2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Effective Local Texture Estimation Using Wavelet Transforms for Arbitrary-Scale Super-Resolution

Baihong Qian[1,2,+], Yu Lu[1,2,+], Dian Ding[1,2,*], Yi-Chao Chen[1,2], Qiaoling Xiao[3], Guanghui Gao[4], Zhengguang Xiao[5], and Guangtao Xue[1,2(✉)]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Key Laboratory of Trusted Data Circulation and Governance, and Web3, Shanghai, China
cherry_qbh@sjtu.edu.cn
[3] Biren Technology, Shanghai, China
[4] Department of Oncology, Shanghai Pulmonary Hospital & Thoracic Cancer Institute, Tongji University School of Medicine, Shanghai, China
[5] Department of Radiology, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Abstract. Image super-resolution (SR) aims to reconstruct high-resolution images from low-resolution inputs, addressing challenges like sensor noise, optical distortions, and compression artifacts. Traditional SR methods often struggle with preserving fine details, particularly in regions with sharp transitions or complex textures. In this work, we propose a novel Local Wavelet Transformer (LWT) framework that leverages the Discrete Wavelet Transform (DWT) to capture both local textures and global structures, improving the accuracy of fine-grained detail restoration. By introducing a magnification factor decomposition strategy, our method enables super-resolution at arbitrary scaling levels, ensuring flexibility and precise detail preservation across different magnifications. We demonstrate the effectiveness of our approach through extensive experiments on multiple benchmark datasets, showing superior performance and achieving state-of-the-art results in high-resolution image reconstruction under diverse conditions. Our results highlight the potential of wavelet-based analysis for enhancing SR tasks, particularly in scenarios requiring fine detail recovery and sharp transitions.

Keywords: Single image super resolution, Discrete wavelet transformation, Local attention mechanism.

1 Introduction

There is a growing demand for clear, realistic, and high-quality images to enhance visual experiences. However, image formation and transmission processes are often hindered by imaging systems' limitations, leading to degraded image quality. Factors such as sensor noise, optical distortions, and compression artifacts collectively contribute to this degradation, posing significant challenges across various domains [17, 23]. Addressing these challenges has fueled extensive research into image super-resolution

(SR) techniques—methods designed to reconstruct high-resolution images from low-resolution inputs—which offer a promising solution to improve image quality and usability.

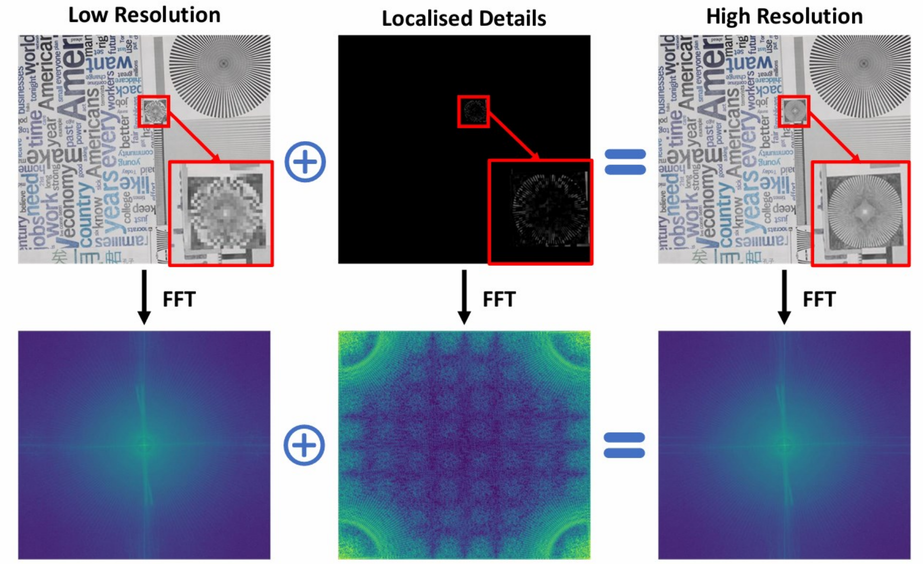

Fig. 1. Local texture details are represented as global features in the frequency domain by FFT

Among the various SR approaches, single image super-resolution (SISR) [7, 14, 22] has gained significant attention due to its potential to recover finer details from a single input image. One notable method in this field is LTE [11], which introduces an innovative approach by enhancing the expressiveness of local implicit functions. This is achieved by transforming spatial coordinates into the Fourier domain, which enables LTE to achieve precise, seamless image reconstruction across different scales, demonstrating the potential of frequency-domain techniques for super-resolution tasks.

