# Lightweight Wavelet-Based Transformer for Image Super-Resolution

Jinye Ran and Zili Zhang(✉)

College of Computer and Information Science, Southwest University,
Chongqing 400715, China
zhangzl@swu.edu.cn

**Abstract.** Suffering from the inefficiency of deeper and wider networks, most remarkable super-resolution algorithms cannot be easily applied to real-world scenarios, especially resource-constrained devices. In this paper, to concentrate on fewer parameters and faster inference, an end-to-end Wavelet-based Transformer for Image Super-resolution (WTSR) is proposed. Different from the existing approaches that directly map low-resolution (LR) images to high-resolution (HR) images, WTSR also implicitly mines the self-similarity of image patches by a lightweight Transformer on the wavelet domain, so as to balance the model performance and computational cost. More specifically, a two-dimensional stationary wavelet transform is designed for the mutual transformation between feature maps and wavelet coefficients, which reduces the difficulty of mining self-similarity. For the wavelet coefficients, a Lightweight Transformer Backbone (LTB) and a Wavelet Coefficient Enhancement Backbone (WECB) are proposed to capture and model the long-term dependency between image patches. Furthermore, a Similarity Matching Block (SMB) is investigated to combine global self-similarity and local self-similarity in LTB. Experimental results show that our proposed approach can achieve better super-resolution performance on the multiple public benchmarks with less computational complexity.

**Keywords:** Transformer · Lightweight network · Wavelet transform · Image super-resolution

## 1 Introduction

Single image super-resolution (SISR) aims to restore the high-resolution (HR) image corresponding to the low-resolution (LR) image. As a low-level computer vision task, SISR enjoys a wide range of applications in many fields, such as remote sense [36], surveillance [39], medical imaging [28], and security [16], amongst others. Essentially, SISR is an ill-posed problem since there are always

S. Khanna et al. (Eds.): PRICAI 2022, LNCS 13631, pp. 368–382, 2022.
https://doi.org/10.1007/978-3-031-20868-3_27

infinite HR images degrading to the same LR image. To minimize the uncertainty, recently, extensive methods [2,8,15,32,33] based on deep neural networks have been proposed and have achieved remarkable performance on many public benchmarks. However, to improve the quality of super-resolution, most mainstream methods concentrate on developing a deeper and wider neural network and neglect the lightweight problem, which may limit the development of super-resolution networks in some resource-constrained devices, like edge computing devices.

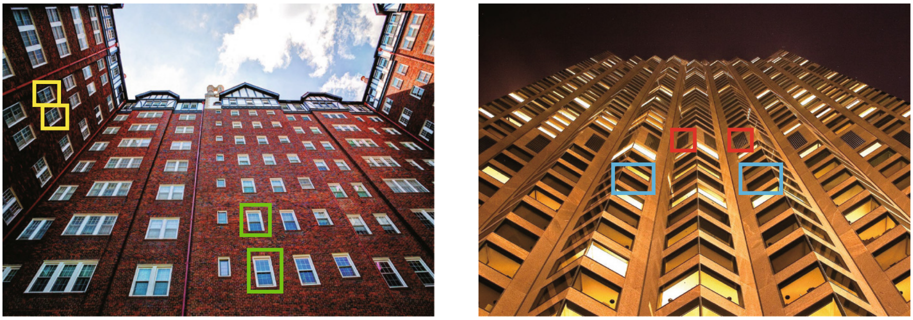

**Fig. 1.** Self-similar information in the image (the same color bounding box region), which can reduce the difficulty of modeling the ill-posed problem. (Color figure online)

Although lightweight and performance are generally considered as a trade-off problem in SISR, there are still two feasible directions to address these issues. One is to build models with a parameter sharing strategy, such as recurrent learning [20,31] and recursive learning [30,33]. The other is to design some elaborate architectures, such as wide activation [37], group convolution [2], and information distillation [12]. However, the above two methods have disadvantages such as long inference time and complicated design of structures, which can not meet universal super-resolution application scenarios. Recently, [27] reported that mining the self-similar information in the images could greatly improve the super-resolution results. As shown in Fig. 1, the ill-posed problems in super-resolution would be greatly alleviated by referring to other similar image patches. Furthermore, we observed that the working principle (self-attention mechanism) of Transformer could effectively capture the self-similarity between the image patches, which encouraged us to build a lightweight super-resolution network by it. However, the existing Transformers cannot be directly used for pixel-level reconstruction tasks because they consume a lot of computing and memory resources, which are not conducive to constructing lightweight super-resolution networks.

To remedy this defect, we make a trade-off between the model performance and computational cost of the Transformer and propose a Wavelet-based Transformer for Image Super-Resolution (WTSR). As shown in Fig. 2, the proposed network is conducted on the wavelet domain because the wavelet transform

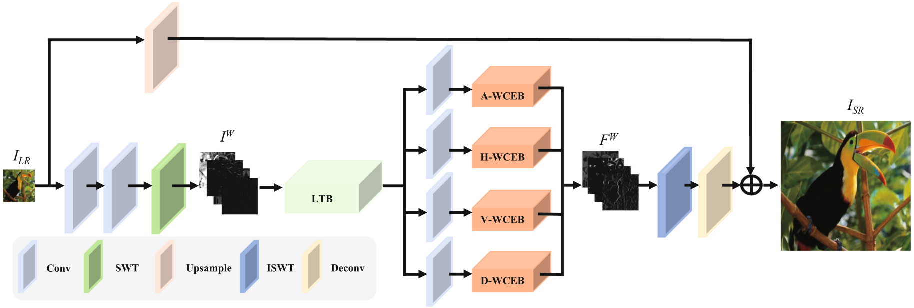

**Fig. 2.** The network of the proposed wavelet-based Transformer for super-resolution. The letters before WCEB denote different ABs for different wavelet coefficient.

