Training Transformer Models by Wavelet Losses Improves Quantitative and Visual Performance in Single Image Super-Resolution

Cansu Korkmaz, A. Murat Tekalp
College of Engineering and KUIS AI Center, Koc University
https://github.com/mandalinadagi/Wavelettention

Abstract

Transformer-based models have achieved remarkable results in low-level vision tasks including image super-resolution (SR). However, early Transformer-based approaches that rely on self-attention within non-overlapping windows encounter challenges in acquiring global information. To activate more input pixels globally, hybrid attention models have been proposed. Moreover, training by solely minimizing pixel-wise RGB losses, such as l_1, have been found inadequate for capturing essential high-frequency details. This paper presents two contributions: i) We introduce convolutional non-local sparse attention (NLSA) blocks to extend the hybrid transformer architecture in order to further enhance its receptive field. ii) We employ wavelet losses to train Transformer models to improve quantitative and subjective performance. While wavelet losses have been explored previously, showing their power in training Transformer-based SR models is novel. Our experimental results demonstrate that the proposed model provides state-of-the-art PSNR results as well as superior visual performance across various benchmark datasets.

1. Introduction

Single-image super-resolution (SR) aims to recover high-frequency (HF) details that are missing in low-resolution (LR) images. Early deep-learning-based models used simple convolutional neural networks (CNN) [15, 25]. Subsequently, various methods introduced residual learning, such as Enhanced deep residual networks (EDSR) [32], and attention mechanisms, such as residual channel attention networks (RCAN) [60]. More advanced models integrated dense connections [44, 61], spatial and channel attention networks [13, 38, 39, 57, 60], yielding remarkable performance in terms of peak signal-to-noise ratio (PSNR) and structural similarity measure (SSIM). However, CNNs encounter challenges in modeling long-range dependencies, prompting the development of Transformer-based image SR networks [7, 8, 31, 39, 58]. For instance, SwinIR [31], based on the Swin Transformer [34], significantly enhanced SR performance. Additionally, a hybrid attention transformer (HAT) [7], which combines channel attention and window-based self-attention with an overlapping cross-attention module, achieved state-of-the-art results.

While Transformer-based SR models demonstrate impressive performance, there is still room for further improvement. For example, HAT [7] provides improved performance over SwinIR [31] by introducing an overlapping cross-attention module to activate more pixels capturing longer-range dependencies. Recent studies [51, 54] have shown that integrating early convolutional layers can enhance visual representation in Transformer-based models. Based on this observation, we propose extending the HAT [7] architecture by incorporating non-local attention (NLSA) blocks [38] to further expand the receptive field and enhance the quality of reconstructed SR images.

It is well-established that merely minimizing RGB domain pixel-wise l_1 or l_2 losses is inadequate for capturing high-frequency details essential for achieving visually pleasing results [20, 27, 47, 64]. Since high-frequency image details are well represented by wavelet coefficients, we employ an additional wavelet loss term during the training to assist reconstruction of high-frequency details and improve both PSNR and visual quality of resulting SR images.

In summary, our main contributions are:

- We propose a new hybrid Transformer-based architecture for image SR that integrates convolutional non-local self-attention (NLSA) blocks with a Transformer-based model aimed at expanding the receptive field of the model.
- We introduce a wavelet loss term for training that enables SR models to better capture high-frequency image details.
- Training the proposed model by the wavelet loss not only improves the PSNR but also the visual quality of images. In particular, we obtain up to 0.72 dB PSNR gain over the state-of-the-art HAT model on the Urban100 test set.
- The proposed framework is generic in the sense that any Transformer-based SR network can be plugged into this framework and trained by wavelet losses for better results.

2. Related Work

2.1. Convolutional Attention Models

Several popular SR models employ convolutional attention mechanisms [13, 38, 39, 57, 59, 60]. One of the pioneer works, deep residual channel attention networks (RCAN) [60] proposed high-frequency channel-wise feature attention using a residual-in-residual (RIR) structure to form a deep CNN architecture. Second-order attention network (SAN) [13] proposed a trainable second-order channel attention module for enhanced feature learning and adaptive channel-wise feature rescaling based on second-order statistics. SAN also integrated non-local operations for long-distance spatial context and features for learning progressively abstract feature representations. Holistic attention network (HAN) [39] proposed capturing the correlation among different convolution layers with layer attention and channel-spatial attention modules. The Non-Local Sparse Attention (NLSA) [38] method introduced a dynamic sparse attention pattern, combining long-range modeling from non-local operations with robustness and efficiency of sparse representation. Recently proposed Efficient Long-Range Attention Network (ELAN) [59] incorporated a shift convolution (shift-conv) method and group-wise multi-scale self-attention module to preserve local structural information and long-range image dependencies. In this work, we incorporated NLSA blocks into a window-based transformer SR model for enlarged receptive field.

2.2. Transformer-Based SR Models

Transformer models, which are extremely successful in natural language processing [45], have also been adopted by the computer vision community for high-level vision tasks [11, 12, 16, 29, 33, 46, 49, 50], as well as low-level vision tasks including image SR [4–8, 28, 30, 31, 43, 48, 58].