While the Fourier Transform (FFT) [17] is a powerful tool for frequency-domain analysis, it operates on a global scale. Specifically, it represents the entire image in the frequency domain, where each frequency component spans the entire spatial domain. As illustrated in Fig. 1, the process requires a comprehensive alignment of features across all frequency bands, which can be computationally intensive and less efficient in capturing localized variations. To overcome these limitations, the Discrete Wavelet Transform (DWT) [15] emerges as a promising alternative. Unlike FFT, DWT inherently supports localized analysis by decomposing an image into multi-resolution sub-bands, as shown in Fig. 2. Each sub-band captures spatial and frequency information at different scales, enabling DWT to simultaneously preserve both global structure and localized details.

Building on the strengths of DWT, we propose a novel framework called the Local Wavelet Transformer (LWT). Unlike traditional approaches that operate directly in the spatial or Fourier domains, LWT projects image features into the wavelet domain before performing further processing and upscaling. By leveraging wavelet-based analysis, LWT enhances the network's ability to capture both localized details and broader structural features. This approach focuses on processing high-frequency textures in a way that supports more effective image reconstruction. As a result, LWT improves the representational capability of the network, leading to better performance in reconstructing high-resolution images, particularly in cases where fine details and sharp transitions are crucial.

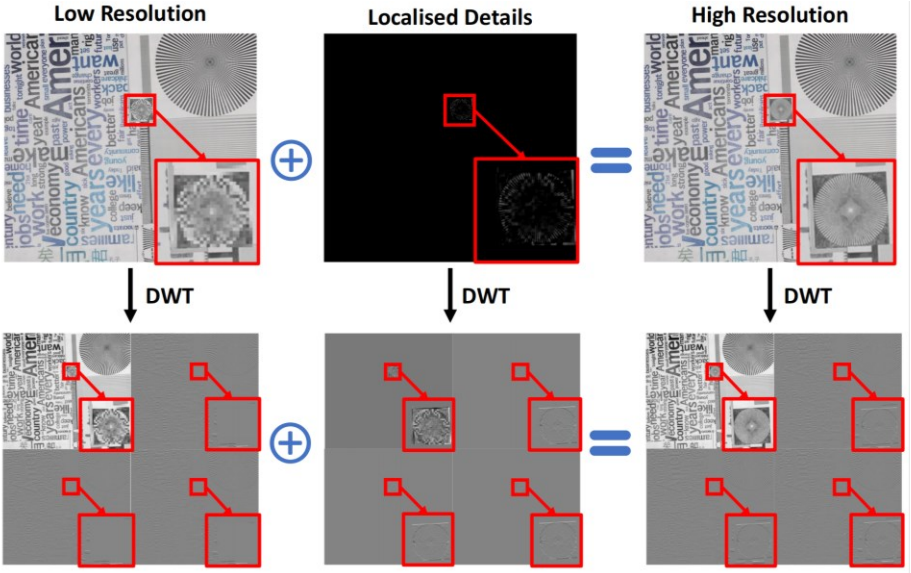

Fig. 2. Local texture details after DWT still behave as local features in the wavelet domain.

Building on the LWT upsampling framework, we adopt a strategy that decomposes the magnification factor, enabling super-resolution at arbitrary scaling levels. This approach not only provides flexibility in scaling but also ensures the preservation of fine-grained details across a range of magnification factors. To rigorously assess the performance and robustness of our method, we conduct extensive experiments on several benchmark datasets, including DIV2K, Set5, Set14, B100, and Urban100. The results demonstrate the effectiveness of our approach, with peak signal-to-noise ratio (PSNR) values reaching 35.26 dB on DIV2K, 38.43 dB on Set5, and 33.93 dB on Urban100 at ×2 magnification. Our method achieves high-quality image reconstruction across various scales, including ×2, ×3, ×4, and up to ×30, and consistently preserves fine details and textures under diverse imaging conditions.

The main contributions of our work are as follows:

- We propose a novel Local Wavelet Transformer (LWT) that leverages the Discrete Wavelet Transform (DWT) to capture local texture details, enhancing the network's ability to reconstruct fine-grained details in high-resolution images.
- We introduce a magnification factor decomposition strategy within the LWT framework, enabling super-resolution at arbitrary scaling levels while maintaining fine details across different magnifications.

