Spatial-enhanced Multi-level Wavelet Patching in Vision Transformers

Fuzhi Wu, Jiasong Wu, Huazhong Shu, Guy Carrault, Lotfi Senhadji

Spatial-enhanced Multi-level Wavelet Patching in Vision Transformers

Fuzhi Wu, Jiasong Wu#, *Member, IEEE*, Huazhong Shu*, *Senior Member, IEEE*, Guy Carrault, Lotfi Senhadji, *Senior Member, IEEE*

***Abstract*—By seamlessly integrating wavelet transforms into the image patching stage of ViT, we leverage the power of multi-level wavelet transforms to decompose images into a diverse array of frequency-domain features. These features, integrated with spatial characteristics at equivalent scales, enrich image details, enhancing ViT's proficiency in delineating intricate textures and distinct edges. Consequently, we registered a notable 2.7% accuracy enhancement on the ImageNet100 dataset in ViT. Our wavelet patching module, designed for versatility, seamlessly fits into various ViT derivatives without necessitating architecture modifications. This advancement has uplifted the performance of several leading vision transformers by 0.46-4.3%, preserving parameter efficiency without notable FLOPs increment.**

***Index Terms*—Vision Transformer, Image patching, Wavelet transform, Low-level feature**

I. Introduction

THE Vision Transformer (ViT) [1] tokenizes images into fixed-size patches, employing Transformer layers akin to language models to determine inter-token relationships for image classification. However, this method often overlooks the vital local nuances [2], [3] within each patch, notably textures [4], edges [5], and lines, requiring larger training datasets to match CNN benchmarks [6]. In signal processing, techniques like the discrete wavelet transform (DWT) can distinguish such features across varied frequency bands and efficiently spotlight these obscured local features. Nevertheless, many ViT variants sidestep patch-processing enhancements.

Existing solutions aim to encode the local structure of tokens, with approaches including Tokens-to-Token [7] module and the Transformer-iN-Transformer (TNT) [8] architecture, as referenced in [9] and [10]; methods that exploit wavelet features through convolution layers, as demonstrated in [11]; and designs that implement parallel channels for high-frequency details, discussed in [12]. However, these approaches frequently entail intricate architectural changes, raising questions on the need for nuanced ViT designs similar to CNNs and the possibility of enhancing ViT without altering its fundamental structure.

Within Vision Transformers, wavelets mainly contribute in: [13] introducing the wavelets into *position embedding* enhances the smoothness in pathological feature maps; [14], [15], [16] improve *self-attention efficiency* via additionally adding wavelet features into multi-head attention; Parallel processing of different wavelet subbands in *texture recognition* [17], *denoising* [18], *super-resolution* [19] tasks leads to better results. Yet, the potential of wavelet-enriched domains to amplify token features remains underexplored, presenting an opportunity for deeper utilization.

Our work introduces an augmented spectrum approach in the patching phase to diversify ViT features. We dissect images across frequency bands by integrating a multi-level wavelet transformation, capturing both macro and micro nuances. This method addresses potential gaps in color and texture direction overlooked by wavelets, further enhanced by a spatial domain module. As shown in Fig. 1, the resulting encoding includes a diverse range of spectral features within patches, with the spatial domain treated as a distinct frequency band. Our wavelet patch module is designed to align

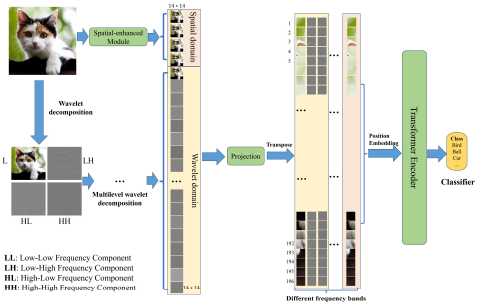

Fig. 1. The Spatial-enhanced Multi-level Wavelet Patch (SMWP) involves wavelet decomposition of an image into four frequency bands: LL, LH, HL, and HH, continued until the number of pixel points aligns with the token count required by the ViT. Concurrently, spatial features are derived from the original image via down-sampling in a spatial-enhanced module, then concatenated with wavelet frequencies as a unique band. This amalgamation of spatial and frequency data creates consolidated tokens for the transformer encoder.

*Corresponding authors: Huazhong Shu. (e-mail: shu.list@seu.edu.cn)

#Co-author. This work was supported in part by the National Key Research and Development Program of China (Nos. 2022YFE0116700, 2021ZD0113202), in part by the National Natural Science Foundation of China under Grants 62171125, 61876037, and in part by the innovation project of Jiangsu Province under grants BZ2023042, BY2022564.