Among the popular transformer-based image SR models, IPT [5] introduced a vision Transformer style network with multi-task pre-training for image SR. SwinIR [31] proposed an image restoration method based on the Swin Transformer [34]. UFormer [48] employed a locally-enhanced Transformer block to reduce computational complexity and introduced a learnable multi-scale restoration modulator for capturing better both local and global dependencies for image restoration. CAT [8] developed rectangle-window self-attention and a locality complementary module to enhance image restoration; later, CRAFT [28] proposed the high-frequency enhancement residual block along with a fusion strategy for Transformer-based SR methods to improve performance. ART [58] enlarged the receptive field using an attention retractable module for improved SR performance. SRFormer [63] suggested permuted self-attention that effectively balances channel and spatial information, enhancing the performance of self-attention mechanisms. Dual Attention Transformer (DAT) [9] combined spatial and channel feature aggregation both between and within Transformer blocks, enabling global context capture and inter-block feature aggregation. More recently, HAT [7] combined channel attention and window-based self-attention for improved global and local feature utilization, achieving state-of-the-art results in image SR tasks. However, existing Transformer-based models still do not fully harness global image correlations and there is room for further improvements [65]. To this effect, we propose to leverage more input pixels by cascading convolutional attention models with window-based transformer SR models aiming to enhance image reconstruction quality.

2.3. Frequency/Wavelet-Domain Models/Losses

Several researchers have recognized the need to treat low and high frequency image features differently and delved into frequency/wavelet domain methodologies to tackle image restoration/SR tasks [17, 40, 42, 52, 56]. Some of them explored the decomposition of features into frequency bands using multi-branch CNN architectures in the Fourier domain [2]. Frequency aggregation networks [40] extracted various frequencies from LR images and feed them into a channel attention-grouped residual dense network to recover HR images with enhanced details and textures. Other approaches include use of Fourier-domain architectures [18, 35, 56] and Fourier-domain loss functions [24]. For instance, SwinFIR [56] extended SwinIR by incorporating image-wide receptive fields using fast Fourier convolution, while DualFormer [35] leveraged spatial and spectral discriminators simultaneously in the Fourier domain. However, approaches that rely on Fourier domain methods and losses lack the ability to localize and capture the scale/orientation of high-frequency image features, thereby show limited improvements in SR performance.

Alternatively, wavelet-domain SR methods, such as DWSR [19], Wavelet- SRNet [21], WDST [14], WIDN [41], wavelet-based dual recursive network [52], WRAN [53] and PDASR [62], predict wavelet coefficients of SR images. In particular, PDASR [62] modified the model architecture to condition reconstructed images on low-level wavelet subbands to reduce visible artifacts and improve performance. In contrast, most recently, WGSR [27] proposed training RGB-domain standard GAN-SR models using a weighted combination of wavelet subband losses, departing from conventional RGB l_1 loss to control artifacts. In this paper, we show that training hybrid transformer SR models by wavelet losses also improves their performance. Our results are superior to others because predicting RGB pixels is easier than predicting sparse wavelet coefficients of detail bands, while unequal weighting of losses in different wavelet subbands enables learning structures with different scales and orientations.

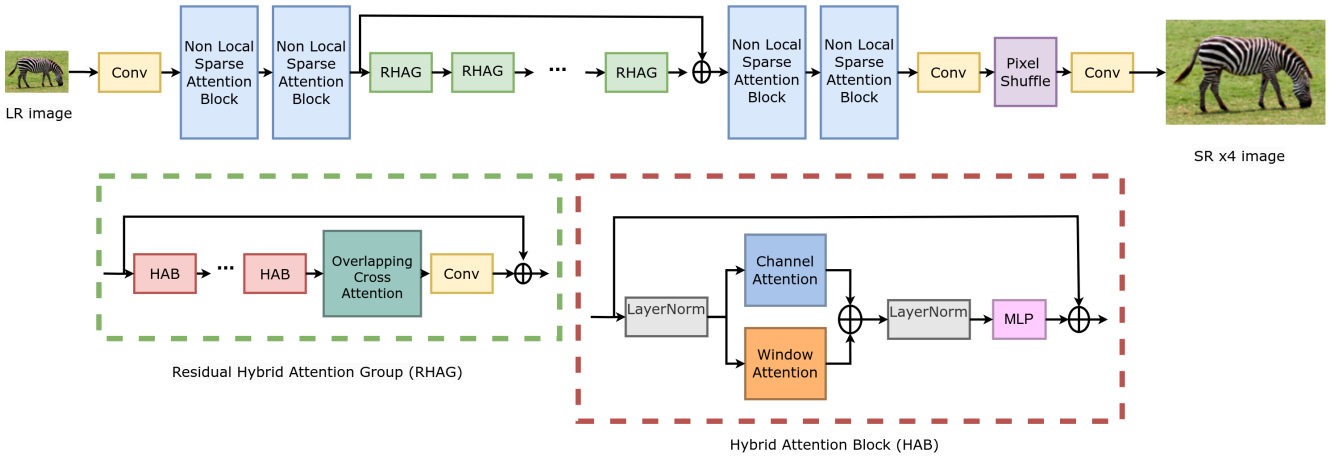

Figure 1. The proposed image SR architecture, which sandwiches HAT [7] in between NLSA blocks [38] for enlarged receptive field.