- We perform extensive experiments on multiple benchmark datasets, demonstrating the effectiveness and robustness of our method in achieving high-quality image reconstruction across various scales and imaging conditions.

2 Related Work

2.1 Wavelet Transformation

The Wavelet Transform enables precise localization of image details and features, making it particularly effective for handling local characteristics such as edges and textures. [2] demonstrates that leveraging CNN representations across wavelet sub-bands can effectively enhance image restoration tasks. A multi-level wavelet transform [16] has also been employed to expand the receptive field while preserving critical information, benefiting image restoration processes. Additionally, Williams et al. [24] apply the Wavelet Transform to perform a second-level decomposition of input features, discarding first-level sub-bands to reduce feature dimensions and improve image recognition efficiency.

2.2 Single Image Super-Resolution and Arbitrary Scale Super-Resolution

Single Image Super-Resolution (SISR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) counterparts by reversing the degradation process, which typically involves blurring, down-sampling, and noise.

Early deep learning-based SISR methods, like SRCNN [5], used CNNs for feature extraction and HR reconstruction. Advances such as VDSR [9] incorporated residual learning, and LapSRN [11] used iterative up-sampling with supervised residuals for better detail preservation. More recent approaches, including diffusion models like SR3 [22], employ iterative refinement for improved image-to-image translation. Further refinements focus on efficiency through residual [13] or latent spaces [21], reducing computational costs and accelerating convergence. However, many methods still rely on fixed degradation models, limiting their real-world applicability.

A key challenge in SISR is the fixed scaling factor, requiring separate models for each upsampling scale. To address this, methods like Meta-SR [7] use meta-networks to enable arbitrary upsampling within a training range, though performance drops at larger scales. LIIF [4] addresses this by using a Multi-Layer Perceptron (MLP) to predict RGB values at arbitrary coordinates, enhancing generalization. UltraSR [25] further improves this by replacing coordinates with embedded ones to mitigate spectral bias, while LTE [12] refines the approach by using a Fourier domain transformation to capture high-frequency details, enabling accurate reconstruction across arbitrary scales.

3 Method

In this section, we introduce our super-resolution method, which consists of the Local Wavelet Transformer (LWT) upsampling framework and the magnification factor decomposition strategy for arbitrary scaling. A detailed description of these components is provided below.

3.1 Local Wavelet Transformer

The Local Wavelet Transformer (LWT) framework introduces a novel local attention mechanism operating in the wavelet domain. By leveraging wavelet-based analysis, LWT efficiently captures both localized high-frequency details and global low-frequency structures, ensuring precise and high-quality image reconstruction.

At its core, LWT adopts a high-level framework rooted in local implicit neural representation. This framework relies on a decoding function, f_θ, parameterized by a multi-layer perceptron (MLP) with trainable weights θ, which is shared across all images. The decoder maps latent tensors and spatial coordinates into RGB values, providing a continuous representation of the image. Mathematically, this mapping is defined as:

$$\mathrm{f}_{\theta(\mathrm{z},\,\mathrm{x})}\colon (\mathrm{Z},\mathrm{X})\to\mathrm{S}, \tag{1}$$

where $\mathrm{z} \in \mathrm{Z}$ is a latent tensor produced by an encoder E_φ, $\mathrm{x} \in \mathrm{X}$ is represents a 2D coordinate in the continuous image domain, and S denotes the space of predicted RGB values. For simplicity, the latent tensor $\mathrm{z} \in \mathrm{R}^{\{\mathrm{H}\times\mathrm{W}\times\mathrm{C}\}}$ is assumed to have the same spatial dimensions (height H and width W) as the low-resolution input image I_{LR}.