F. Wu, J. Wu, and H. Shu are with LIST, Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing 210096, China.

F. Wu, G. Carrault and L. Senhadji are with Univ Rennes, Inserm, LTSI - UMR 1099, Rennes F-35000, France.

F. Wu, J. Wu, H. Shu, G. Carrault and L. Senhadji are with Centre de Recherche en Information Biomédicale Sino-français (CRIBs), Univ-Rennes, Inserm, Southeast University, Rennes F-35042, France, and with Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing 210096, China.

seamlessly with numerous ViT derivatives without requiring structural changes, offering three key advantages:
(1) Enhanced spectrum characterization, demonstrated by a 2.7% improvement in ImageNet100 accuracy with ViT-B/16.
(2) A versatile design compatible with several state-of-the-art ViT variants (T2T-ViT [7], TNT [8], Swin-ViT [22], Pyramid TNT [23], Wave-ViT [14], Dual-ViT [24]), positioning it as a potential better backbone network for Transformer-based vision networks.
(3) Reduced reliance on position embedding, enhancing ViT models' robustness.

II. METHOD

A. Wavelet Decomposition

1) 1D-DWT

For a 1D signal $\boldsymbol{s} = \{s_j\}_{j\in\mathbb{Z}}$, DWT decomposes it into its low-frequency component $\boldsymbol{s_1} = \{s_{1k}\}_{k\in\mathbb{Z}}$ and high-frequency component $\boldsymbol{d_1} = \{d_{1k}\}_{k\in\mathbb{Z}}$, where

$$\begin{cases} s_{1k} = \sum_j l_{j-2k} s_j, \\ d_{1k} = \sum_j h_{j-2k} s_j, \end{cases} \tag{1}$$

and $\boldsymbol{l} = \{l_j\}_{j\in\mathbb{Z}}, \boldsymbol{h} = \{h_k\}_{k\in\mathbb{Z}}$ are the low-pass and high-pass filters of a given wavelet basis. In expressions with matrices and vectors, Eq. (1) can be rewritten as

$$\boldsymbol{s_1} = \boldsymbol{Ls}, \boldsymbol{d_1} = \boldsymbol{Hs}, \tag{2}$$

$$\boldsymbol{s} = \boldsymbol{L}^T\boldsymbol{s_1} + \boldsymbol{H}^T\boldsymbol{d_1}, \tag{3}$$

where **L** and **H** are the wavelet matrix filters.

2) 2D-DWT

For 2D signal **X**, the DWT does 1D-DWT on every row and column, i.e.,

$$\mathbf{X}_{ll} = \mathbf{LXL}^T; \mathbf{X}_{lh} = \mathbf{HXL}^T; \mathbf{X}_{hl} = \mathbf{LXH}^T; \mathbf{X}_{hh} = \mathbf{HXH}^T, \tag{4}$$

3) n-level 2D-DWT

We can obtain its features from a 1-level discrete wavelet decomposition using Eq. (4):

$$\mathbf{X}^1 = [\mathbf{X}_{ll}, \mathbf{X}_{lh}, \mathbf{X}_{hl}, \mathbf{X}_{hh}]. \tag{5}$$

For a signal **X**, its features from an *n*-level discrete wavelet decomposition can be obtained as follows:

$$\begin{aligned} \mathbf{X}^n &= [\mathbf{X}_{ll}^{n-1}, \mathbf{X}_{lh}^{n-1}, \mathbf{X}_{hl}^{n-1}, \mathbf{X}_{hh}^{n-1}] \\ &= [\mathbf{LX}^{\boldsymbol{n-1}}\mathbf{L}^T, \mathbf{HX}^{\boldsymbol{n-1}}\mathbf{L}^T, \mathbf{LX}^{\boldsymbol{n-1}}\mathbf{H}^T, \mathbf{HX}^{\boldsymbol{n-1}}\mathbf{H}^T] \end{aligned} \tag{6}$$

B. Spatial-enhanced Multi-level Wavelet Patch

1) Multi-level Wavelet Decomposition module

In the wavelet decomposition module, the initial step involves an adaptive resizing of the input image to a resolution of 448×448. This strategic resizing facilitates a deeper exploration into the nuanced capabilities offered by intricate wavelet decompositions. Notably, this preprocessing action is optional, and its adoption depends on the requirements of the target application and the specific characteristics of the dataset.