can efficiently generate detailed information of images, which is beneficial for the mining of self-similarity in the images and the lightweight design of the model. Meanwhile, considering that the resolution of wavelet coefficients can significantly influence the difficulty of modeling, we finally decide to learn the mapping between LR and HR on the stationary wavelet domain. In addition, WTSR also includes a Lightweight Transformer Backbone (LTB) and a Wavelet Coefficient Enhancement Backbone (WCEB). For the LTB, inspired by [22] and considering lightweight, an Efficient Transformer (ET) encoder that adopts the design of wavelet coefficients partitioning and channel compression is proposed, which can significantly alleviate the requirements on hardware resources. Meanwhile, noting the effect of the partitioning method and partitioning size on self-similarity modeling, a Similarity Matching Block (SMB) is proposed to connect two ET encoders with different partitioning sizes. In other words, SMB combines global and local self-similarity to overcome the network receptive field constraints imposed by partitioning. For the WCEB, recovering each wavelet coefficient's structural information and channel redundancy is its main target. In particular, four WCEBs with different Asymmetric Blocks (ABs) are utilized to reconstruct corresponding wavelet coefficients, such as $1 * 3$ AB for the horizontal wavelet coefficient. For then, a coordinate attention layer [10] is used to balance the lightweight and performance. To make WTSR trainable end-to-end, the real stationary wavelet transform is realized by referring to the discrete wavelet transform based on the PyTorch Library.

The main contributions of this paper are summarized as follows:

- An end-to-end WTSR is proposed to solve the lightweight problem in the super-resolution. The competitive super-resolution results are achieved through a two-dimensional stationary wavelet transform and Transformer.
- A LTB is proposed to capture the long-term dependency between image patches in the images. Meanwhile, a SMB and an ET are designed to balance the performance of the model and the consumption of computing resources.

- Different WECBs are proposed to recover different wavelet coefficient's structural information and channel redundancy. Notably, four ABs for different wavelet coefficient characteristics are investigated.

The rest of the paper is organized as follow: Sect. 2 introduces the related work of this paper. Section 3 elaborates the details of our proposed approach. Section 4 presents the experimental particulars and results, along with the ablation experiments for each part of the method, and Sect. 5 concludes the paper.

## 2 Related Work

Our approach is closely related to wavelet transform for image super-resolution and Transformer in the computer vision. A brief introduction to these two aspects will be presented in this section.

### 2.1 Wavelet Transform for Image Super-Resolution

Wavelet transform has been widely employed for image super-resolution as an essential technology in traditional image processing [3,13]. With the arrival of the wave of neural networks, the wavelet transform, as a powerful content and texture representation tool, has been diffusely employed in many deep learning-based super-resolution networks. Kumar et al. [18] proposed a three-layer convolution neural network to map the wavelet coefficients of LR images to those of HR images. Inspired by encoder-decoder architecture and wavelet inverse transform, Liu et al. [21] focused on multi-level wavelet transform and proposed a multi-level wavelet convolution. By imposing constraints on the network in the wavelet domain, the performance of super-resolution reconstruction was greatly improved. Xue et al. [34] introduced the attention mechanism into the residual network based on wavelet transform and proposed a wavelet-based residual attention network. Significant performance was gained across multiple baselines with a well-designed multi-kernel network. Zhang et al. [38] and Xin et al. [33] designed recurrent and recursive structures in image super-resolution tasks. Through the parameter sharing mechanism and wavelet transform, the computational complexity of the model was remarkably reduced while maintaining the performance of the model. Above mentioned works, the resolution of the feature maps before and after the wavelet transform is different (the resolution of feature map is half after wavelet transform). For the image generation tasks, like super-resolution, generally, this means that the parameters of network will increase. To some extend, it is not conducive to lightweight.

### 2.2 Transformer in the Computer Vision

Transformer has become the most important technology in natural language processing due to its powerful model capability and efficient parallel capability. Recently, some high-level computer vision tasks have also tried to utilize the

Transformer as a backbone in the network to extract features, such as visual recognition [7,29] and object detection [5,24]. Carion et al. [5] proposed an end-to-end object detection with Transformer, which embedded images as sequential data and sent them to encoder and decoder, creating a precedent for Transformers in computer vision applications. Dosovitskiy et al. [7] proposed a Transformer for image recognition at scale, which greatly reduced the computational resource consumption of the Transformer by partitioning and serializing the image into a combination of some patches. Liu et al. [22] added the hierarchical structure to the Transformer and proposed Swin Transformer. As a new backbone in high-level computer vision tasks, it further reduced the computational complexity of the Transformer and could extract more useful feature information. Obviously, reducing the computation cost of Transformer is one of the most important problems in its computer vision applications. However, low-level computer vision tasks are generally more computation-intensive. In the field of super-resolution, Lu et al. [23] proposed an efficient Transformer for SISR by the lightweight design of the Transformer for image super-resolution. Meanwhile, they also designed a lightweight convolution backbone to reduce the need for a large amount of data for Transformer training. In other words, Transformer is still relatively underused on the super-resolution task due to the lightweight issue.