RGB value predictions $\mathrm{s} \in \mathrm{R}^3$ at a given coordinate $\mathrm{x} \in \mathbb{R}^2$ are calculated using:

$$s(x, I_{\mathrm{LR}};\, \Theta) = \sum_{\{\mathrm{j}\,\in\mathrm{J}\}\mathrm{w}_\mathrm{j}} \mathrm{f}_{\theta(\mathrm{z}_\mathrm{j},\mathrm{x}\,-\,\mathrm{x}_\mathrm{j})}, \tag{2}$$

where the latent tensor z is obtained as:

$$\mathrm{z} \;=\; \mathrm{E}_\phi(I_{\mathrm{LR}}),\;\; \Theta = [\theta;\phi]. \tag{3}$$

Similarly, our proposed method generates a high-resolution (HR) image $\mathrm{I}^{\mathrm{HR}} \in \mathrm{R}^{\mathrm{r_h H}\times\mathrm{r_w W}\times 3}$ from a low-resolution (LR) image $\mathrm{I}^{\mathrm{LR}} \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 3}$, where r ={$\mathrm{r_h}$, $\mathrm{r_w}$} represents the scaling factors. The workflow can be summarized as follows: an encoder E_ϕ first extracts feature embeddings $\mathrm{z} \in \mathrm{R}^{\{\mathrm{H}\times\mathrm{W}\times\mathrm{C}\}}$ from I^{LR}. These embeddings, along with the 2D coordinates of I^{LR}, are then passed into the LWT module to compute the RGB values of a residual HR image $\mathrm{I}^{\mathrm{HR}} \in \mathrm{R}^{\mathrm{r_h H}\times\mathrm{r_w W}\times 3}$ pixel by pixel.

To further enhance this process, we integrate the wavelet prediction network with residual networks. The residual networks align well with wavelet coefficients due to their sparsity-promoting nature and are further optimized to infer residuals. The resulting residual image $\mathrm{I}_\mathrm{r}^{\mathrm{HR}}$ is combined element-wise with an upsampled version of the LR image $\mathrm{I}_\uparrow^{\mathrm{HR}} \in \mathrm{R}^{\mathrm{r_h H}\times\mathrm{r_w W}\times 3}$ to produce the final HR image:

$$\mathrm{I}^{\mathrm{HR}} \;=\; \mathrm{I}_\mathrm{r}^{\mathrm{HR}} + \mathrm{I}_\uparrow^{\mathrm{HR}}. \tag{4}$$

This approach ensures that the network effectively utilizes both the residual information and the upsampled base image, enabling accurate reconstruction of fine-grained textures and high-frequency details.

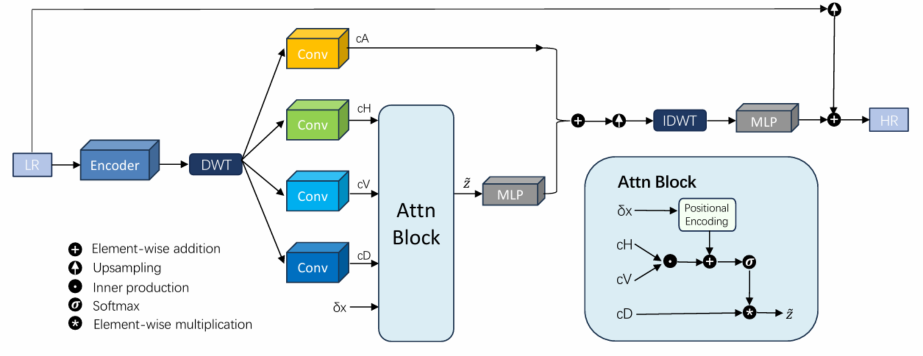

Fig. 3. Detailed framework of LWT.

As illustrated in Fig.3, the framework begins with an LR image, which is processed by an encoder to extract feature representations capturing its structural and textural details. LWT first projects the feature map into four sub-bands in wavelet domain through DWT: approximation coefficients cA, representing the low-frequency global structure, and detail coefficients cH, cV and cD, capturing high-frequency details in horizontal, vertical, and diagonal orientations, respectively. Using four separate convolutional layers, the four coefficients are processed into distinct latent embeddings, and the three detail coefficients cH, cV and cD correspond to the query q, key k, and value v in local attention block. Alongside this decomposition, local relative coordinates δx are generated to facilitate spatially aware processing. The detail coefficients and δx are then forwarded into a local attention block, which will be described in detail in the following subsection.

The output features z̃ are further projected into the same shape as cA using a multilayer perceptron (MLP). After concatenating four components, a convolutional layer is used to upsample them into certain scale, the upsampled features are then passed to an inverse DWT (IDWT) module, which reconstructs the residual HR image structure. Finally, the residual HR image is added to directly upsampled LR image (through bilinear upsampling) to construct a complete HR image.