After the resizing phase, our methodology invokes a multi-level wavelet decomposition. As visualized in Fig. 1, a single iteration of wavelet decomposition segregates the image into four distinct frequency bands: **LL**, **LH**, **HL**, and **HH**. Each of these bands captures different aspects of the image:

LL (Low-Low): This band retains the approximate coefficients and primarily embodies the lower frequency components, representing the more global structures and broader contours of the image.

LH (Low-High) & HL (High-Low): These bands capture the horizontal and vertical details, respectively, often resonating with the edges and directional features within the image.

HH (High-High): Representing the diagonal details, this band captures the high-frequency components, typically linked with textural nuances and finer granularity in the image content.

As the image undergoes multiple layers of wavelet decomposition, the process is repeated until the pixel points' count corresponds precisely with the intended token count in the ViT. At this critical juncture, a synthesis is performed using the information extracted from the varied frequency bands to construct and enrich each token. Consequently, these comprehensive tokens, replete with multispectral features from distinct decomposition bands, are seamlessly integrated into the transformer encoder.

2) Spatial-enhanced module

Wavelet decomposition excels in extracting hidden frequency details from images, such as texture frequencies and edge delineations, which is particularly effective in grayscale images. However, its application to RGB color channels in parallel independently can diminish the color richness and struggle with capturing directional texture patterns.

We integrate a spatial-enhanced module into our wavelet decomposition process to address these issues. Our method innovatively combines joint convolution and down-sampling by altering the convolution stride. This simultaneous process efficiently preserves essential spatial domain features, treating them as a unique frequency component. The down-sampled spatial information is then merged with frequency-domain features, enriching the network encoder's feature space.

This approach maintains spatial and wavelet nuances while preserving inter-channel correlations in RGB images. The stride-modified convolution strategy mitigates the loss of color richness and bolsters texture direction cues. Our method thus enhances the encoder's capability to integrate comprehensive spatial information, offering a robust solution for maintaining color fidelity and textural details in wavelet-transformed images. This combined technique of spatial enhancement and strategic convolution presents a balanced solution, addressing the inherent limitations of standard

Algorithm 1: Spatial-enhanced Multi-level Wavelet Patch

Input:
X: original images of shape [B, 3, 224, 224];
n: levels of wavelet decomposition.
Output:
T: tokens of shape [B, N, C]. ((Batch,196,768) in ViT-B/16).
1: $\mathbf{X}_r$ = Resize (**X**, (448,448))
2: $\mathbf{X}_n = DWT^n$ ($\mathbf{X}_r, n$)
3: $\mathbf{X_S}$ = Conv2d (input = 3, output = 4^n, kernel = 2^n, stride = 2^n)($\mathbf{X}_r$)
3: $\mathbf{X_T}$= Conv2d (input = $4 * 4^n$, output = C, kernel = 2^{5-n}, stride = 2^{5-n})([$\mathbf{X_S}$, $\mathbf{X}_n$])
4: **T** = Rearrange ('b e (h) (w)' -> 'b (h w) e') ($\mathbf{X_T}$)
5: **Return T**

wavelet decomposition in color image processing.

3) Projection & Transpose

After merging features from wavelet and spatial domains, we adopt a convolution projection layer, mirroring standard ViT processing, to adjust channel sizes and ensure token dimensions align with the subsequent encoder's requirements. Following ViT's structure, a Transpose operation immediately follows, yielding the final tokens. The specifics of our method are outlined in Algorithm 1, presenting our Spatial-Enhanced Multi-level Wavelet Patch (SMWP).

A distinguishing characteristic of our method is its ability to capture many feature types adeptly. With transformer encoders fed these enhanced tokens that span various spectra, they can keenly identify intricate phase texture patterns (evident in the high-frequency HH band) and broader structural interrelationships stemming from the subtle low-frequency LL band. Fig. 2 contrasts the information within each signal token from a standard ViT image patch versus our wavelet patch.

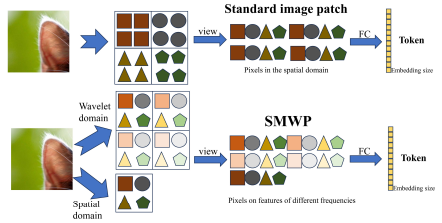

Fig. 2. The difference between the standard image patch and our wavelet patch. In standard ViT image patching, pixels of a local spatial domain are projected into a 'word' (token) according to the embedding size. In a wavelet patch, the features of different frequency bands of the entire image are projected into a 'word' based on the image block.