## 3 Proposed Method

### 3.1 Overall Network

As shown in Fig. 2, our WTSR mainly consists of six parts: shallow feature extractor, two-dimensional Stationary Wavelet Transform (SWT), LTB, WCEB, two-dimensional Inverse Stationary Wavelet Transform (ISWT) and reconstruction block. The input and output of WTSR are defined as $I_{LR}$ and $I_{SR}$. As a result, the shallow feature $F_0$ is extracted from $I_{LR}$ by two convolution layers:

$$F_0 = f_{1*1}(f_{3*3}(I_{LR})) \tag{1}$$

where $f_{1*1}$ and $f_{3*3}$ denote a $1*1$ convolution layer and a $3*3$ convolution layer, respectively. Then, $F_0$ is fed to a two-dimensional SWT to obtain the stationary wavelet coefficients for each channel, which can be formulated as:

$$I^W = concat(SWT(F_0)) \tag{2}$$

$I^W$ represents the stationary wavelet coefficients after concatenation. All outputs of SWT are sent to LTB to capture and model the long-term dependency between the image patches:

$$F_L = \phi^5(\psi(\phi^8(f_{group}(I^W)))) \tag{3}$$

where $f_{group}$, $\phi$ and $\psi$ denote a group convolution, ET encoder and SMB. The superscript of $\phi$ indicates the partitioning size. The output of LTB is split by a split layer and then sent to WCEBs:

$$F_A, F_H, F_V, F_D = split(F_L) \tag{4}$$

$$F^W = concat(\sigma_{A,H,V,D}(F_A, F_H, F_V, F_D)) \tag{5}$$

where $\sigma_{A,H,V,D}$ denotes four different WCEBs, and $F^W$ stands for the wavelet coefficient after recovering the redundancy of channels and the structure information of different wavelet coefficients. Then, all outputs of WCEBs are concatenated and passed to two-dimensional ISWT:

$$F_D = ISWT(F^W) \tag{6}$$

Finally, $F_D$ and $I_{LR}$ are sent into the reconstruction block simultaneously to get $I_{SR}$:

$$I_{SR} = f_{Deconv}(f_{3*3}(F_d)) + f_{up}(I_{LR}) \tag{7}$$

where $f_{Deconv}$, $f_{3*3}$, and $f_{up}$ express a deconvolution layer, a $3*3$ convolution layer, and an upsample layer, respectively.

### 3.2 Lightweight Transformer Backbone

The LTB in Fig. 3 (a) consists of two ET encoders in Fig. 3 (b) with different partitioning size and a SMB. There are some modifications in ET encoder to make it more lightweight than the standard Transformer. Suppose the input wavelet coefficient $S_i$ has the shape of $B \times C \times H \times W$, where $B$, $C$, $H$, and $W$ denote batch size, channels, height, and width of wavelet coefficients. Firstly, a reduction layer is employed to reduce the number of channels by quarter $(B \times C_1 \times H \times W, C_1 = \frac{C}{4})$. Then, a partitioning layer with a partitioning size of K is utilized to reduce the resolution of sub-bands $((K * K * B) \times C_1 \times \frac{H}{K} \times \frac{W}{K})$. After that, a standard Transformer, including Normalization layer, Multi-Head Attention, and Multi-Layer Perceptrons, is used to capture and model the long-term dependencies. Finally, a reverse layer is designed to restore original resolution of wavelet coefficients $(B \times C_1 \times H \times W, C_1 = \frac{C}{4})$. Assume the output wavelet coefficients are $S_o$, and the output wavelet coefficients $S_o$ can be obtained by:

$$S_{m1} = f_{partitioning}(f_{reduction}(S_i)) \tag{8}$$

$$S_{m2} = MHA(Norm(S_{m1})) + S_{m1} \tag{9}$$

$$S_o = f_{reverse}(MLP(Norm(S_{m2})) + S_{m2}) \tag{10}$$

where MHA(·), Norm(·), and MLP(·) represent Multi-Head Attention, Normalization layer, and Multi-Layer Perceptrons in the standard Transformer respectively. Meanwhile, $f_{partitioning}$, $f_{reduction}$ and $f_{reverse}$ denote a wavelet coefficients partitioning layer, a reduction layer and a wavelet coefficients reverse layer.

Although the ET encoder reduces a lot of computational resource consumption compared to the standard Transformer, the partitioning method and partitioning size have a great impact on the mining of self-similarity in the images. To deal with this issue, two ET encoders with different partitioning sizes are connected through a SMB, which can combine local self-similarity and global self-similarity. Notably, the reduction layer in the second ET encoder is dropped. For SMB, an unfold layer with both kernel size and stride size of M is adopted to unfold the output of the first ET encoder into multiple patches of shape $\mathrm{B} \times \mathrm{M} * \mathrm{M} \times 1$. Then, the most similar patch except itself for each patch is found through the normalized inner product and a new wavelet coefficient is generated, which can be formulated as:

$$p_{i1c,j1c} = \arg\max_{p_{i2,j2}} \langle \frac{p_{i1,j1}}{\|p_{i1,j1}\|}, \frac{p_{i2,j2}}{\|p_{i2,j2}\|} \rangle \quad s.t. \quad |i1 - i2| + |j1 - j2| \neq 0 \tag{11}$$

where $p_{i1c,j1c}$ denotes the patch in the new wavelet coefficient, $p_{i1,j1}$ and $p_{i2,j2}$ represent the patch in the original wavelet coefficient. Finally, a concatenation layer and two convolution layers are applied to integrate and map the features.