3.2 Local Attention Block

The Local Attention Block begins by computing the inner product between q and k, incorporating the relative positional bias B to yield an intermediate attention score. The relative position coordinates is derived by first computing a displacement matrix, which represents the offset between neighboring pixels. This matrix is then added to the initial pixel coordinates, resulting in a new set of coordinates that account for both the absolute position and the relative offset.

The attention score is normalized through a Softmax function to generate a local attention map. The attention map is then applied to v via element-wise multiplication to produce z̃. The entire process can be described mathematically as:

$$\tilde{z}= \mathrm{softmax}\left(\frac{\mathrm{qk}^{\mathrm{T}}}{\sqrt{\mathrm{C}}}+\mathrm{B}\right)\times \mathrm{v}, \tag{5a}$$

$$\mathrm{B} = \mathrm{FC}\big(\gamma(\delta \mathrm{x})\big), \tag{5b}$$

$$\gamma(\delta \mathrm{x})= [\sin(2^{0}\delta \mathrm{x}), \cos(2^{0}\delta \mathrm{x}), \ldots ,\sin(2^{\mathrm{L}\text{-}1}\delta \mathrm{x}), \cos(2^{\mathrm{L}\text{-}1}\delta \mathrm{x})]\ , \tag{5c}$$

where C is the channel dimension of the key embedding k, FC denotes a fully-connected layer, $\gamma(\bullet)$ is the positional encoding function, and L is a hyperparameter set to 10. Additionally, the framework leverages a multi-head attention mechanism, which is defined as:

$$\tilde{z} = \mathrm{concat}\left(\mathrm{softmax}\left(\frac{\mathrm{q_i k_i^T}}{\sqrt{\left\{\frac{\mathrm{C}}{\mathrm{H}}\right\}}}+ \mathrm{B_i}\right)\times \mathrm{v_i}\right), \tag{6}$$

where H is the number of attention heads, and $i \in \{1, \ldots, H\}$.

3.3 Arbitrary-Scale Super-Resolution Strategy

In the proposed Local Wavelet Transformer (LWT) framework, we initially implement super-resolution upsampling for fixed magnifications of 2x and 3x. For arbitrary magnification factors A_s, we introduce a decomposition strategy to efficiently achieve the desired upsampling.

Given any target magnification factor $\mathrm{A_s}$, we seek a combination of powers of 2 and 3, denoted by $k = 2^{m} \bullet 3^{\mathrm{n}}$, that minimizes the absolute difference between k and $\mathrm{A_s}$. Specifically, we find the optimal values of m and n through an exhaustive search:

$$(\mathrm{m,n}) = \mathrm{argmin_{m,n}}|2^{\mathrm{m}} \bullet 3^{\mathrm{n}} - \mathrm{A_s}| \tag{7}$$

Once the closest $\mathrm{k} = 2^{m} \bullet 3^{\mathrm{n}}$ is identified, we apply m iterations of 2x upsampling and n iterations of 3x upsampling using the LWT framework. This decomposition ensures that the super-resolution process remains computationally efficient while preserving fine-grained details.

Finally, after performing the m and n upsampling steps, the final magnification of $\mathrm{A_s}$ is achieved through interpolation. This strategy allows for flexible and precise super-resolution scaling, enabling the LWT framework to handle arbitrary magnification factors with high accuracy and computational efficiency.

4 Experiments

In this section, we present the experimental results and discuss their implications.

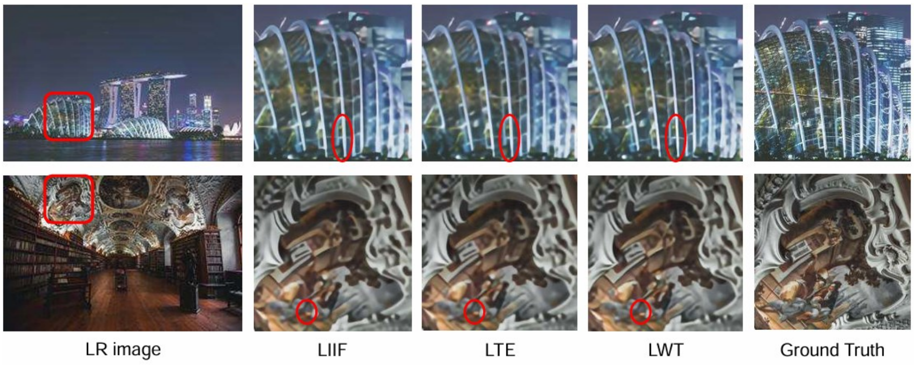

Fig. 4. Qualitative comparison to other arbitrary-scale SR. RDN [27] is used as an encoder for all methods.

