DOI: 10.19884/j.1672-5220.202412001

A Wavelet Transform and Spatial Positional Enhanced Method for Vision Transformer

HU Runyu[1], TANG Xuesong[1,2*], HAO Kuangrong[1,2]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: In the vision transformer (ViT) architecture, image data are transformed into sequential data for processing, which may result in the loss of spatial positional information. While the self-attention mechanism enhances the capacity of ViT to capture global features, it compromises the preservation of fine-grained local feature information. To address these challenges, we propose a spatial positional enhancement module and a wavelet transform enhancement module tailored for ViT models. These modules aim to reduce spatial positional information loss during the patch embedding process and enhance the model's feature extraction capabilities. The spatial positional enhancement module reinforces spatial information in sequential data through convolutional operations and multi-scale feature extraction. Meanwhile, the wavelet transform enhancement module utilizes the multi-scale analysis and frequency decomposition to improve the ViT's understanding of global and local image structures. This enhancement also improves the ViT's ability to process complex structures and intricate image details. Experiments on CIFAR-10, CIFAR-100 and ImageNet-1k datasets are done to compare the proposed method with advanced classification methods. The results show that the proposed model achieves a higher classification accuracy, confirming its effectiveness and competitive advantage.

Keywords: transformer; wavelet transform; image classification; computer vision
CLC number: TP183 **Document code**: A
Article ID: 1672-5220(2025)03-0330-09

Open Science Identity (OSID)

0 Introduction

Convolutional neural networks (CNNs) are leading frameworks for modeling the human visual system, achieving notable breakthroughs in computer vision tasks like image classification[1-4], object detection[5-8] and semantic segmentation[9-11]. CNNs primarily use convolutional layers to iteratively extract multi-scale features from images, capturing texture, brightness and color information across pixel levels to facilitate efficient feature representation learning. These features are abstracted through the CNN's layered structure, enabling complex pattern modeling and recognition in higher-level feature spaces. To perform downstream visual tasks, CNNs typically include a few fully connected layers after the convolutional layers to integrate and classify the extracted features. However, achieving high performance in applications often demands extensive, high-quality and well-annotated training datasets. Furthermore, CNN architectures have become increasingly complex in design. The optimization process requires careful hyperparameter tuning to ensure the generalization and stability across tasks and datasets.

Transformer-based models have recently attracted significant attention owing to the success of their self-attention mechanism. Initially developed for natural language processing (NLP) tasks, transformers[12], with suitable modifications, have shown outstanding performance in the computer vision. The vision transformer (ViT)[13] is the first model to successfully adapt the transformer architecture for visual tasks, introducing a fresh perspective to traditional computer vision paradigms. In the ViT model, images are divided into patches treated as sequential data, allowing the transformer to process visual information similarly to word sequences in natural language. This method effectively encodes the image content within the transformer architecture. Unlike CNNs, transformers lack translational invariance, resulting in a reduced efficiency for capturing local features and structured information in images. This limitation is especially pronounced in cases where the dataset is small or lacks sufficient diversity. Furthermore, transformers generally demand larger datasets or advanced data augmentation strategies to achieve optimal training performance.

The wavelet transform[14] is a versatile mathematical

Received date: 2024-12-02
Foundation item: National Natural Science Foundation of China (No. 62176052)
* Correspondence should be addressed to TANG Xuesong, email: tangxs@dhu.edu.cn
Citation: HU R Y, TANG X S, HAO K R. A wavelet transform and spatial positional enhanced method for vision transformer[J]. *Journal of Donghua University (English Edition)*, 2025, 42(3): 330-338.

tool that decomposes signals into components at different time scales, simultaneously providing time and frequency information. In the image processing, wavelet transform offers notable advantages due to its multi-scale analysis and sparse representation properties. Wavelet transform facilitates multi-scale image decomposition, capturing detailed features across resolutions and excelling in processing complex textures or edges. This multi-scale capability preserves global structures and enhances local details, providing a comprehensive image representation. This enables superior performance in tasks like the image denoising, edge detection and texture analysis.

This paper presents a novel feature enhancement module leveraging wavelet transform to capture image features across multiple frequency scales. This design substantially enhances the model's feature extraction capabilities while effectively mitigating the limitations of the ViT model in processing fine-grained information. Furthermore, a spatial positional enhancement module (SPEM) is introduced to address the loss of spatial positional information incurred during the image serialization process. Comprehensive experiments are conducted on CIFAR-10, CIFAR-100 and ImageNet-1k[15] datasets to demonstrate the efficacy and robustness of the proposed model.