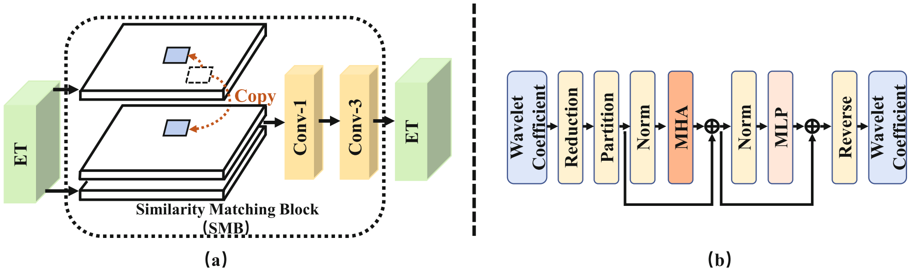

**Fig. 3.** (a) The structure of LTB in the WTSR. The dashed line area is SMB. (b) The structure of ET encoder in the LTB.

### 3.3 Stationary Wavelet Transform

The core of Transformer in computer vision applications is to model serialized image patches. From the perspective of super-resolution, it can implicitly mine self-similarity in the images. Meanwhile, there is no doubt that improving the embedding quality of images can effectively help the convergence of Transformer. As an efficient technique, wavelet transform depicts the contextual and textural information of an image without increasing parameters, which is beneficial for building a lightweight Transformer. For then, our experiments have shown that the resolution of wavelet coefficients also significantly influences the performance of lightweight Transformer, so WTSR learns the mapping between LR images and HR images on the stationary wavelet domain. Notably, stationary wavelet transform can also guarantee translation invariance on features and the stability of two-dimensional inverse stationary wavelet transformation (More details can be seen in the ablation study).

The process of stationary wavelet transform is shown in Fig. 4. $F_0$ extracted by the shallow extractor is filtered two times in different directions by a particular high-pass filter and a particular low-pass filter, yielding four distinct sets of wavelet coefficients, namely approximation (A), horizontal (H), vertical (V), diagonal (D), respectively. For the two-dimensional inverse stationary wavelet transform, by reversing the process and module of the two-dimensional stationary wavelet transform, almost the same data as before transformation can be obtained without changing any wavelet coefficients. In addition, the “Haar” kernel is employed to improve the efficiency of wavelet transform.

### 3.4 Wavelet Coefficient Enhance Backbone

The WCEB in Fig. 5 consists of two ABs, a $1*1$ convolution layer, and a coordinate attention layer [10]. For the AB, four particular convolution layers are designed to process wavelet coefficients with different characteristics. More specifically, approximation, horizontal, vertical, and diagonal coefficients are processed by a $3*3$ convolution layer, a $1*3$ asymmetric convolution layer, a $3*1$ asymmetric convolution layer, and an oblique asymmetric convolution layer. For the $1*1$ convolution layer, the channels of the wavelet coefficients after a concatenation layer are reduced by half. For the coordinate attention layer, it is an efficient and lightweight attention block that can improve the expressiveness of the model. For more details, suppose that one of the WCEBs has an input $M$ with a shape of $B \times C \times H \times W$. Firstly, an AB layer and a PRule layer are used to reconstruct a preliminary wavelet-specific structure without changing the number of channels and spatial resolution of the wavelet coefficients. Secondly, a concatenation layer is employed to contact the output and $M$, where the wavelet coefficients have a shape size of $B \times 2C \times H \times W$. Next, a $1*1$ convolution layer is utilized to compress the channel information, giving the wavelet coefficients the shape of $B \times C \times H \times W$. Then, a coordinate attention layer, a PRule layer, and an AB layer are used in turn for the more superior wavelet structure reconstruction. Finally, to constrain the WCEB modeling performance, the input M is added to the final output. The final output shape of the WCEB is $B \times C \times H \times W$.

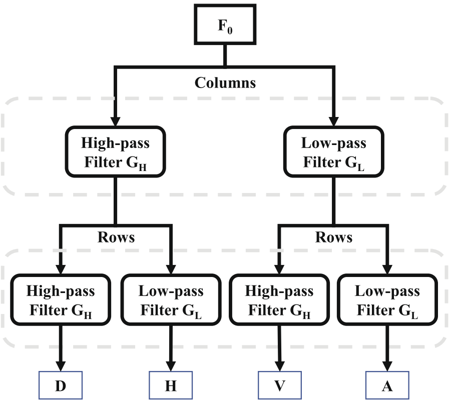

**Fig. 4.** Process of the two-dimensional stationary wavelet transform.

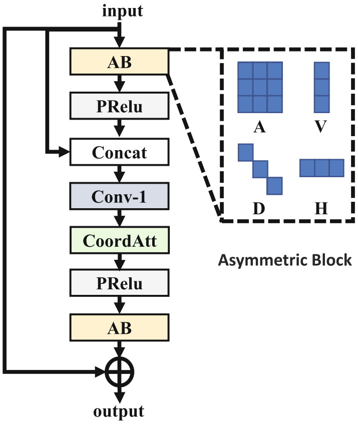

**Fig. 5.** The structure of WECB in the WTSR.

# 4 Experiments

## 4.1 Datasets and Metrics

To follow the previous literature [20], Flickr2K and DIV2K [1] are used as our training data. In our experiments, multiple data augmentations are adopted to make full use of training data, including image rotation, image flipping, and image scaling. For evaluation, five standard benchmark datasets: Set5 [4], Set14 [35], BSD100 [25], Urban100 [11], and Manga109 [26] are utilized. Meanwhile, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) are used to evaluate the performance of the reconstructed super-resolution images. Furthermore, the results are calculated on the Y channel of the YCbCr color space.