3. Proposed Method

3.1. Hybrid Attention Architecture

The proposed approach enhances the HAT architecture [7] by sandwiching it in between non-local sparse attention (NLSA) blocks [38] as depicted in Figure 1. The baseline HAT architecture utilized residual-in-residual approach [47] and developed hybrid attention block (HAB) similar to the standard Swin Transformer block [31] to enhance performance. In addition, the NLSA module [38] combines several techniques to enhance efficiency and global modeling in attention mechanisms. Specifically, NLSA uses Spherical Locality Sensitive Hashing method to divide input features into buckets to calculate attention. This module incorporates the strengths of Non-Local Attention, which allows for global modeling and capturing long-range dependencies in the image. Additionally, it leverages advantages of sparsity and hashing [38], which lead to high computational efficiency. Therefore, by sandwiching the HAT [7] architecture in between 2 or 4 NLSA blocks, we aim to increase the effectiveness of the attention mechanism, making the framework suitable for various applications in deep learning models especially for SR.

3.2. Training Loss Function

The wavelet loss is proposed to capture high-frequency details essential for visually pleasing SR results. This modified hybrid Transformer-based SR architecture is trained using wavelet losses along with the conventional l_1 RGB loss. The Stationary Wavelet Transform (SWT) is a technique that enables the multi-scale decomposition of images [23]. The illustration of SWT decomposition is depicted in Figure 2, which results in one low-frequency (LF) subband, called LL, and multiple high-frequency (HF) subbands, called LH, HL, and HH. The number of HF subbands is determined by the number of decomposition levels, and

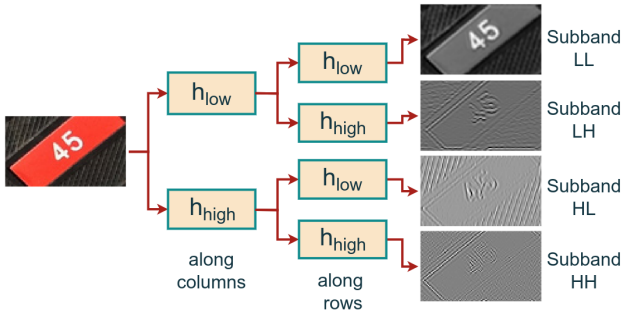

Figure 2. Illustration of the Stationary Wavelet Transform (SWT). SWT uses low-pass and high-pass decomposition filter pairs to compute the wavelet coefficients without subsampling subbands.

each HF subband contains detailed information in one of horizontal, vertical, or diagonal directions. SWT inherently combines scale/frequency information with spatial location, making it particularly suitable for tasks where preserving spatial details across different scales is essential, such as in SR applications. This combination of SWT loss with the proposed Transformer model aims to improve the overall performance of the baseline HAT model [7] by incorporating non-local attention mechanisms and leveraging wavelet losses to enhance image quality during training. To the best of our knowledge, this is the first work to use a wavelet loss function to train Transformer-based image SR models.

SWT is known for its efficiency and intuitive ability to represent and store multi-resolution images effectively [23, 36]. This means that it can capture both contextual and textural information of an image across different levels of detail. The understanding of how WT operates and its capacity to handle varying levels of image details inspired us to integrate wavelet losses into a Transformer-based super-resolution system. In other words, wavelet subbands have capability of handling different aspects of information en-

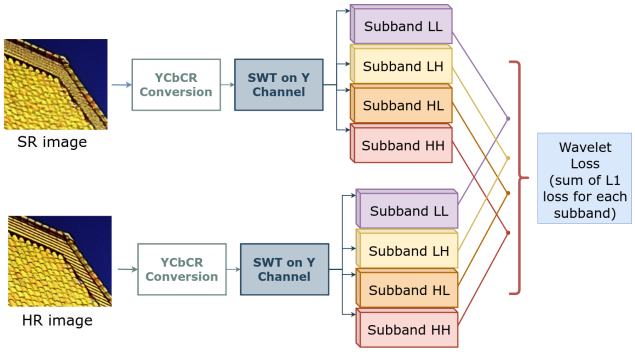

Figure 3. Computation of pixel-wise l_1 loss on SWT subbands. Training hybrid transformer architectures by a weighted combination of RGB and SWT losses results in remarkable quantitative and qualitative performance improvements.