1 Related Work

The success of transformers in NLP has driven extensive exploration of their applications in the computer vision. Since the transformer architecture was first introduced in visual classification models, transformers have rapidly evolved into one of the core frameworks in the field of the computer vision. Despite their impressive modeling capabilities, transformers face challenges in practical applications due to their high computational complexity.

Deformable-detection transformer (Deformable-DETR)[16] leverages deformable convolutions, introducing a flexible attention mechanism to reduce the computational load and improve the model adaptability. Axial attention performs self-attention along a single axis of the input tensor, avoiding global computations on flattened high-dimensional data and significantly reducing computational costs. Compact vision transformer (CVT)[17] introduces the SeqPool method, which efficiently aggregates token sequence outputs to preserve key information from each part of the input image. SeqPool outperforms traditional learnable class token approaches by significantly improving model performance without adding extra parameters. While ViT demonstrates the strong potential of transformers in visual tasks, its reliance on large-scale datasets limits its generalization and applicability across diverse scenarios. To address this limitation, the data-efficient image transformer (DeiT)[18] uses a teacher-student distillation training strategy and incorporates token distillation to enhance the training efficiency. Compared to ViT, DeiT speeds up training by introducing a distillation token that closely aligns with the teacher model's output. This model improves the training efficiency and preserves performance, offering a more practical solution for applying ViT.

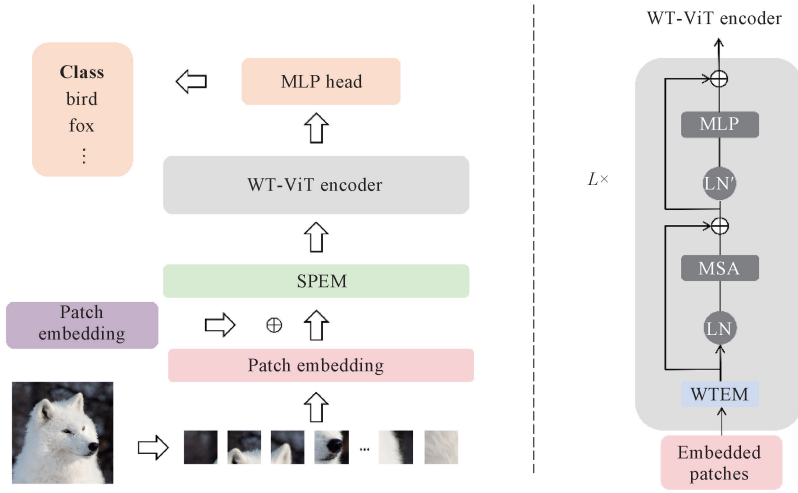
Patch embedding is a fundamental operation that converts image data into sequential information. This process divides an input image into patches of the defined size, segmenting it along its width and height. Each patch undergoes a linear transformation, projecting it into one-dimensional (1D) space and flattening the two-dimensional (2D) image into a 1D vector sequence. The compact convolutional transformer (CCT)[17] replaces traditional patch embedding with convolutional operations, directly applying convolutions to the input image and flattening the outputs into the input sequence. This model inherently captures positional information through convolutions. Building on this, the convolution-enhanced image transformer (CeiT)[19] enhances the process by replacing the conventional patch embedding with convolutional and pooling layers. The resulting tokens are rearranged into feature maps, and feature maps undergo local processing via depthwise separable convolutions. During this processing, linear layers generate new tokens, improving spatial relationships among adjacent tokens. This method enhances the self-attention mechanism, significantly strengthening model performance in visual tasks.

The conditional position encoding for ViT (CPVT)[20] introduces a hybrid architecture that integrates CNNs and transformers. By employing convolutions to capture spatial information in place of positional encoding, CPVT effectively enriches positional information for sequences. The convolutional neural networks meet ViT(CMT)[21] builds on the strengths of both CNNs and transformers, introducing a novel hybrid network architecture. This design leverages transformers to capture long-range dependencies and CNNs for local feature extraction, striking an improved balance between performance and efficiency. Expanding on CMT, the retentive networks meet ViT (RMT)[22] incorporates RetNet's retention mechanism, and employs the explicit decay to model 1D distance priors. RMT extends this concept to two-dimensional space by introducing Manhattan self-attention, thereby enhancing spatial modeling capabilities for visual tasks. The ViT with convolutional multi-scale feature interaction for dense predictions (ViT-CoMer)[23] integrates multi-scale convolutional features into the ViT architecture, thereby enhancing its performance on dense prediction tasks. The revisiting mobile CNN from ViT perspective (RepViT)[24] incorporates the efficient architectural designs of the lightweight ViT into CNNs, achieving superior performance over existing lightweight ViTs across various vision tasks. Image retrieval based on ViT

and masked learning[25] proposes an image retrieval framework by using an masked auto-encoder, achieving significant accuracy improvements.