## 4.2 Implementation Details

All WTSRs in experiments are trained with a mini-batch size of 64 for 1000 epochs. To exploit self-similarity more in the images, a batch of RGB images are cropped to 80 × 80 patch size randomly and sent to the network in each step. Warmup [9] strategy is employed to increase the learning rate to $4 \times 10^-3$ after 10 iterations and decrease half for every 200 epochs. The network parameters are initialized according to [14] and optimized by Adam [17]. Meanwhile, L1 Loss is used as the loss function for our network training to avoid over-smoothing. Our experiments are run on the PyTorch platform with a single RTX3090 GPU.

## 4.3 Evaluation

Since our WTSR mainly focuses on implementing a lightweight and efficient super-resolution network, we compare it to other super-resolution methods with parameters within 1M, including Bicubic, SRCNN [32], FSRCNN [6], VDSR [15], DWSR [8], LapSRN [19], CARN-M [2], MemNet [31], SRFBN-S [20] and WDRN-S [33]. Many of them achieve competitive super-resolution quality with an efficient and lightweight network. The quantitative evaluation performance are presented in Table 1. The highest score is represented by a highlighted number, while the second-highest score is represented by an underlined number. The comparison results show that our WTSR has a very obvious advantage over the comparative approaches on five public benchmarks in ×4 scale factor. Meanwhile, our WTSR also achieves a competitive result in ×2 and ×3 scale factors. Notably, despite having fewer parameters than our WTSR, the PSNR/SSIM results of SRCNN, FSRCNN, and DWSR lag far behind WTSR. As shown in Fig. 6, we also visualize the trade-off analysis between the number of the model parameters and the model performance among these lightweight super-resolution networks. Under the condition of model parameters less than 1M, our WTSR is a better choice than SRFBN-S, WDSR-S, and CARN-M. So, all experiments fully give evidence that our WTSR achieves a better trade-off between model sizes and model performance.

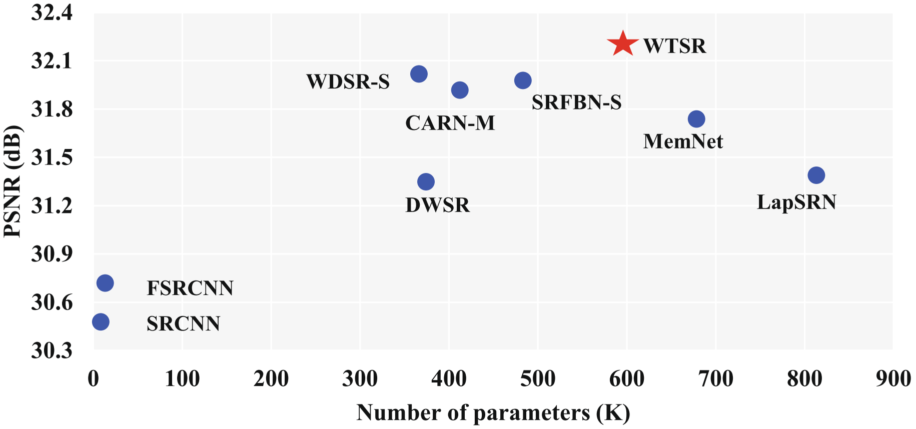

**Fig. 6.** The relationship between model sizes and model performance of different lightweight super-resolution networks on the Set5 with ×4 scale factor.

In Fig. 7, we provide a super-resolution visual comparison between WTSR and other lightweight super-resolution networks on the ×4 scale factor. Since WTSR can implicitly mine the self-similarity from the whole LR, the super-resolution results restored by our WTSR have more sophisticated particulars in some with multiple repeating structures, especially in the edges and lines.

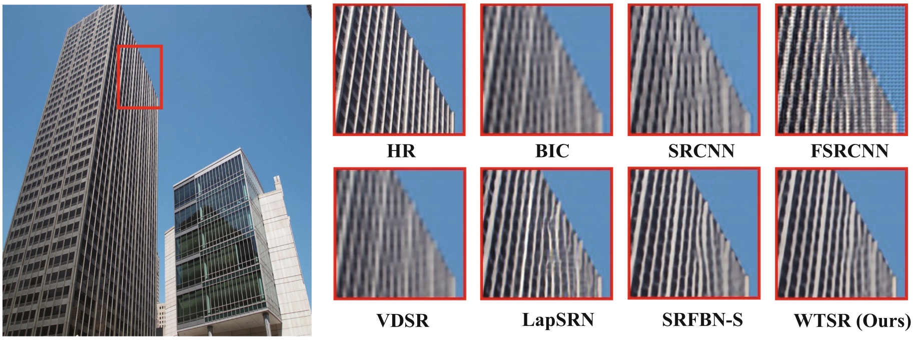

**Fig. 7.** Subjective visual quality compared with other super-resolution networks for ×4 scale factor on the pictures from BSD100.