III. Experiments

To evaluate SMWP's advantage over ViT's standard spatial patching, we used ImageNet100, a 100-class subset of ImageNet1000 [20] (details in the Supplementary material). Our SMWP, consistent with ViT, utilizes a single convolution for dimension projection, preserving the original computational and parameter footprint. This module ensures parameter consistency with the original network's public code. Given their variety, the network parameters are documented in the Supplementary material. Experiments were conducted on an RTX4090 GPU with PyTorch [21]. The primary batch size is 128, with exceptions set at 64.

A. Ablation Study

In this section on the study of SMWP architecture, we focus exclusively on variations within the image patch component, maintaining all other parameters constant to guarantee result accuracy.

1) Wavelet features

Objective: To systematically assess wavelet features' efficacy in tasks compared to original spatial features (ViT-B/16) and examine the impact of different wavelet decomposition layers on overall framework performance.

Theoretical Insight: Wavelet transform's strength lies in its capability for multi-level signal decomposition. The challenge is determining the optimal number of decomposition layers and selecting spectral images for the most compelling feature representation.

Empirical Outcomes: Experiments conducted without the spatial-enhanced module, using a multi-level wavelet patch (MWP) module instead, are presented in Table I. Results indicate peak performance with four wavelet decompositions, surpassing the spatial domain features (ViT-B/16). Each additional decomposition level decreases resolution but enriches frequency domain features. For example, Level 1 decomposition reduces feature resolution to a quarter, with further reductions at each subsequent level. At Level 5, resolution is minimal, with tokens representing distinct spectra. Four decompositions emerge as the optimal balance; fewer levels are too similar to spatial domain patches, while more levels obscure image details and frequency clarity. *(The wavelet transformation in MWP, conducted independently during image preprocessing, does not participate in backpropagation (BP), thus not impacting the network's training time.)*

TABLE I. WAVELET PATCH OF DIFFERENT LEVEL

Network	ViT-B/16	MWP 1-level	MWP 2-level	MWP 3-level	MWP 4-level	MWP 5-level
Top-1 Accuracy	70.71% (0.018)	71.27% (0.0072)	71.8% (0.04)	71.87% (0.02)	72.38% (0.18)	71.58% (0.34)
Training Time	17.54h	14.24h	14.191h	14.23h	14.33h	15.12h

2) Spatial Domain Integration

Objective: To explore the potential gains of merging information from the spatial domain.

Theoretical Insight: The fusion of spatial domain intricacies into tokens potentially enriches the mosaic of encapsulated features, offering a more panoramic feature perspective.

Empirical Outcomes: Table II shows that merging spatial-domain features with wavelet patching consistently enhances model performance. This suggests the value of spatial-domain details like color and texture direction, which might be overlooked with only wavelet features. Notably, this integration leads to a significant 2.4% accuracy increase over the baseline ViT-B/16 with four wavelet decompositions. The SMWP variant displays notable variance attributed to downsampling spatial attributes to match wavelet resolutions. We used a single-layer downsampling approach to keep computational demands consistent, though a multi-layer method might reduce this variance, a subject for future research. The Spatial-enhanced module is involved in BP, and the DWT module can still be left out of the operation, and we compare the training time of networks.

TABLE II. SPATIAL-ENHANCED WAVELET PATCH

Network	ViT-B/16	SMWP 1-level	SMWP 2-level	SMWP 3-level	SMWP 4-level	SMWP 5-level
Top-1 Accuracy	70.71% (0.018)	72.45% (0.71)	72.89% (0.27)	72.90% (0.66)	73.11% (0.004)	72.78% (0.56)
Training Time	17.54h	14.54h	14.73h	14.68h	15.32h	15.17h

3) Exploration of Wavelet Basis Functions

Objective: To navigate the landscape of wavelet basis functions and discern the most congruous role for the task.

Theoretical Insight: Wavelet bases with distinct operational strengths—ranging from Haar wavelets' edge discernment capabilities to the multifaceted attributes of Symlets and

DMeyer wavelets—potentially harbor the key to optimal model performance.