2 Methods

In the patch embedding process of ViT, the image is segmented into multiple patches and transformed into sequential information, inevitably resulting in the loss of spatial positional information. As shown in Fig. 1, to address this issue, this paper designs an SPEM that leverages convolutional operations and multi-scale feature capture to enrich the spatial information within the sequence. Additionally, a wavelet transform enhancement module (WTEM) is introduced to improve the performance in feature extraction. With its multi-scale analysis and frequency decomposition capabilities, wavelet transform significantly enhances the robustness of the ViT. When processing images, wavelet transform effectively extracts features at different frequencies, ensuring the preservation of critical structural information even in the presence of noise or missing data. Furthermore, wavelet transform excels in extracting shape and texture features. By decomposing the image, it can independently capture shape and texture characteristics, enabling the model to better understand and distinguish object boundaries and details in image recognition and classification tasks. This reinforcement of shape and texture feature learning leads to more robust and accurate performance of the model in complex scenes. The integration of these two modules aims to strengthen the model's feature extraction capacity and its sensitivity to spatial positional information. We name the entire model wavelet transform vision transformer (WT-ViT).



MLP—multilayer perceptron; MSA—multi-head self-attention; LN—layer normalization.

Fig. 1 Overall architecture of WT-ViT

2.1 SPEM

This paper proposes an SPEM designed to improve the capability capturing spatial position information. As shown in Fig. 2, this module employs a convolutional structure, combining the characteristics of different convolutional layers to extract both fine details and contextual information from the input feature map. Specifically, the core of the SPEM consists of multiple convolutional layers, including a 3×3 convolutional layer and a 3×3 dilated convolutional layer. The use of the dilated convolution enables the model to effectively capture multi-scale features and rich spatial positional information, enhancing its perceptual ability regarding target objects.

After feature extraction, the module utilizes an attention mechanism that computes the mean and maximum values of the feature map to generate dynamic attention weights. The module further processes the combined features through convolutional layers to ensure compatibility with the output feature map dimensions, strengthening the model's focus on key features. The SPEM multiplies the weighted and fused feature map with the input feature map, thereby reinforcing the features and effectively extracting spatial position information.

Firstly, features $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ extracted using convolutional kernels with different receptive field ranges are concatenated to form a new tensor $\boldsymbol{x}$:

$$\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2]. \tag{1}$$

Secondly, channel-based average $P_{\text{avg}}(\cdot)$ and max pooling $P_{\max}(\cdot)$ are applied to effectively capture spatial relationships:

$$\begin{cases} \boldsymbol{x}_{\text{avg}} = P_{\text{avg}}(\boldsymbol{x}), \\ \boldsymbol{x}_{\max} = P_{\max}(\boldsymbol{x}), \end{cases} \tag{2}$$

where $\boldsymbol{x}_{\text{avg}}$ and $\boldsymbol{x}_{\max}$ denote the spatial features obtained from average pooling and max pooling, respectively. To enable information interaction between different spatial features, the two pooled features are concatenated and passed through a convolutional layer $F_1(\cdot)$ to transform the pooled features (with two channels) into two spatial attention maps $\boldsymbol{A}$:

$$\boldsymbol{A} = F_1([\boldsymbol{x}_{\text{avg}}, \boldsymbol{x}_{\max}]). \tag{3}$$

For the two spatial attention maps $\boldsymbol{A}_i$, the sigmoid activation function $\sigma(\cdot)$ is applied to obtain individual spatial selection masks at two distinct scales:

$$\boldsymbol{A}'_i = \sigma(\boldsymbol{A}_i),\ i = 1,\ 2. \tag{4}$$

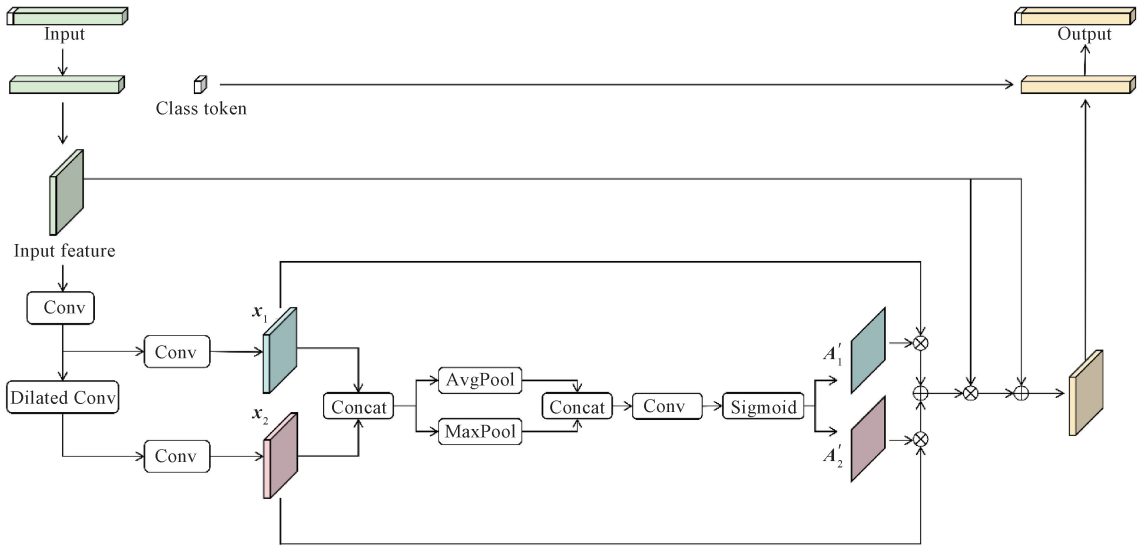
Then, the features extracted by convolutional kernels with different receptive field ranges are weighted by their respective spatial selection masks and fused through a convolutional layer $F_2(\cdot)$ to obtain the attention features $\boldsymbol{S}$:

$$\boldsymbol{S} = F_2(\boldsymbol{A}'_1 \cdot \boldsymbol{x}_1 + \boldsymbol{A}'_2 \cdot \boldsymbol{x}_2). \tag{5}$$

Finally, an element-wise product is performed between the input features $\boldsymbol{U}$ and the attention features $\boldsymbol{S}$, followed by a residual connection to produce the final output $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \boldsymbol{U} + \boldsymbol{U} \cdot \boldsymbol{S}. \tag{6}$$

By introducing spatial information at different time scales, this method addresses the limitations of the ViT model in capturing spatial positional information. This enhancement provides subsequent encoders with sequence information that is more semantically cohesive and spatially dependent. Such an improvement plays a crucial role in enhancing the model's sensitivity to local contexts, thereby improving the accuracy and efficiency of overall feature representation.



Conv—convolutional layers; AvgPool—average pooling; MaxPool—max pooling.

Fig. 2 Specific details of SPEM

2.2 WTEM

Wavelet transform enables the decomposition of input signals, allowing the network to capture information across different frequency and spatial scales, making it more effective for handling images with complex textures or edge details. This paper introduces a WTEM, combining wavelet transforms with standard convolutional operations to capture information across various frequency and spatial scales. This module significantly extends the receptive field, allowing small convolutional kernels to effectively cover a broader context. We adopt the Haar wavelet as the chosen wavelet transform implementation. The Haar wavelet is composed of simple rectangular functions and relies solely on addition and subtraction, thus eliminating the need for complex numerical calculations and offering a high computational efficiency. Furthermore, the orthogonality among Haar wavelet functions effectively prevents information redundancy. Owing to the demonstrated effectiveness of the Haar wavelet's in encoding the visual information, its features enrich image representation and enhance the performance of feature extraction.

In the wavelet transform operation, for an input image $\boldsymbol{X}$, a single-layer wavelet transform is implemented in 1D space (either width or height) using the depthwise convolution. To achieve the 2D wavelet transform, this operation is applied in both dimensions, using four filters in stride-2 depthwise convolutions: a low-pass filter $\boldsymbol{F}_{\text{ll}}$, and a set of high-pass filters $\boldsymbol{F}_{\text{lh}}$, $\boldsymbol{F}_{\text{hl}}$ and $\boldsymbol{F}_{\text{hh}}$.

$$\begin{cases} \boldsymbol{F}_{\text{ll}} = \dfrac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \\ \boldsymbol{F}_{\text{lh}} = \dfrac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \\ \boldsymbol{F}_{\text{hl}} = \dfrac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \\ \boldsymbol{F}_{\text{hh}} = \dfrac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \end{cases} \tag{7}$$

For each input channel, the convolution output comprises four channels, with each channel size of the feature map in each spatial dimension being half that of the input $\boldsymbol{X}$. $\boldsymbol{X}_{\text{ll}}$ represents the low-frequency component of