### 4.4 Ablation Study

To explore the impact of embedding quality on Transformer, three versions of WTSR are trained. The first version (Version 1) is to use Discrete Wavelet Transform (DWT) to process the feature map, the second version (Version 2) is to

**Table 1.** Quantitative results of WTSR compared with other lightweight super-resolution network, the best model performance is **highlighted** and the second performance is underlined.

| Methods | Scales | Params | Set5 | Set14 | BSD100 | Urban100 | Manga109 |
|---|---|---|---|---|---|---|---|
| | | | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM |
| Bicubic | x2 | - | 33.66/0.930 | 30.24/0.869 | 29.56/0.843 | 26.88/0.840 | 30.30/0.934 |
| SRCNN [32] | | 8K | 36.66/0.954 | 32.45/0.907 | 31.36/0.888 | 29.50/0.895 | 35.60/0.966 |
| FSRCNN [6] | | 13K | 37.00/0.956 | 32.63/0.909 | 31.53/0.892 | 29.88/0.902 | 36.67/0.971 |
| VDSR [15] | | 666K | 37.53/0.959 | 33.03/0.912 | 31.90/0.896 | 30.76/0.914 | 37.22/0.975 |
| DWSR [8] | | 374K | 37.43/0.957 | 33.07/0.911 | 31.80/0.894 | 31.46/0.916 | -/- |
| LapSRN [19] | | 813K | 37.52/0.959 | 32.99/0.912 | 31.80/0.895 | 30.41/0.910 | 37.27/0.974 |
| MemNet [31] | | 678K | 37.78/0.960 | 33.28/0.914 | 32.08/0.898 | 31.31/0.920 | 37.72/0.974 |
| CARN-M [2] | | 412K | 37.53/0.958 | 33.26/0.914 | 31.92/0.896 | 31.23/0.919 | -/- |
| SRFBN-S [20] | | 483K | 37.78/0.960 | 33.35/0.916 | 32.00/0.897 | 31.41/0.921 | 38.06/0.976 |
| WDRN-S [33] | | 344K | 37.93/**0.961** | 33.42/**0.916** | 32.08/**0.898** | 31.80/0.924 | -/- |
| WTSR | | 528K | **37.95**/0.961 | **33.51**/0.915 | **32.09**/0.893 | **31.91/0.928** | **38.42/0.976** |
| Bicubic | x3 | - | 30.39/0.868 | 27.55/0.774 | 27.21/0.739 | 24.46/0.735 | 26.95/0.856 |
| SRCNN [32] | | 8K | 32.75/0.909 | 29.30/0.822 | 28.41/0.786 | 26.24/0.799 | 30.48/0.912 |
| FSRCNN [6] | | 13K | 33.18/0.914 | 29.37/0.824 | 28.53/0.791 | 26.43/0.808 | 31.10/0.921 |
| VDSR [15] | | 666K | 33.66/0.921 | 29.77/0.831 | 28.82/0.798 | 27.14/0.828 | 32.01/0.934 |
| DWSR [8] | | 374K | 33.82/0.922 | 29.83/0.831 | -/- | -/- | -/- |
| LapSRN [19] | | 813K | 33.81/0.922 | 29.79/0.833 | 28.82/0.798 | 27.07/0.828 | 32.21/0.935 |
| MemNet [31] | | 678K | 34.09/0.925 | 30.00/0.835 | 28.96/0.800 | 27.56/0.838 | 32.51/0.937 |
| CARN-M [2] | | 412K | 33.99/0.924 | 30.08/0.837 | 28.91/0.800 | 27.55/0.839 | -/- |
| SRFBN-S [20] | | 483K | 34.20/0.926 | 30.10/**0.837** | 28.96/0.801 | 27.66/0.842 | 33.02/0.94 |
| WDRN-S [33] | | 366K | 34.18/0.925 | **30.17**/0.837 | 28.98/0.802 | **27.82/0.844** | -/- |
| WTSR | | 558K | **34.27/0.925** | 30.12/0.836 | **28.98/0.802** | 27.69/0.841 | **33.11/0.941** |
| Bicubic | x4 | - | 28.42/0.810 | 26.00/0.703 | 25.96/0.668 | 23.14/0.658 | 24.89/0.787 |
| SRCNN [32] | | 8K | 30.48/0.863 | 27.50/0.751 | 26.90/0.710 | 24.52/0.722 | 27.58/0.856 |
| FSRCNN [6] | | 13K | 30.72/0.866 | 27.61/0.755 | 26.98/0.715 | 24.62/0.728 | 27.90/0.861 |
| VDSR [15] | | 666K | 31.35/0.884 | 28.01/0.767 | 27.29/0.725 | 25.18/0.752 | 28.83/0.887 |
| DWSR [8] | | 374K | 31.39/0.883 | 28.04/0.767 | 27.25/0.724 | 25.26/0.755 | -/- |
| LapSRN [19] | | 813K | 31.54/0.885 | 28.09/0.770 | 27.32/0.728 | 25.21/0.756 | 29.09/0.890 |
| MemNet [31] | | 678K | 31.74/0.889 | 28.26/0.772 | 27.40/0.728 | 25.50/0.763 | 29.42/0.894 |
| CARN-M [2] | | 412K | 31.92/0.890 | 28.42/0.776 | 27.44/0.730 | 25.62/0.769 | -/- |
| SRFBN-S [20] | | 483K | 31.98/0.892 | 28.45/0.778 | 27.44/0.731 | 25.71/0.772 | 29.91/0.901 |
| WDRN-S [33] | | 366K | 32.02/0.890 | 28.47/0.774 | 27.47/0.730 | 25.82/0.776 | -/- |
| WTSR | | 593K | **32.16/0.895** | **28.57/0.781** | **27.56/0.735** | **26.03/0.784** | **30.44/0.908** |