Table 1. Quantitative comparison with state-of-the-art methods for arbitrary-scale SR on DIV2K dataset (PSNR (dB)).

Method	×2	×3	×4	×6	×12	×18	×24	×30
Bicubic	31.01	28.22	26.66	24.82	22.27	21.00	20.19	19.59
EDSR-baseline	34.55	30.90	28.94	-	-	-	-	-
EDSR-baseline-MetaSR [7]	34.64	30.93	28.92	26.61	23.55	22.03	21.06	20.37
EDSR-baseline-LIIF [4]	34.67	30.96	29.00	26.75	23.71	22.17	21.18	20.48
EDSR-baseline-LTE [12]	34.72	31.02	29.04	26.81	23.78	22.23	21.24	20.53
EDSR-baseline-LWT (ours)	34.78	31.00	29.04	26.89	23.90	22.41	21.04	20.61
RDN-MetaSR [7]	35.00	31.27	29.25	26.88	23.73	22.18	21.17	20.47
RDN-LIIF [4]	34.99	31.26	29.27	26.99	23.89	22.34	21.31	20.59
RDN-LTE [12]	35.04	31.32	29.33	27.04	23.95	22.40	21.36	20.64
RDN-LWT (ours)	35.09	31.34	29.36	27.06	23.95	22.41	21.05	20.62
SwinIR-MetaSR [7]	35.15	31.40	29.33	26.94	23.80	22.16	21.26	20.54
SwinIR-LIIF [4]	35.17	31.46	29.46	27.15	24.02	22.43	21.40	20.67
SwinIR-LTE [12]	35.24	31.50	29.51	27.20	24.09	22.50	21.47	20.73
SwinIR-LWT (ours)	**35.26**	**31.59**	**29.63**	**27.24**	**24.16**	**22.58**	**21.49**	**20.75**

4.1 Experimental Setup

Dataset We train our network using the DIV2K dataset [1], which was introduced in the NTIRE 2017 Challenge. This dataset contains 800 high-quality training images, covering a wide variety of scenes, such as natural landscapes, urban environments, and indoor settings, making it ideal for training generalizable super-resolution models. For evaluation, we report the performance of our model in terms of Peak Signal-to-Noise Ratio (PSNR) on several benchmark datasets: the DIV2K validation set [1], Set5 [3], Set14 [26], B100 [18], and Urban100 [8].

Implementation Details For training our network, we use 48 × 48 patches as low-resolution inputs and perform bicubic resizing to generate high-resolution images for down-sampling. The network is optimized using the Mean Squared Error (MSE) loss function, which is commonly used for image reconstruction tasks. The optimization process is carried out with the Adam optimizer [10], using a learning rate of 1e-4, $\beta_1 = 0.9$, $\beta_1 = 0.999$, and $\varepsilon = 1e-8$. These hyperparameters were chosen based on standard practices for image super-resolution tasks, balancing stability and convergence speed.

To demonstrate the flexibility of our approach, we train the Local Wavelet Transformer (LWT) with three different encoder architectures: the EDSR baseline [15], RDN [27], and SwinIR [14]. These encoders represent different types of architectures, ranging from traditional residual networks to more recent transformer-based models. Each encoder is integrated into the LWT framework to evaluate the impact of different feature extraction mechanisms on SR performance.

The model is implemented using PyTorch [20] and trained on an NVIDIA Tesla A800 GPU, providing sufficient computational power for training large-scale networks. We ensure that all experiments are performed with consistent hardware and software environments to ensure the reproducibility of results.

4.2 Quantitative and Qualitative Comparison

We evaluate the performance of our proposed Local Wavelet Transformer (LWT) on the DIV2K dataset, comparing it with state-of-the-art methods for arbitrary-scale SR. The results are reported in terms of Peak Signal-to-Noise Ratio (PSNR) at various scaling factors: ×2, ×3, ×4, ×6, ×12, ×18, ×24, and ×30.

As shown in Tab.1, our method consistently outperforms traditional bicubic interpolation and various other advanced methods across all scaling factors. At the ×2 scale, our SwinIR-LWT achieves the highest PSNR of 35.26 dB, surpassing the next best method, SwinIR-LTE, by 0.02 dB. This trend is maintained across the other scales, with our SwinIR-LWT model achieving the highest PSNR at every scaling factor, including 29.63 dB at ×4, 27.24 dB at ×6, and 24.16 dB at ×12.