$\boldsymbol{X}$, while channels $\boldsymbol{X}_{\mathrm{lh}}$, $\boldsymbol{X}_{\mathrm{hl}}$ and $\boldsymbol{X}_{\mathrm{hh}}$ correspond to the high-frequency components in the horizontal, vertical and diagonal directions, respectively:

$$[\boldsymbol{X}_{\mathrm{ll}}, \boldsymbol{X}_{\mathrm{lh}}, \boldsymbol{X}_{\mathrm{hl}}, \boldsymbol{X}_{\mathrm{hh}}] = \mathrm{Conv}([\boldsymbol{F}_{\mathrm{ll}}, \boldsymbol{F}_{\mathrm{lh}}, \boldsymbol{F}_{\mathrm{hl}}, \boldsymbol{F}_{\mathrm{hh}}], \boldsymbol{X}). \quad (8)$$

Since the filters in Eq. (7) constitute an orthogonal basis, the inverse wavelet transform can be obtained through transposed convolutional operator Conv_trans(·):

$$\boldsymbol{X} = \mathrm{Conv\_trans}([\boldsymbol{F}_{\mathrm{ll}}, \boldsymbol{F}_{\mathrm{lh}}, \boldsymbol{F}_{\mathrm{hl}}, \boldsymbol{F}_{\mathrm{hh}}], [\boldsymbol{X}_{\mathrm{ll}}, \boldsymbol{X}_{\mathrm{lh}}, \boldsymbol{X}_{\mathrm{hl}}, \boldsymbol{X}_{\mathrm{hh}}]). \quad (9)$$

Cascade wavelet decomposition recursively decomposes the low-frequency components. The output of each decomposition layer consists of the following components, where WT(·) represents the wavelet transform operator:

$$[\boldsymbol{X}_{\mathrm{ll},i}, \boldsymbol{X}_{\mathrm{lh},i}, \boldsymbol{X}_{\mathrm{hl},i}, \boldsymbol{X}_{\mathrm{hh},i}] = \mathrm{WT}(\boldsymbol{X}_{\mathrm{ll},i-1}). \quad (10)$$

The WTEM employs a two-level wavelet transform, as illustrated in Fig. 3. Firstly, the class token is separated from the sequence. The remaining sequence elements are then reshaped into a feature map to facilitate further computations and processing. Secondly, wavelet transform is applied to the input feature map, decomposing it into low-frequency and high-frequency features. Through convolutional operations, the model extracts multi-scale features, and the features are then recombined using the inverse wavelet transform (IWT). During the feature reconstruction phase, the model retains information across various scales and further enhances the low-frequency features through convolutional operations. This process ensures the consistency in the number of input and output channels to facilitate subsequent processing of the sequence by encoder layers. Repeating this step two times enables the model to capture feature information across multiple frequency scales, thereby enhancing its feature extraction performance. Thirdly, the output feature map is reshaped back to the structure of the original sequence, and the class token is reintegrated at the beginning of the sequence. This module not only improves the model's comprehensive understanding of both global and local image structures but also significantly enhances its ability to perceive rich details and complex structures within images.

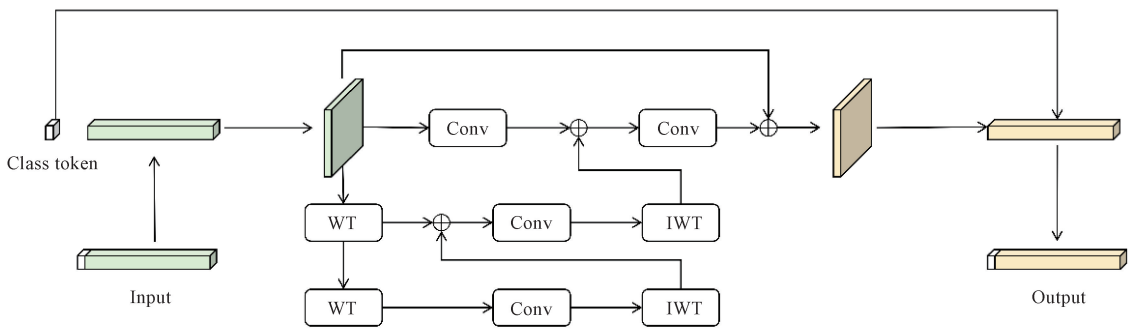


Fig. 3 Specific details of WTEM based on wavelet transform

3 Experiments

3.1 Datasets and experimental settings

To validate the effectiveness of the proposed model, tests are conducted on three public datasets: CIFAR-10, CIFAR-100 and ImageNet-1k.

CIFAR-10 is a standard dataset commonly used for image classification tasks. It contains 60 000 color images (32 × 32 pixels) across 10 categories, with 6 000 images per category. The dataset is split into 50 000 training images and 10 000 test images.

CIFAR-100 is another standard image classification dataset, similar to CIFAR-10, but it includes 100 categories with 600 images per category, totaling 60 000 color images (32 × 32 pixels). These images are divided into 50 000 training images and 10 000 test images. CIFAR-100 provides a more fine-grained classification task compared to CIFAR-10.