coded in images that enables it a promising addition to enhance the performance of generated SR images. Hence, rather than employing the typical RGB-domain fidelity loss seen in conventional Transformer methods, we introduce the SWT fidelity loss denoted as L_{SWT} along with corresponding tuning parameter. The l_1 fidelity loss in wavelet-domain is calculated between the SWT subbands of the generated images x and the ground truth (GT) image y, is illustrated in Figure 3. The total wavelet loss is averaged over a minibatch size represented by $\mathbb{E}[.]$. This formulation allows for a more nuanced optimization process that can enhance the overall quality of the super-resolved images produced by the image Transformer models.

$$L_{SWT} = \mathbb{E}[\sum_j \lambda_j \left\| SWT(G(x))_j - SWT(y)_j \right\|_1] \quad (1)$$

where G denotes the proposed Transformer-based SR model and λ_j are appropriate scaling factors to control the generated HF details.
The overall loss for the training is given by

$$L_G = L_{RGB} + L_{SWT} \quad (2)$$

where L_{RGB} denotes the l_1 fidelity loss calculated on RGB domain, measuring pixel-wise errors in the image space.

The wavelet l_1 loss landscape differs from Y channel l_1 loss if i) we use a non-orthogonal wavelet transform, and/or ii) we use different weights for different subbands. It is worth noting that failing to adjust the balance between the wavelet subband loss terms and the RGB fidelity loss term (using scaling parameters λ_j) may result in chroma artifacts (producing greenish images), as the wavelet loss is computed based on the Y-channel only and favors Y channel. An alternative approach to address this potential issue would be incorporating an additional chroma loss term to explicitly preserve the color balance.

4. Experimental Results

4.1. Implementation Details

We configured the chunk size for the non-local sparse attention as 144 and added 2 consecutive NLSA [38] layers before and after the HAT [7] architecture. We utilized pre-trained HAT-L with default configuration. Hence the embedding dimension is set to 180 and patch embedding is set to 4. The total parameter number of proposed method is 41.3M. During training, we randomly crop 64x64 patches from the LR images from LSDIR [30] and DIV2K [1] datasets to form a mini-batch of 8 images. The training images are further augmented via horizontal flipping and random rotation of 90, 180, and 270 degrees. We optimize the model by ADAM optimizer [26] with default parameters. The learning rate is set to $4e^5$ and reduced by half after 125k, 200k and 240k iterations. The final model is obtained after 250k iterations. While calculating the pixel-wise loss of wavelet subbands, we utilize a Symlet filter "sym19" to compute wavelet coefficients. In the computation of the wavelet loss (Eqns. 1 and 2), we set all $\lambda_j = 0.05$ in order to avoid chroma artifacts. Our model is implemented with PyTorch and trained on Nvdia A40 GPUs.

4.2. Quantitative Results

To evaluate the generalization capability of our approach, we present results on validation benchmarks such as Set5 [3], Set14 [55], BSD100 [37], and Urban100 [22]. We employ fidelity metrics like PSNR and SSIM to gauge the performance of our proposed method alongside various state-of-the-art image SR techniques, including traditional attention models like EDSR [32], RCAN [60], SAN [13], HAN [39], NLSA [38], and ELAN [59]. Furthermore, we conduct comparisons with the state-of-the-art Transformer-based SR methods such as SwinIR [31], CAT [8], CRAFT [28], ART [58], SRFormer [63], DAT [9], and HAT [7]. Table 1 demonstrates the PSNR and SSIM performances of those methods for $\times 4$ image SR task on benchmark datasets. As depicted in Table 1, our method, Wavelettention, exhibits superior performance across all four benchmarks. In contrast to existing state-of-the-art Transformer-based methods such as SwinIR [31], CAT [8], ART [58], DAT [9] and HAT [7], our method achieves notable performance enhancements for $\times 4$ SR. Particularly, our method exhibits a PSNR gain of 0.3 dB on Set14[55], 0.12 dB on BSD100 [37], and 0.72 dB on Urban100 [22] compared to the competitive HAT [7] method.

To summarize, the observed PSNR improvements stem from enlarged receptive field of the proposed hybrid transformer model, our training strategy including wavelet losses, and utilization of larger datasets for training. These findings affirm that our Wavelettention stands as a robust hybrid transformer-based image SR model.

Table 1. Quantitative comparison of the proposed wavelet decomposition-based optimization objective vs. other state-of-the-art methods for $\times 4$ SR task. The best and the second-best are marked in **bold** and underlined, respectively.

Benchmark	Set5		Set14		BSD100		Urban100	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033
RCAN	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087
SAN	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068
HAN	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094
NLSA	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109
ELAN	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167
SwinIR	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254
CAT-R	32.89	0.9044	29.13	0.7955	27.95	0.7500	27.62	0.8292
CAT-A	33.08	0.9052	29.18	0.7960	27.99	0.7510	27.89	0.8339
CRAFT	32.52	0.8989	28.85	0.7872	27.72	0.7418	26.56	0.7995
ART	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321
SRFormer	32.93	0.9041	29.08	0.7953	27.94	0.7502	27.68	0.8311
SRFormer+	33.09	0.9053	29.19	0.7965	28.00	0.7511	27.85	0.8338
DAT	33.08	0.9055	29.23	0.7973	28.00	0.7515	27.87	0.8343
DAT+	33.15	0.9062	29.29	0.7983	28.03	0.7518	27.99	0.8365
HAT	33.04	0.9056	29.22	0.7973	28.00	0.7517	27.97	0.8368
Ours	**33.27**	**0.9082**	**29.53**	**0.8020**	**28.12**	**0.7549**	**28.69**	**0.8506**