Additionally, the performance of our method is further confirmed through visual comparisons. As shown in Fig.4 and Fig.5, our method excels in preserving fine edge details and textures, especially in regions with sharp transitions and intricate patterns. The results demonstrate that LWT effectively restores high-frequency components, leading to more accurate and visually appealing image reconstructions compared to other state-of-the-art techniques.

In addition to the experiments on the DIV2K dataset, we also evaluate our Local Wavelet Transformer (LWT) on several other benchmark datasets: Set5, Set14, B100, and Urban100. The results, presented in Fig. 5, further demonstrate the effectiveness of our method across different datasets and scaling factors.

For Set5, our SwinIR-LWT achieves the highest PSNR at all scaling factors, with notable improvements, especially at the ×4 scale, where we reach a PSNR of 32.95 dB, surpassing the second-best SwinIR-LTE by 0.14 dB. Similarly, on other datasets, our method consistently outperforms the others.

These results indicate that our method outperforms the state-of-the-art across multiple datasets, consistently achieving better PSNR values at various scaling factors, and further emphasizes the robustness and adaptability of the LWT framework in handling arbitrary-scale super-resolution tasks.

Table 2. Quantitative comparison with state-of-the-art methods for arbitrary-scale SR on benchmark datasets (PSNR).

Method	Set5					Set14					B100					Urban100				
	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8	×2	×3	×4	×6	×8
RDN [27]	38.24	34.71	32.47	-	-	34.01	30.57	28.81	-	-	32.34	29.26	27.72	-	-	32.89	28.80	26.61	-	-
RDN-MetaSR [7]	38.22	34.63	32.38	29.04	26.96	33.98	30.54	28.78	26.51	24.97	32.33	29.26	27.71	25.90	24.83	32.92	28.82	26.55	23.99	22.59
RDN-LIIF [4]	38.17	34.68	32.50	29.15	27.14	33.97	30.53	28.80	26.64	25.15	32.32	29.26	27.74	25.98	24.91	32.87	28.82	26.68	24.20	22.79
RDN-LTE [12]	38.23	34.72	32.61	29.32	27.26	34.09	30.58	28.88	26.71	25.16	32.36	29.30	27.77	26.01	24.95	33.04	28.97	26.81	24.28	22.88
RDN-LWT (ours)	38.23	34.78	32.90	29.43	27.24	34.36	30.62	28.88	26.77	25.43	32.48	29.33	27.79	26.09	25.09	33.16	28.97	26.80	24.68	23.21
SwinIR [14]	38.35	34.89	32.72	-	-	34.14	30.77	28.94	-	-	32.44	29.37	27.83	-	-	33.40	29.29	27.07	-	-
SwinIR-MetaSR [7]	38.26	34.77	32.47	29.09	27.02	34.14	30.66	28.85	26.58	25.09	32.39	29.31	27.75	25.94	24.87	33.29	29.12	26.76	24.16	22.75
SwinIR-LIIF [4]	38.28	34.87	32.73	29.46	**27.36**	34.14	30.75	28.98	26.82	25.34	32.39	29.34	27.74	26.07	25.01	33.36	29.33	27.15	24.59	23.14
SwinIR-LTE [12]	38.33	34.89	32.81	29.50	27.35	34.25	30.80	29.06	26.86	25.42	32.44	29.39	27.86	26.09	25.03	33.50	29.41	27.24	24.62	23.17
SwinIR-LWT (ours)	**38.43**	**34.92**	**32.95**	**29.63**	27.34	**34.68**	**30.83**	**29.16**	**26.87**	**25.75**	**32.60**	**29.39**	**27.87**	**26.14**	**25.12**	**33.93**	**29.42**	**27.28**	**24.80**	**23.30**

Table 3. Comparison of wavelet functions for super-resolution on different datasets.