ImageNet-1k is a large-scale image classification dataset with 1 000 categories, approximately 1.2 million training images, and 50 000 test images. These categories encompass a wide range of objects, animals and scenes. Since 2009, ImageNet-1k has served as a crucial benchmark in computer vision, significantly impacting large-scale image classification and the training of deep learning models.

The model used in this study features 12 layers, an embedding dimension of 768, and an input image size of 224 × 224 pixels. The programming environment includes Python 3.8, CUDA 11.8 and the deep learning framework PyTorch 2.1.0. The hardware environment includes an Intel (R) Xeon (R) Silver 4110 CPU @ 2.10 GHz and an RTX 3090 GPU with 24 GB memory.

To ensure a fair comparison, all models are trained on their respective training sets, and the accuracy is evaluated on the test set. During training, the AdamW optimizer is employed with an initial learning rate of 5×10^{-5}, decayed following a cosine annealing schedule. Momentum is set to 0.9, weight decay to 0.05 and batch size to 128, with training conducted over 150 epochs.

3.2 Experimental analyses

As shown in Table 1, the WT-ViT is comprehensively compared with advanced models for

image classification on the CIFAR-10 and CIFAR-100 datasets. On the CIFAR-10 dataset, the proposed model achieves a classification accuracy of 96.67%, demonstrating the exceptional performance. This result highlights the model's effectiveness in natural image classification tasks. On the more complex and diverse CIFAR-100 dataset, the model demonstrates robust generalization with a classification accuracy of 76.67%. This result underscores the model's capability to handle a significantly larger number of classes. The success on smaller datasets addresses the limitations of traditional ViT models in low-data scenarios, optimizing their generalization capabilities.

Additionally, the proposed model is evaluated on the large-scale ImageNet-1k dataset, and the results are shown in Table 2. The model achieves 76.66% of a Top-1 accuracy and 93.65% of a Top-5 accuracy, demonstrating the strong performance. These results validate the method's effectiveness in enhancing performance on smaller datasets and achieving improvements on large datasets like ImageNet-1k. This demonstrates that the approach enhances the model's capability to handle complex datasets and improves feature extraction, enabling the accurate recognition of key features in diverse and large-scale data.

Table 1 Comparison of WT-ViT with other models on CIFAR-10 and CIFAR-100 datasets

Model	Number of parameters/MB	Top-1 accuracy/%	
		CIFAR-10	CIFAR-100
VGG19[3]	144	91.93	72.13
ResNet50[2]	25	92.15	72.05
EfficientNet-B5[26]	30	93.43	74.43
ViT-B/16[13]	87	88.17	67.63
CaiT[27]	68	90.33	69.35
CPVT[20]	88	91.15	71.80
CF-ViT[28]	27	90.65	71.13
Evo-VIT[29]	88	92.70	73.37
WT-ViT	95	**96.67**	**76.67**

Table 2 Comparison of WT-ViT with other models on ImageNet-1k dataset

Model	Top-1 accuracy/%	Top-5 accuracy/%
VGG19[3]	72.32	90.82
ResNet50[2]	74.12	92.84
EfficientNet-B5[26]	73.37	90.97
ViT-B/16[13]	72.74	91.55
CaiT[27]	73.51	90.35
CPVT[20]	74.21	93.14
CF-ViT[28]	72.66	91.93
Evo-VIT[29]	74.13	92.44
WT-ViT	**76.66**	**93.65**

3.3 Ablation studies

This subsection presents detailed ablation studies on the CIFAR-10 and CIFAR-100 datasets to evaluate the effectiveness of the proposed model. In this section, the ViT model is employed, featuring a depth of 12 layers and a hidden dimension of 768. Specific ablation experiments are designed to evaluate the contributions of the SPEM and the WTEM to model performance. The role of each module in enhancing the model performance is individually verified. Two modules are removed from the model separately to evaluate their impact on the classification accuracy and generalization capability. Then, experiments are performed on models with different wavelet transform depths to assess their impact on the model performance.

As shown in Table 3, introducing SPEM and WTEM individually improves performance on both CIFAR-10 and CIFAR-100 datasets. The model achieves optimal performance when both modules are integrated, demonstrating the proposed model's effectiveness in enhancing feature extraction.

Table 4 presents the experimental results for models with varying wavelet transform depths and highlights their impact on performance. A two-layer wavelet transform outperforms a single-layer configuration, suggesting that extracting features across multiple frequency scales enhances the model's ability to capture the image information and improve the classification performance. However, increasing the wavelet transform depth beyond two layers results in performance degradation, confirming that the two-layer configuration is optimal. Wavelet transform reduces the size of feature maps, and with the increased depth, the maps become excessively small, hindering effective feature learning during convolutional operations and ultimately degrading model performance.