4.3. Qualitative Results

We present challenging examples for visual comparison across three benchmark datasets in Figure 5. When compared with prominent Transformer-based methods such as SwinIR [31], CAT [8], ART [58], DAT [9] and HAT [7], our Wavelettention demonstrates superior restoration of detailed edges and textures. Specifically, our method exhibits a stronger ability to restore blurred text characters in Set14 [55] and BSD100 [37]. Additionally, Wavelettention successfully restores bricks of the architectures in the image patch from BSD100 [37] which SwinIR [31] struggles with. Furthermore, our method manages to reconstruct the parallel stripes with small intervals in the Urban100 dataset [22], however the other Transformer-based methods including SwinIR [31], CAT [8], ART [58], DAT [9] and HAT [7] focus only on simpler textures due to their limited receptive field, resulting in undesirable visual outcomes.

To summarize, by sandwiching the HAT architecture in between NLSA blocks and introducing the wavelet loss, our Wavelettention model excels in the image SR task, showcasing outstanding visual performance.

4.4. Power of Training by Wavelet Losses

This subsection demonstrates that training by a weighted combination of RGB and SWT losses contributes to improved reconstruction quality with other Transformer-based SR models as well. We take SwinIR [31] as a case study in Figure 4. Figure 4 shows SwinIR trained by l_1 loss only exhibits artifacts and erroneous reconstructions, notably in hallucinating roof bricks as parallel lines. However, incorporating the proposed wavelet losses in the training without altering any other configuration or additional training data, SwinIR+SWT results in more accurate reconstructions, particularly in preserving structural details akin to the ground-truth HR image. While training the SwinIR+SWT, we scale LL and HH subbands with 0.05, whereas LH and HL subbands are multiplied by 0.01. We observe that SwinIR+SWT showcases improved capabilities in overcoming aliasing artifacts and recovering correct line orientations in both images.

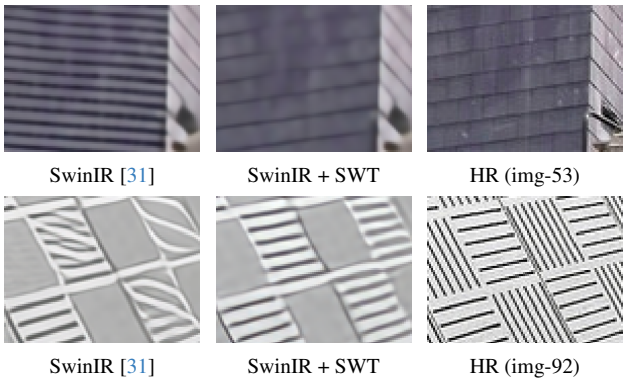

Figure 4. Visual comparison of SwinIR trained by l_1 loss [31] vs. trained by SWT losses on images 53 & 92 from Urban100 dataset [22]. Observe that training SwinIR by l_1 loss results in hallucinated edge directions, whereas SwinIR trained by weighted l_1 and SWT losses (SwinIR+SWT) recovers all structures correctly.

To summarize, training by wavelet losses improves the performance of the SwinIR model [31] by mitigating artifacts without any bells and whistles.

Figure 5. Visual comparison of $\times 4$ SR results on images selected from Set14 [55], BSD100 [37] and Urban100 [22] benchmarks. Observe that SwinIR trained by l_1 loss only (second column) generates aliasing artifacts, while CAT shows extreme blurring on some patches of the image in the last row. Other models show moderate blurring, while our models shows the best visual results on all images.

Table 2. Ablation studies to demonstrate the contributions of adding NLSA blocks and using wavelet loss in $\times 4$ SR results on benchmarks.

	Baseline	Baseline	+NLSA blocks	+L_{SWT}	+NLSA+ L_{SWT}
Training Set	ImageNet +DF2K	ImageNet +DF2K +LSDIR	ImageNet +DF2K +LSDIR	ImageNet +DF2K +LSDIR	ImageNet +DF2K +LSDIR
Benchmark	PSNR - SSIM	PSNR - SSIM	PSNR - SSIM	PSNR - SSIM	PSNR - SSIM
Set5	33.039 - 0.9054	33.143 - 0.9071	33.204 - 0.8893	33.268 - 0.9084	33.273 - 0.9082
Set14	29.246 - 0.7974	29.387 - 0.8007	29.524 - 0.8071	29.527 - 0.8020	29.524 - 0.8020
BSD100	27.996 - 0.7515	28.091 - 0.7547	27.795 - 0.7422	28.118 - 0.7548	28.119 - 0.7549
Urban100	27.954 - 0.8366	28.247 - 0.8299	28.548 - 0.8434	28.684 - 0.8504	28.692 - 0.8506
DIV2K	31.219 - 0.8547	31.393 - 0.8575	31.363 - 0.8482	31.426 - 0.8577	31.430 - 0.8578
LSDIR	26.995 - 0.7883	27.247 - 0.7955	26.879 - 0.7780	27.292 - 0.7956	27.296 - 0.7958

Table 3. Analysis of the effect of the number of SWT-decomposition levels used to compute the wavelet loss term on the model performance over the BSD100 [37] dataset.