employ SWT, and the last one (Version 3) does not utilize wavelet transform. The experimental results in Table 2 indicate that WTSR based on the SWT performs better than the other two in all benchmarks without any additional parameters. By comparing the experimental results of Version 2 and Version 3, it can be found that the feature map processed by wavelet transform at the same resolution is more conducive to the convergence of WTSR. Equally, wavelet transform is beneficial for Transformer to mine self-similarity in the images. By comparing the experimental results of Version 1 and Version 2, SWT has a noticeable promotion effect on the performance of super-resolution reconstruction, instead of DWT is harmful to the performance of super-resolution reconstruction. This is because the SWT does not change the resolution of the feature map, which guarantees translation invariance on the features and the stability of inverse wavelet transformation. Furthermore, SWT can greatly reduce the modeling difficulty of WTSR, which is beneficial for lightweight.

**Table 2.** Comparisons on PSNR/SSIM of WTSR with different wavelet transform. Best results are **highlighted**.

| Wavelet transform | Params | PSNR/SSIM | | | | |
|---|---|---|---|---|---|---|
| Type | | Set5 | Set14 | BSD100 | Urban100 | Manga109 |
| None | 593K | 32.01/0.893 | 27.47/0.781 | 27.51/0.734 | 25.84/0.777 | 20.22/0.905 |
| DWT | 593K | 27.56/0.790 | 25.51/0.682 | 25.54/0.647 | 22.69/0.635 | 24.19/0.767 |
| SWT | 593K | **32.16/0.895** | **28.57/0.781** | **27.56/0.735** | **26.03/0.784** | **30.44/0.908** |

To study the effect of different LTB structures on super-resolution reconstruction performance, another five experiments are designed by changing the partitioning size and arrangement order of the ET encoder. In Table 3, it can be found that using two ET encoders with different partitioning sizes can obtain better super-resolution reconstruction performance than using two ET encoders with the same partitioning size, and the SMB significantly improves the super-resolution reconstruction results. More details, in the case of two ET encoders with the same partitioning size, the impact of partitioning size on super-resolution reconstruction results is not apparent. On the contrary, in the case of two ET encoders with different partitioning sizes, the ET with a larger partitioning size is ranked first, which can obtain better super-resolution performance in all benchmarks.

**Table 3.** Comparisons on PSNR/SSIM of WTSR with different network of LTB. Best results are **highlighted**. The number after T indicates the partitioning size of ET encoder, S represents SMB, and the arrow denotes the direction of data flow.

| The network of LTB | Params | PSNR/SSIM | | | | |
|---|---|---|---|---|---|---|
| | | Set5 | Set14 | BSD100 | Urban100 | Manga109 |
| T5 → T5 | 567K | 31.92/0.892 | 27.43/0.779 | 27.46/0.734 | 25.78/0.777 | 30.06/0.904 |
| T8 → T8 | 569K | 31.96/0.892 | 28.46/0.779 | 27.48/0.733 | 25.79/0.776 | 30.09/0.904 |
| T5 → T8 | 568K | 31.97/0.893 | 28.44/0.779 | 27.48/0.733 | 25.79/0.776 | 30.04/0.903 |
| T8 → T5 | 568K | 32.00/0.893 | 28.47/0.779 | 27.50/0.733 | 25.87/0.778 | 30.11/0.904 |
| T8 → S → T5 | 593K | **32.16/0.895** | **28.57/0.781** | **27.56/0.735** | **26.03/0.784** | **30.44/0.908** |

To investigate the impact of different WCEB on super-resolution performance, we set the usage of WCEM on different wavelet coefficients and conduct five experiments in Table 4. By comparison, it demonstrates that using different WCEMs on different wavelet coefficients can significantly improve the super-resolution reconstruction quality of WTSR in all benchmarks.