Table 3 Ablation study with respect to SPEM and WTEM

Model	Top-1 accuracy/%	
	CIFAR-10	CIFAR-100
ViT	88.17	67.63
ViT+SPEM	91.02	70.15
ViT+WTEM	94.32	74.96
ViT+SPEM+WTEM	96.67	76.67

Table 4 Impact of wavelet transform depth on model

Depth	Top-1 accuracy/%	
	CIFAR-10	CIFAR-100
Raw	91.02	70.15
Single-layer	95.17	75.56
Two-layer	96.67	76.67
Three-layer	95.03	74.93

3.4 Visualization

This subsection employs Grad-CAM[30] to visualize the model's attention maps, illustrating the regions that the network focuses on for a given class. As shown in

Fig. 4, ViT's attention is scattered across background regions; WT-ViT's attention concentrates more effectively on the target class, improving conditions for accurately capturing object features.

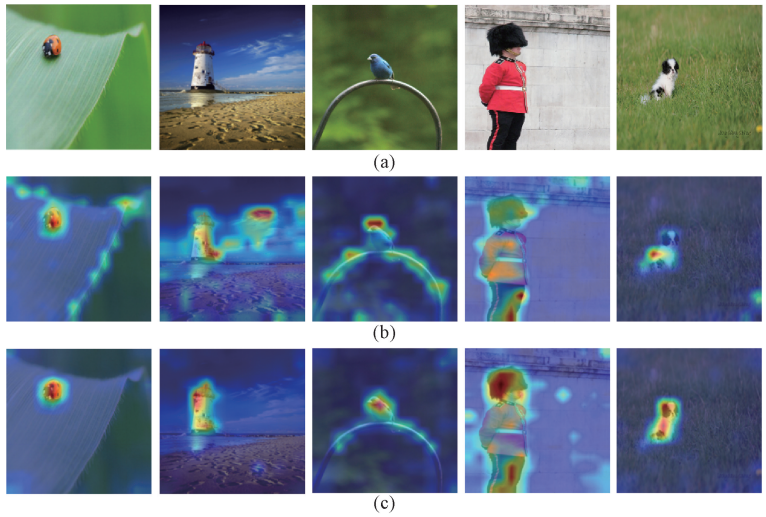


Fig. 4 Visualization of attention maps by different models: (a) origin images; (b) ViT's attention maps; (c) WT-ViT's attention maps

The focused attention distribution demonstrates that WT-ViT equips the model with enriched spatial positional information and multi-scale frequency-domain features, significantly enhancing its feature extraction capability. The enhanced focus on target areas improves the classification accuracy and suggests better interpretability, as the model now emphasizes key image regions relevant to the target class. This concentrated attention mechanism enables the ViT model to more effectively utilize both local and global contextual information, achieving a robust representation of complex visual data.

4 Conclusions

The proposed spatial positional enhancement and wavelet transform feature modules enhance the ViT's ability to capture spatial position and frequency-domain information, effectively addressing its limitations in positional encoding and local feature extraction. Experiments on CIFAR-10, CIFAR-100 and ImageNet-1k reveal notable performance improvements over state-of-the-art models, demonstrating the model's robustness across both small-scale and large-scale datasets. Grad-CAM visualizations confirm that our method effectively directs the model's attention to critical object regions, enhancing feature focus and interpretability. Overall, WT-ViT improves the classification accuracy and the model's adaptability to complex image data, providing a foundation for further exploration of spatial and frequency-domain integration in transformer-based architectures.

WT-ViT paves the way for integrating multi-scale and frequency-domain information into transformer-based architectures for computer vision tasks. Future work could extend these enhancement modules to other transformer-based models and evaluate their performance on diverse datasets or in tasks like object detection and segmentation. Moreover, optimizing the computational efficiency of the wavelet transform module for faster processing on large datasets could further improve its practical utility. Overall, WT-ViT lays a strong foundation for advancing spatial position and frequency-domain integration in deep learning models, with the goal of enhancing their robustness and precision in complex visual scenarios.

References

[1] WANG K P, ZHAO M B. Region-aware fashion contrastive learning for unified attribute recognition and composed retrieval[J]. *Journal of Donghua University (English Edition)*, 2024, 41(4): 405-415.

[2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 770-778.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2024-11-20]. https://arxiv.org/abs/1409.1556v6.