Benchmark	BSD100	Urban100
Method	PSNR - SSIM	PSNR - SSIM
Baseline	28.088 - 0.756	28.589 - 0.849
1-level SWT	28.119 - 0.755	28.697 - 0.851
2-level SWT	27.514 - 0.742	28.699 - 0.851

4.5. The Effect of Wavelet-Decomposition Levels

The selection of the number of levels in the SWT decomposition plays a crucial role and directly affects the overall performance of the SR outputs. This decision is influenced by various factors, including the scale and orientation of structures within LR images. Our investigation focuses on assessing the impact of the decomposition level within our auxiliary wavelet-loss term. Specifically, we experiment with a 2-level SWT loss by further decomposing the LL subband of the 1-level SWT into 4 subbands (L-LL, L-LH, L-HL, L-HH), resulting in a total loss term that involves 8 distinct pixel-wise loss calculations (1 for RGB pixel-loss and 7 for wavelet subbands). Our findings, detailed in Table 3, indicate that while the utilization of a 2-level wavelet loss does not significantly improve network generalization performance on the BSD100 [37] benchmark, it does lead to performance improvements on the Urban100 [22] dataset.

To summarize, the effect of the number of SWT decomposition levels varies depending on the properties of images, such as the scale/frequency and orientation of structures, in the training/test datasets.

4.6. Ablation Study

In this subsection, we conduct a series of ablation study experiments to analyze the individual impact of fine-tuning the HAT-L architecture with each of the following steps: i) using the large-scale LSDIR dataset [30], ii) inclusion of NLSA blocks, and iii) our training strategy with wavelet losses. The evaluation is performed on six benchmark datasets, including Set5 [3], Set14 [55], BSD100 [37], Urban100 [22], DIV2K [1], and LSDIR [30], with results presented in Table 2.

We first fine-tune our baseline model HAT-L [7] with the recently published large-scale dataset LSDIR containing diverse natural images [30]. This fine-tuning step yields a gain of +0.1 dB across all benchmarks, prompting us to utilize LSDIR dataset for the other experiments, including sandwiching in between the NLSA blocks and training with wavelet losses. Hence, as a baseline model, we consider HAT-L architecture that is fine-tuned on LSDIR [30] dataset. Subsequently, we introduce +4 NLSA blocks [38] to the baseline architecture to augment the receptive field and the inclusion of NLSA blocks improves Set5 [3], Set14 [55] and Urban100 [22] datasets up to 0.3 dB. Then, we train baseline architecture with SWT loss only without including NLSA blocks which significantly enhances overall results across all benchmarks, with an improvement of approximately 0.2 dB. Additionally, by incorporating the NLSA block and introducing an auxiliary wavelet loss function to refine our Wavelettention model, we achieve a notable enhancement, demonstrating a significant PSNR gain of 0.32 dB across almost every validation benchmark, surpassing the baseline performance.

4.7. NTIRE 2024 Single Image Super-Resolution (x4) Challenge Results

With the proposed Wavelettention model, we participated in the Image Super-Resolution $\times 4$ subtrack of the New Trends in Image Restoration and Enhancement (NTIRE) 2024 challenge [10]. This challenge aims to design SR methods with the best PSNR and SSIM performance. It comprises of two datasets: DIV2K [1] and LSDIR [30]. Specifically, the DIV2K training dataset contains 800 pairs of high-resolution (HR) and low-resolution (LR) images and the validation set consists of 100 LR-HR pairs. During the final phase of the challenge, DIV2K test dataset containing 100 diverse LR images has been released to generate SR

Table 4. NTIRE 2024 Image Super-Resolution Challenge Results. PSNR and SSIM scores for $\times 4$ SR on validation and test phases.

	PSNR	SSIM
Validation Dataset	31.43	0.86
Test Dataset	31.13	0.86

results. Our Wavelettention SR model actively participated in both the validation and testing phases of this challenge with the outcomes shown in Table 4.

4.8. Analysis of Model Complexity vs. Performance

We conduct experiments to analyze the computational complexity versus performance gain of additional NLSA blocks to the baseline HAT-L [7] architecture. As illustrated in Table 5, adding single NLSA block before and after the feature extraction layers of HAT-L [7] model increases model performance 0.14 dB on Set14 [55]. Furthermore, our Wavelettention model that contains +4 NLSA blocks obtains a performance gain by 0.2 dB with an increase of parameters and Multi-Adds. However, further addition of NLSA blocks does not lead to a performance gain even though the parameter size and Multi-Add increases. This problem may be addressed by providing additional training data.

