] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Referenceless image spatial quality evaluation engine," in *Proc. 45th Asilomar Conf. Signals, Syst. Comput.*, vol. 38, Nov. 2011, pp. 53–54.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[53] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop Autodiff*, 2017.
[54] J. W. Soh, G. Y. Park, J. Jo, and N. I. Cho, "Natural and realistic single image super-resolution with explicit natural manifold discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8114–8123.
[55] A. Lugmayr, M. Danelljan, and R. Timofte, "NTIRE 2020 challenge on real-world image super-resolution: Methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2058–2076.



SOOBIN PARK was born in Republic of Korea, in 1996. She received the B.S. degree in IT convergence from Hanyang University, Seoul, South Korea, in 2022, where she is currently pursuing the M.S. degree with the Department of Artificial Intelligence. Her research interests include computer vision and image super-resolution.

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. From 1997 to 2001, he was with the Telecommunication Research Laboratory, Samsung Electronics Company. In 2001, he joined Hanyang University, Seoul, where he is currently a Professor with the Department of Computer Science and Engineering. His research interests include ontology, knowledge-based systems, computer vision, dialogue generation systems, social media analysis and visualization, and multi-modal.