**Table 4.** Study the effect of each WCEB on PSNR/SSIM, Best results are **highlighted**.

| WCEM type | Params | PSNR/SSIM | | | | |
|---|---|---|---|---|---|---|
| | | Set5 | Set14 | BSD100 | Urban100 | Manga109 |
| None | 416K | 31.91/0.892 | 28.44/0.778 | 27.47/0.732 | 25.75/0.7735 | 29.92/0.901 |
| A | 482K | 32.01/0.893 | 28.42/0.779 | 27.48/0.734 | 25.80/0.7782 | 29.99/0.924 |
| A + H | 519K | 32.06/0.894 | 28.54/0.780 | 27.52/0.733 | 25.91/0.7791 | 30.27/0.905 |
| A + H + V | 556K | 32.10/0.894 | 28.54/0.781 | 27.54/0.735 | 25.97/0.7823 | 30.31/0.907 |
| A + H + V + D | 593K | **32.16/0.895** | **28.57/0.781** | **27.56/0.735** | **26.03/0.784** | **30.44/0.908** |

## 5 Conclusion

In this article, a lightweight network called WTSR was proposed to extend the application scenarios of super-resolution algorithm. WTSR used SWT to enhance the representation of feature map, which was beneficial to the convergence and lightweight of the network. In the WTSR, a LTB was employed to capture and model the long-term dependency between similar image patches on the stationary wavelet domain. Then a set of WCEBs were utilized to recover the redundancy of the channels and the structure information of the wavelet coefficients. In the LTB, two ET encoders were investigated to mine local self-similarity while reducing the consumption of computing and memory resources. Meanwhile, a SMB was designed to combine global self-similarity and local self-similarity. In the WCEB, different ABs were designed according to the structural characteristics of different wavelet coefficients. Extensive experimental results in the public benchmarks showed that our WTSR achieved better trade-off between model performance and computation cost. However, WTSR still has some potential limitations, such as poor scalability of the network structure, difficulty in achieving better super-resolution reconstruction performance and higher super-resolution magnification by simple structure stacking. In the future, we will extend the proposed WTSR to specific mobile devices.

## References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: dataset and study. In: CVPR Workshops, pp. 126–135 (2017)
2. Ahn, N., Kang, B., Sohn, K.-A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 256–272. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_16
3. Anbarjafari, G., Demirel, H.: Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image. ETRI J. **32**(3), 390–394 (2010)
4. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: BMVC, pp. 1–10. BMVA Press (2012)

5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
6. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Guo, T., Seyed Mousavi, H., Huu Vu, T., Monga, V.: Deep wavelet prediction for image super-resolution. In: CVPR Workshops, pp. 104–113 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
10. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: CVPR, pp. 13713–13722 (2021)
11. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR, pp. 5197–5206 (2015)
12. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: ACMM, pp. 2024–2032 (2019)
13. Ji, H., Fermüller, C.: Robust wavelet-based super-resolution reconstruction: theory and algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 649–660 (2008)
14. K. He, X. Zhang, S. Ren, and J. Sun,: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: ICCV, pp. 1026–1034. IEEE Computer Society (2015)
15. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR, pp. 1646–1654 (2016)
16. Kim, J., Li, G., Yun, I., Jung, C., Kim, J.: Edge and identity preserving network for face super-resolution. Neurocomputing **446**, 11–22 (2021)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kumar, N., Verma, R., Sethi, A.: Convolutional neural networks for wavelet domain super resolution. Pattern Recogn. Lett. **90**, 65–71 (2017)
19. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep Laplacian pyramid networks for fast and accurate super-resolution. In: CVPR, pp. 624–632 (2017)
20. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: CVPR, pp. 3867–3876 (2019)
21. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-CNN for image restoration. In: CVPR, pp. 773–782 (2018)
22. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
23. Lu, Z., Liu, H., Li, J., Zhang, L.: Efficient transformer for single image super-resolution. arXiv preprint arXiv:2108.11084 (2021)
24. Ma, T., et al.: Oriented object detection with transformer. arXiv preprint arXiv:2106.03146 (2021)
25. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV, vol. 2, pp. 416–423. IEEE (2001)
26. Matsui, Y., et al.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools Appl. **76**(20), 21811–21838 (2017)

27. Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T.S., Shi, H.: Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: CVPR, pp. 5690–5699 (2020)
28. Oktay, O., et al.: Multi-input cardiac image super-resolution using convolutional neural networks. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 246–254. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46726-9_29
29. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: CVPR, pp. 16519–16529 (2021)
30. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR, pp. 3147–3155 (2017)
31. Tai, Y., Yang, J., Liu, X., Xu, C.: MemNet: a persistent memory network for image restoration. In: ICCV, pp. 4539–4547 (2017)
32. Ward, C.M., Harguess, J., Crabb, B., Parameswaran, S.: Image quality assessment for determining efficacy and limitations of super-resolution convolutional neural network (SRCNN). In: Applications of Digital Image Processing XL, vol. 10396, p. 1039605. International Society for Optics and Photonics (2017)
33. Xin, J., Li, J., Jiang, X., Wang, N., Huang, H., Gao, X.: Wavelet-based dual recursive network for image super-resolution. IEEE Trans. Neural Netw. Learn. Syst. **33**(2), 707–720 (2022)
34. Xue, S., Qiu, W., Liu, F., Jin, X.: Wavelet-based residual attention network for image super-resolution. Neurocomputing **382**, 116–126 (2020)
35. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
36. Yıldırım, D., Güngör, O.: A novel image fusion method using Ikonos satellite images. J. Geodesy Geoinf. **1**(1), 75–83 (2012)
37. Yu, J., et al.: Wide activation for efficient and accurate image super-resolution. arXiv preprint arXiv:1808.08718 (2018)
38. Zhang, H., Jin, Z., Tan, X., Li, X.: Towards lighter and faster: learning wavelets progressively for image super-resolution. In: ACMM, pp. 2113–2121 (2020)
39. Zou, W.W., Yuen, P.C.: Very low resolution face recognition problem. IEEE Trans. Image Process. **21**(1), 327–340 (2011)