To summarize, we observe that inserting 2 or 4 NLSA blocks both before and after the feature extraction module of HAT-L leads to an increase in the performance of the state-of-the-art HAT model with a small increase in complexity.

Table 5. Complexity vs. performance comparison for $\times 4$ SR. Each row shows the cost and benefit of adding NLSA blocks [38] to the baseline HAT-L [7] architecture on Set14 [55] benchmark.

Method	# Params.	# Multi-Adds.	PSNR
Baseline	40.8M	79.61G	29.349
+2 NLSA Blocks	41.1M	80.47G	29.485
+4 NLSA Blocks	41.3M	81.33G	29.524
+8 NLSA Blocks	41.7M	83.06G	29.496

5. Conclusion

This paper introduces a new hybrid transformer model for image SR by sandwiching a transformer architecture in between convolutional NLSA blocks aiming to further increase the receptive field of the model and enhance the quality of the generated SR images. Furthermore, we address the well-known limitation of RGB pixel-wise losses in capturing high-frequency details that are crucial for visual quality, by introducing wavelet subband losses. Specifically, we propose training SR models by a weighted combination of RGB and wavelet losses to preserve the scale and orientation of high-frequency image details for improved SR performance. Through extensive ablation studies, we show that both the proposed architectural improvement and use of wavelet losses in training help us achieve better quantitative (PSNR and SSIM) scores and qualitative (visual) results compared to state-of-the-art transformer-based methods for the SR task. We also demonstrated that training by wavelet losses can improve the performance of other transformer-based SR models, such as the SwinIR model.

6. Acknowledgements

This work is supported by TUBITAK 2247-A Award No. 120C156, TUBITAK 2232 International Fellowship for Outstanding Researchers Award No. 118C337, KUIS AI Center, and Turkish Academy of Sciences (TUBA).

References

[1] E Agustsson and R. Timofte. NTIRE 2017 Challenge on single image super-resolution: Dataset and study. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR) Workshops*, 2017. 4, 7

[2] Sangwook Baek and Chulhee Lee. Single image super-resolution using frequency-dependent convolutional neural networks. In *IEEE Int. Conf. on Industrial Technology (ICIT)*, pages 692–695. IEEE, 2020. 2

[3] Marco Bevilacqua, Aline Roumy, C. Guillemot, and M. Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference*, pages 135.1–135.10, 2012. 4, 7

[4] Jiezhang Cao, Yawei Li, Kai Zhang, Jingyun Liang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2

[5] Hanting Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, Siwei Ma, C. Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 12299–12310, 2021. 2

[6] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023.

[7] Xiangyu Chen, X. Wang, J. Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 22367–22377, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[8] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NeurIPS*, 2022. 1, 2, 4, 5, 6

[9] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 12312–12321, 2023. 2, 4, 5, 6

[10] Zheng Chen, Zongwei Wu, Eduard-Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. Ntire 2024 challenge on image super-resolution (x4): Methods and results. In *Computer Vision and Pattern Recognition Workshops*, 2024. 7

[11] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS 2021*, 2021. 2

[12] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *ICLR 2023*, 2023. 2

[13] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 11057–11066, 2019. 1, 2, 4

[14] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 3076–3085, 2019. 2

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conf. Comp. Vision (ECCV)*, pages 184–199, 2014. 1

[16] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *IEEE/CVF Conf. on Comp. Vis. and Patt. Recog. (CVPR)*, pages 12124–12134, 2022. 2

[17] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog.*, pages 203–212, 2021. 2

[18] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 2340–2349, 2021. 2

[19] T. Guo, H. S. Mousavi, T. H. Vu, and V. Monga. Deep wavelet prediction for image super-resolution. In *IEEE/CVF Conf. Comp. Vis. and Patt. Recog. (CVPRW)*, pages 104–113, 2017. 2

[20] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1698–1706, 2017. 1

[21] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet-SRnet: A wavelet-based CNN for multi-scale face super resolution. In *IEEE Int. Conf. Comp. Vis. (ICCV)*, pages 1689–1697, 2017. 2

[22] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 5197–5206, 2015. 4, 5, 6, 7

[23] Bjorn Jawerth and Wim Sweldens. An overview of wavelet based multiresolution analyses. *SIAM Review*, 36(3):377–412, 1994. 3

[24] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Int. Conf. Comp. Vision (ICCV)*, 2021. 2

[25] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 1646–1654, 2016. 1

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[27] Cansu Korkmaz, A Murat Tekalp, and Zafer Dogan. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. *arXiv preprint arXiv:2402.19215*, 2024. 1, 2

[28] Ao Li, Le Zhang, Yun Liu, and Ce Zhu. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In *IEEE/CVF Int. Conf. on Comp. Vision*, pages 12514–12524, 2023. 2, 4

[29] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 2

[30] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog (CVPR.) Workshops*, pages 1775–1787, 2023. 2, 4, 7

[31] Jingyun Liang, Jie Cao, G. Sun, K. Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *IEEE/CVF Int. Conf. on Comp. Vision (ICCV) Workshops*, pages 1833–1844, 2021. 1, 2, 3, 4, 5, 6