Empirical Outcomes: Rigorous evaluations, framed within the context of 4-level decomposition, brought forth the performance metrics of these wavelet bases, all of which have been meticulously cataloged in Table III. Upon innovatively integrating wavelet patching into the ViT image patching module, a consistent improvement was observed across various wavelet functions over the baseline ViT-B/16 model (70.71%). Notably, the Sym4 and Bior1.1 wavelets delivered top-tier performance, achieving 73.4% and 73.28% accuracy, respectively. Even wavelet functions with modest accuracy increments, like the Daubechies series, still surpassed the baseline. This consistent enhancement across multiple wavelets underscores the potential and efficacy of wavelet patching, suggesting its promising role in future ViT adaptations.

TABLE III. DIFFERENT WAVELET BASIS

Wavelets basis	ViT-B/16	Haar	Biorthogonal				Dmeyer
			Bior1.1	Bior2.2	Bior3.3	Bior3.5	
Top-1 Accuracy	70.71%	73.11%	73.28%	72.48%	71.76%	71.83%	72.99%
Wavelets basis	Daubechies			Symlets			
	Db1	Db4	Db8	Sym2		Sym4	Sym8
Top-1 Accuracy	72.13%	72.21%	71.95%	73.14%		**73.4%**	72.48%

B. Extended Study

1) Performance Improvement w.r.t. Network Depth.

Our SMWP seamlessly integrates into ViT networks across varying complexities. Table IV elucidates the performance enhancement of the wavelet patch across diverse ViT sizes, underscoring the sustained benefits of spatial-enhanced wavelet patching.

TABLE IV. DIFFERENT ViT SIZE

Different Size (Model size/Patch size)	Base/32	Base/16	Large/32	Large/16
ViT	63.76%	70.71%	53.98%	68.87%
MWP ViT	62.78%	72.56%	59.39	69.13%
SMWP ViT	**64.2%**	**73.4%**	**60.52%**	**70.13%**

2) Wavelet Patch in other Advanced Transformers

Our module serves as a patching method, and unlike previous networks that improve the local feature extraction ability of ViT through a complex structure, we are seamlessly implanting Transformer-based networks without changing the design of the network, which can be used as an alternative to ViT's backbone network. We validate our ability as a new backbone on multiple ViT variants. Table V provides a comprehensive breakdown of the results. The tags '-Ti', '-S', and '-B' differentiate models by size, while '-7' and '-14' indicate the depth variations in T2T models. Notably, our method elevates T2T-7 performance from 84.82% to 85.38%. There is an impressive gain of 4.32% for TNT-S, a 1.82 % improvement on Pyramid-TNT-Ti, and notable increments ranging from 0.76% to 1.26% percentage points on Wave-ViT and Dual-ViT. Such outcomes underscore the adaptability and efficacy of our wavelet patches across transformer networks. Despite these networks having more complex image patching procedures than ViT, we detail a straightforward approach to integrate our wavelet patch in the Supplementary material, ensuring the original network architectures remain undisturbed.

3) Low Sensitivity to Position Embedding of Wavelet Patch

From the wavelet transform defined in Eq. 1, it is evident that features from various spectra are derived from the comprehensive spatial domain information of the original image. As evidenced by the experimental results, a consequential advantage of our wavelet patch is its capability to attenuate the ViT's dependence on location encoding. To substantiate this, we removed the position embedding module from the ViT architecture, and the outcomes are cataloged in Table VI. Despite forgoing location encoding, our network maintains a commendable accuracy rate in the ballpark of 69-70%. This indicates that our wavelet-infused enhancement bestows the ViT with low sensitivity to location encoding, potentially eliminating the need for such modules in specific scenarios, thus streamlining the architecture and improving efficiency.

TABLE V. SMWP IN LOCAL ATTENTION-ENHANCED TRANSFORMERS

Network size	-7	-14	Network size	-S	-B (Batch 64)
T2T[7]	84.82%	86.00%	TNT[8]	69.28%	71.22%
MWP T2T	85.34%	85.76%	MWP TNT	72.82%	72.55%
SMWP T2T	**85.38%**	**86.16%**	SMWP TNT	**73.60%**	**73.30%**
Network size	-Ti	-S	Network size	-Ti	-S (Batch 64)
Swin [22]	86.72%	87.18%	Pyramid TNT[23]	76.88%	78.64%
MWP Swin	86.70%	87.00%	MWP Pyramid	**78.70%**	77.74%
SMWP Swin	**87.18%**	**87.58%**	SMWP Pyramid	78.60%	**79.52%**
Network size	-S	-B (Batch 64)	Network size	-S	-B (Batch 64)
Wave-ViT [14]	77.54%	79.34%	Dual-ViT [24]	78.20%	76.26%
MWP Wave	77.44%	**80.60%**	MWP Dual	78.91%	**77.48%**
SMWP Wave	**78.30%**	80.24%	SMWP Dual	**79.01%**	77.10%