[4] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 11966-11976.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on

Computer Vision and Pattern Recognition (CVPR). New York:IEEE, 2016: 779-788.
[6] GIRSHICK R. Fast R-CNN[EB/OL]. (2015-09-27)[2024-11-20]. https://arxiv.org/abs/1504.08083.
[7] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
[8] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 2980-2988.
[9] WENG W H, ZHU X. INet: convolutional networks for biomedical image segmentation[J]. *IEEE Access*, 2021, 9: 16591-16603.
[10] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 3431-3440.
[11] QIN X B, ZHANG Z C, HUANG C Y, et al. U2-Net: Going deeper with nested U-structure for salient object detection[J]. *Pattern Recognition*, 2020, 106: 107404.
[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach: Curran Associates, Inc., 2017.
[13] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recongnition at scale[EB/OL]. (2020-09-04)[2024-11-20]. https://arxiv.org/abs/12010.11929.
[14] DAUBECHIES I. Ten lectures on wavelets[M]. [S. l.]: Society for Industrial and Applied Mathematics, 1992.
[15] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248-255.
[16] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. (2020-10-08)[2024-11-20]. https://arxiv.org/abs/2010.04159v4.
[17] HASSANI A, WALTON S, SHAH N, et al. Escaping the big data paradigm with compact transformers[EB/OL]. (2021-04-12)[2024-11-20]. https://arxiv.org/abs/2104.05704v4.
[18] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. New York: PMLR, 2021: 10347-10357.
[19] YUAN K, GUO S P, LIU Z W, et al. Incorporating convolution designs into visual transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021: 559-568.
[20] CHU X X, TIAN Z, ZHANG B, et al. Conditional positional encodings for vision transformers[EB/OL]. (2021-02-22)[2024-11-20]. https://arxiv.org/abs/2102.10882v3.
[21] GUO J Y, HAN K, WU H, et al. CMT: convolutional neural networks meet vision transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 12165-12175.
[22] FAN Q H, HUANG H B, CHEN M R, et al. RMT: retentive networks meet vision transformers[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 5641-5651.
[23] XIA C L, WANG X L, LV F, et al. ViT-CoMer: vision transformer with convolutional multi-scale feature interaction for dense predictions[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 5493-5502.
[24] WANG A, CHEN H, LIN Z J, et al. Rep ViT: revisiting mobile CNN from ViT perspective[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 15909-15920.
[25] LI F, PAN H S, SHENG S X, et al. Image retrieval based on vision transformer and masked learning[J]. *Journal of Donghua University (English Edition)*, 2023, 40(5): 539-547.
[26] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C/OL]//International Conference on Machine Learning. [S. l.]: PMLR, 2019: 6105-6114[2024-11-20]. https://proceedings.mlr.press/v97.
[27] TOUVRON H, CORD M, SABLAYROLLES A, et al. Going deeper with image transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 32-42.
[28] CHEN M Z, LIN M B, LI K, et al. CF-ViT: a general coarse-to-fine method for vision transformer[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37(6): 7042-7052.
[29] XU Y F, ZHANG Z J, ZHANG M D, et al. Evo-ViT: slow-fast token evolution for dynamic vision transformer[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(3): 2964-2972.
[30] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep

networks *via* gradient-based localization [C]// 2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 618-626.

一种用于视觉 Transformer 的小波变换与空间位置增强方法

胡润宇[1], 唐雪嵩[1,2*], 郝矿荣[1,2]

1. 东华大学 信息科学与技术学院, 上海 201620

2. 东华大学 数字化纺织服装技术教育部工程研究中心, 上海 201620

摘　要: 在视觉 transformer (vision transformer, ViT) 架构中, 图像数据需被转换为序列化数据以进行处理, 这可能导致图像的空间位置信息的丢失。虽然自注意力机制增强了模型捕获全局特征的能力, 但却在细粒度局部特征信息的保留上有所妥协。为了解决这些问题, 提出了一个基于 ViT 的包含空间位置增强模块和小波变换增强模块的模型。这些模块旨在减少补丁嵌入过程中空间位置信息的损失, 并提升模型的特征提取能力。空间位置增强模块通过卷积操作和多尺度特征提取来强化序列数据中的空间信息; 与此同时, 小波变换增强模块利用多尺度分析和频率分解来提升模型对图像全局和局部结构的理解能力。这种增强还提高了模型处理复杂结构和图像细节的能力。在 CIFAR-10、CIFAR-100 和 ImageNet-1k 数据集上, 将所提出的模型与先进方法进行了比较。结果表明, 所提出的模型在分类准确率上表现更优, 其有效性和优势得到了验证。

关键词: transformer; 小波变换; 图像分类; 计算机视觉