Wavelet	DIV2K		Set5		Set14		B100		Urban100	
	×2	×4	×2	×4	×2	×4	×2	×4	×2	×4
Haar	35.26	29.63	38.43	32.90	34.68	28.88	32.48	27.79	33.93	27.28
Db2	34.80	29.62	37.86	31.90	33.46	28.37	32.10	27.49	31.71	25.74
Symlet	34.82	29.52	37.85	31.74	33.48	28.39	32.11	27.60	31.72	25.76
Morlet	35.10	29.86	37.96	31.98	34.11	28.54	32.25	27.87	32.92	26.54

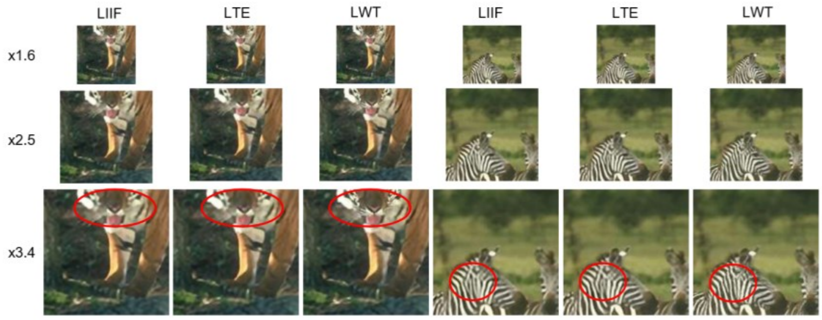

Fig. 5. Qualitative comparison to other arbitrary-scale SR. EDSR [7] is used as an encoder for all methods.

4.3 Impact of Wavelet Basis Functions

To further investigate the impact of wavelet functions on the performance of the Local Wavelet Transformer (LWT), we compare several commonly used wavelet bases, including Haar, Daubechies (Db2), Symlet, and Morlet. Tab. 3 shows the quantitative results of these wavelet functions on various benchmark datasets, including DIV2K, Set5, Set14, B100, and Urban100. Our findings indicate that the Haar wavelet performs the best across most datasets, achieving the highest PSNR values for both ×2 and ×4 upscaling factors.

5 Conclusion

We proposed the LWT, a novel framework that uses the DWT to capture both local textures and global structures for improved image super-resolution. By introducing a magnification factor decomposition strategy, our method enables super-resolution at arbitrary scales while preserving fine details. Extensive experiments on benchmark datasets demonstrate its effectiveness and robustness across different magnifications and conditions. Future work will focus on optimizing LWT for real-time applications and exploring its integration with other vision tasks.

Acknowledgements. This work is supported in part by National Natural Science Foundation of China (No. 61936015), Natural Science Foundation of Shanghai (No. 24ZR1430600) and Shanghai Key Laboratory of Trusted Data Circulation and Governance, and Web3.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW. pp. 126–135 (2017)
2. Bae, W., Yoo, J., Chul Ye, J.: Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In: CVPRW. pp. 145–153(2017)
3. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012)
4. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: CVPR. pp. 8628–8638 (2021)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence 38(2), 295–307 (2015)
6. Guo, T., Seyed Mousavi, H., Huu Vu, T., Monga, V.: Deep wavelet prediction for image super-resolution. In: CVPRW. pp. 104–113 (2017)
7. Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., Sun, J.: Meta-sr: A magnification arbitrary network for super-resolution. In: CVPR. pp. 1575–1584 (2019)
8. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR. pp. 5197–5206 (2015)
9. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR. pp. 1646–1654 (2016)
10. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE transactions on pattern analysis and machine intelligence 41(11), 2599–2613 (2018)
12. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function. In: CVPR. pp. 1929–1938 (2022)
13. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing 479, 47–59 (2022)

14. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCV. pp. 1833–1844 (2021)
15. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW. pp. 136–144 (2017)
16. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-cnn for image restoration. In: CVPRW. pp. 773–782 (2018)
17. Lu, Y., Ding, D., Pan, H., Fu, Y., Zhang, L., Tan, F., Wang, R., Chen, Y.C., Xue, G., Ren, J.: M3cam: Extreme super-resolution via multi-modal optical flow for mobile cameras. In: Sensys. pp. 744–756 (2024)
18. 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)
19. Nussbaumer, H.J., Nussbaumer, H.J.: The fast Fourier transform. Springer (1982)
20. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32 (2019)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
22. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence 45(4), 4713–4726 (2022)
23. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCV. pp. 1905–1914 (2021)
24. Williams, T., Li, R.: Wavelet pooling for convolutional neural networks. In: ICLR (2018)
25. Xu, X., Wang, Z., Shi, H.: Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. arXiv preprint arXiv:2103.12716 (2021)
26. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7. pp. 711–730. Springer (2012)
27. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR. pp. 2472–2481 (2018)