TABLE VI. LOW SENSITIVITY TO POSITION EMBEDDING

Network		1-level	2-level	3-level	4-level	5-level
With position embedding	MWP_ViT-B/16	71.27%	71.8%	71.87%	72.38%	71.58%
	SMWP_ViT-B/16	72.45%	72.89%	72.90%	73.11%	72.78%
Without position embedding	MWP_ViT-B/16	69.28%	69.26%	69.54%	70.53%	69.10%
	SMWP_ViT-B/16	69.02%	70.66%	70.95%	71.11%	69.28%

IV. Conclusion

Our research introduces a groundbreaking enhancement in the ViT texture and local feature extraction capabilities through the integration of a SMWP into its image patching module. This advancement significantly outperforms traditional spatial domain patching methods. Our approach allows for more effective capture and analysis of textures and intricate local details in images, surpassing previous ViT models that depended on larger datasets or complex structural modifications.

The key feature of our method is its exceptional effectiveness in texture feature extraction, enabling ViT to discern fine-grained details. This capability is vital across various applications, especially in critical detail and texture interpretation areas. Our patching method's adaptability across different ViT architectures showcases its potential as a universal upgrade, offering compatibility with various ViT variants without necessitating extensive architectural changes. This versatility positions our approach as a potential new backbone network for ViT models, representing an update to their core functionality.

REFERENCES

[1] Dosovitskiy, Alexey, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations*, 2020.

[2] H. Yin and S. Ma, "CSformer: Cross-Scale Features Fusion Based Transformer for Image Denoising," *IEEE Signal Processing Letters*, vol. 29, pp. 1809–1813, 2022.

[3] J. Li, J. Zhu, C. Li, X. Chen, B. Yang, and Y. Bin, "CGTF: Convolution-Guided Transformer for Infrared and Visible Image Fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[4] N. Cheng, Z. Sun, X. Zhu, and H. Wang, "A transformer-based network for perceptual contrastive underwater image enhancement," *Signal Processing: Image Communication*, vol. 118, p. 117032, 2023.

[5] Y. Sun, R. Ni, and Y. Zhao, "ET: Edge-enhanced Transformer for Image Splicing Detection," *IEEE Signal Processing Letters*, vol. 29, pp. 1232–1236, 2022.

[6] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A Survey of Visual Transformers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2023.

[7] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[8] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in Transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15908–15919, 2021.

[9] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision Outlooker for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2022.

[10] T. Yu, G. Zhao, P. Li, and Y. Yu, "BOAT: Bilateral Local Attention Vision Transformer," *arXiv*:2201.13027, 2022.

[11] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the Design of Spatial Attention in Vision Transformers," *Neural Information Processing Systems*, 2021.

[12] Zhang, Guiwei, et al., "PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2023.

[13] Ding, Meidan, et al., "An enhanced vision transformer with wavelet position embedding for histopathological image classification," *Pattern Recognition*, 2023, 140: 109532.

[14] Yao, Ting, et al., "Wave-vit: Unifying wavelet and transformers for visual representation learning," *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

[15] Zhuang, Yufan, et al., "Waveformer: Linear-Time Attention with Forward and Backward Wavelet Transform," *arXiv*:2210.01989, 2022.

[16] Huang, Hao, and Yi Fang, "Adaptive wavelet transformer network for 3d shape representation learning," I*nternational Conference on Learning Representations*, 2021.

[17] Tao, Zhiyong, Tong Wei, and Jie Li, "Wavelet multi-level attention capsule network for texture classification," *IEEE Signal Processing Letters*, vol. 28, pp. 1215-1219, 2021.

[18] Li, Hao, et al., "DnSwin: Toward real-world denoising via a continuous Wavelet Sliding Transformer," *Knowledge-Based Systems*, vol. 255, pp. 109815, 2022.

[19] Ai, Yuang, et al., "SOSR: Source-Free Image Super-Resolution with Wavelet Augmentation Transformer," *arXiv*:2303.17783, 2023.

[20] Russakovsky, Olga, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211-252, 2015.

[21] Paszke, Adam, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[22] Liu, Ze, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[23] Han, Kai, et al., "Pyramidtnt: Improved transformer-in-transformer baselines with pyramid architecture," *arXiv*:2201.00978, 2022.

[24] Yao, Ting, et al., "Dual vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, pp. 10870, 2023.