[32] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF CVPR Workshops*, 2017. 1, 4

[33] Xiaoning Liu, Ao Li, Zongwei Wu, Yapeng Du, Le Zhang, Yulun Zhang, Radu Timofte, and Ce Zhu. Pasta: Towards flexible and efficient hdr imaging via progressively aggregated spatio-temporal aligment. *arXiv preprint arXiv:2403.10376*, 2024. 2

[34] 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 *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 9992–10002, 2021. 1, 2

[35] Xin Luo, Y Zhu, S. Xu, and D. Liu. On the effectiveness of spectral discriminators for perceptual quality improvement. In *IEEE/CVF Int. Conf. on Comp. Vision (ICCV)*, 2023. 2

[36] S. Mallat. Wavelets for a vision. *Proceedings of the IEEE*, 84(4):604–614, 1996. 3

[37] 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 *IEEE Int. Conf. on Computer Vision. (ICCV)*, pages 416–423 vol.2, 2001. 4, 5, 6, 7

[38] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3516–3525, 2021. 1, 2, 3, 4, 7, 8

[39] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conf. Comp Vis.ion (ECCV)*, page 191–207, 2020. 1, 2, 4

[40] Yingxue Pang, Xin Li, Xin Jin, Yaojun Wu, Jianzhao Liu, Sen Liu, and Zhibo Chen. Fan: Frequency aggregation network for real image super-resolution. In *Euro. Conf. Comp (ECCV) Workshops: Glasgow, UK*, pages 468–483, 2020. 2

[41] F. Sahito, P. Zhiwen, J. Ahmed, and R. A. Memon. Wavelet-integrated deep networks for single image super-resolution. *Electronics*, 8:553, 2019. 2

[42] Zehua Sheng, Xiongwei Liu, Si-Yuan Cao, Hui-Liang Shen, and Huaqi Zhang. Frequency-domain deep guided image denoising. *IEEE Trans. on Multimedia*, 2022. 2

[43] Chunwei Tian, M. Zheng, W. Zuo, S. Zhang, Y. Zhang, and Chia-Wen Lin. A cross transformer for image denoising. *Information Fusion*, 102:102043, 2024. 2

[44] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *IEEE Int. Conf. on Comp. Vision (ICCV)*, pages 4809–4817, 2017. 1

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[46] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 548–558, 2021. 2

[47] X. Wang, Ke Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *European Conf. on Comp. Vision (ECCV) Workshops*, 2018. 1, 3

[48] Zhendong Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. Uformer: A general u-shaped transformer for image restoration. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 17683–17693, 2022. 2

[49] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CVT: Introducing convolutions to vision transformers. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 22–31, 2021. 2

[50] Sitong Wu, Tianyi Wu, Haoru Tan, and Guodong Guo. Pale transformer: A general vision transformer backbone with pale-shaped attention. In *AAAI Conf. on Artificial Intelligence*, pages 2731–2739, 2022. 2

[51] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 1

[52] Jingwei Xin, Jie Li, Xinrui Jiang, Nannan Wang, Heng Huang, and Xinbo Gao. Wavelet-based dual recursive network for image super-resolution. *IEEE Trans. on Neural Networks and Learning Systems*, 33(2):707–720, 2020. 2

[53] S. Xue, W. Qiu, Fan Liu, and X. Jin. Wavelet-based residual attention network for image super-resolution. *Neurocomputing*, 382:116–126, 2020. 2

[54] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 579–588, 2021. 1

[55] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer. 4, 5, 6, 7, 8

[56] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022. 2

[57] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Haiyin Piao, Haiyang Mei, Xin Yang, and Baocai Yin. A two-stage attentive network for single image super-resolution. *IEEE Trans. on Circuits and Systems for Video Tech.*, 2021. 1, 2

[58] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *Int. Conf. Learning Representations (ICLR)*, 2023. 1, 2, 4, 5, 6

[59] Xindong Zhang, Hui Zeng, Shi Guo, and L. Zhang. Efficient long-range attention network for image super-resolution. In *Euro. Conf. Comp. Vision (ECCV)*, 2022. 2, 4

[60] Y. Zhang, K. Li, Kai Li, L. Wang, B. Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conf. Comp. Vision (ECCV)*, 2018. 1, 2, 4

[61] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE/CVF conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 2472–2481, 2018. 1

[62] Yuehan Zhang, Bo Ji, Jia Hao, and Angela Yao. Perception-distortion balanced admm optimization for single-image super-resolution. In *European Conf. on Comp. Vision (ECCV)*, 2022. 2

[63] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 12780–12791, 2023. 2, 4

[64] Qiuyu Zhu, Hu Wang, and R. Zhang. Wavelet loss function for auto-encoder. *IEEE Access*, 9:27101–27108, 2021. 1

[65] Qiang Zhu, Pengfei Li, and Qianhui Li. Attention retractable frequency fusion transformer for image super resolution. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 1756–1763, 2023. 2