Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

A Wavelet Compression-Fused Vision Transformer Architecture for Meteorological Image Recognition

Mingyue Li[1], Yuanyuan Huang[1]*, Chengmao Wu[2], Lixin Zhao[1], Lijia Liu[1]

[1]Chengdu University of Information Technology, Chengdu, 610225, P.R. China
[2]Xichang University, Xichang 615013, Sichuan, China

DOI: https://doi.org/10.36347/sjet.2026.v14i05.001 | **Received:** 26.02.2026 | **Accepted:** 13.04.2026 | **Published:** 09.05.2026

***Corresponding author:** Yuanyuan Huang
Chengdu University of Information Technology, Chengdu, 610225, P.R. China

Abstract | **Original Research Article**

The current remote-sensing satellites and ground-based systems of observations generate meteorological images at a high spatial resolution and temporal frequency, and manual interpretation becomes less and less acceptable in real-time. Existing deep learning algorithms are too expensive in terms of computation and slow to be trained on such data. In order to address this, a vision transformer-based (WCViT-LL) architecture is created using low-frequency wavelet compression. Haar wavelet transform divides the input into low and high-frequency subbands, and the LL component (where most of the semantic information is found) is retained, which decreases the data volume and sequence length at the input stage. This tiny representation is then fed through an ordinary Vision Transformer (ViT) in order to obtain features, which can be utilized to categorize the meteorological images. Complexity analysis shows that the calculation of WCViT-LL of the self-attention is about 6.25% of that of a typical ViT, which leads to faster inference. This paper presents a solution to the real-time processing and edge deployment of high-resolution meteorological images in a concise and uncomplicated way.

Keywords: Wavelet transform, image compression; meteorological image recognition; lightweight model; Vision Transformer.

1. INTRODUCTION

Practically, the meteorological information measured by remote sensing and ground-based observation systems through satellites is currently of high quality, constantly updated, and multi-mode. The proper recognition of serious convective phenomena, centers of typhoons, and shapes of clouds on satellite images and radar echoes will become a significant problem of weather alerts, agricultural output, and aviation security in this situation. However, the traditional processes of manual interpretation cannot meet high levels of real-time weather prediction because of the sheer amount of high-resolution weather images generated daily, and are vulnerable to any errors.

The automated identification of meteorological elements through the application of deep learning is gaining significant interest in the framework of the current transformation to smart meteorology. It has been found that deep learning models are highly useful in feature extraction and classifying weather pictures. They can be improved to increase the recognition accuracy of the intricate weather phenomena [1]. Weather images, however, usually have large-scale weather systems structure, as well as fine local texture structure. Even though high-resolution input data is better than low-resolution input data in terms of recognition performance, it increases the computation and memory load of the model. The problem of managing the complexity of the computation without losing the ability to model the complex weather globally is a significant issue in the area of weather image recognition [2].

It is in this context that this paper evaluates the subject matter topic on compression of input data. The goal is to add an algorithm of wavelet compression in order to diminish the ViT input data without major meteorological information. This leads to a reduction in the computational complexity in the input stage, and real-time meteorological image recognition is possible with fewer computational resources.

2. RELATED WORK

Meteorological image recognition process has been gradually transferred to deep learning methods, as opposed to the conventional methods. Earlier studies

mainly relied on physical thresholds or handcrafted feature extraction, combined with algorithms such as SVM [3]. Even though these techniques were found to be effective in relatively simple cases, they were not good at generalization. They were found to be challenged by more difficult or faster-changing weather patterns. The introduction of deep learning has made CNNs the dominant approach and can be trained to find features automatically by using local receptive fields and weight-sharing. ResNet and other architectures are very precise in activities like image classification [4]. Nevertheless, the inductive bias inherent in CNNs constrains their ability to capture global contextual information, and thus long-term dependencies are difficult to learn, particularly in large-scale meteorological systems (e.g., typhoons).

ViT models have performed well on different visual tasks in the past few years. This is because they have a self-attention mechanism that enables them to model world dependency and hence circumvent the limitations of local receptive fields [5]. They work well under most circumstances, but images of high resolution are difficult to handle. This is largely due to the fact that the self-attention calculation is quadratic with the number of input tokens, which can easily cause bottlenecks in the computations in real-life situations. In order to reduce this problem, several lightweight variants have been suggested. In Swin Transformer [6] , the local window attention mechanism is utilized as an example, but MobileViT is resource-friendly and mobile [7]. These methods render things more efficient, though the majority of them are largely connected with changes in the interior design of the model. They are still fed with high-resolution images, but the multiscale character of the information is not used to the full.

The wavelet transform is different from model structure optimization and is widely used as a time-frequency analysis tool. It reduces data redundancy through multiscale decomposition, while important structural information is still kept. Because of this, it has been commonly used in image compression and denoising for a long time [8]. In recent years, some studies have started to combine wavelet transforms with deep neural networks. For example, wavelet convolutional neural networks introduce wavelet decomposition into CNNs to improve multi-scale feature extraction [9]. In transformer-based research, applying frequency-domain processing to the input is a noteworthy direction. This can reduce redundant information in the input. It also provides new insights for applying ViT to high-resolution image tasks.

3. METHODOLOGY

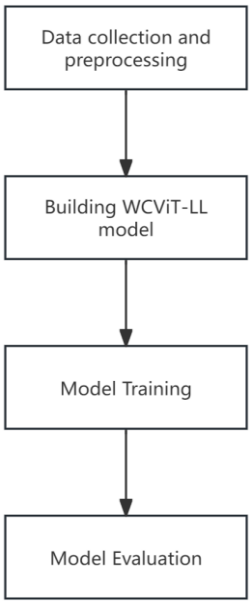

Figure 1: Method flowchart

3.1 Data Collection and Preprocessing

A meteorological image dataset containing six categories of weather phenomena was constructed in this study, including sunny, cloudy, rainy, foggy, snowy, and other, with a total of 183,798 images. We split the dataset into 70% for training and 30% for testing. Every image was resized to 224 × 224 pixels. To improve generalization performance, we applied several simple methods, such as random rotation, horizontal flipping, and color jittering during training. Finally, normalization was utilized to maintain stable training and accelerate convergence.

3.2 Building WCViT-LL model

The proposed WCViT-LL architecture is built upon the ViT framework, with the key design objective of reducing computational complexity through the introduction of a wavelet-based compression module. Figure 2 presents the overall architecture, which contains three main stages: wavelet compression, sequence serialization and embedding, and ViT-based feature learning.

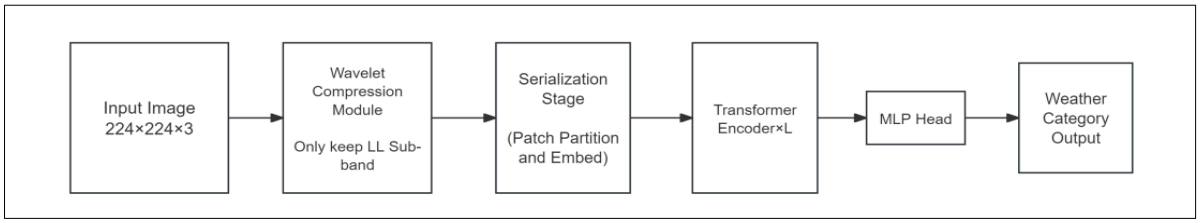

Figure 2: WCViT-LL Model Design Schematic

3.2.1 Low-Frequency Wavelet Compression Module

Given an input meteorological image $I \in R^{224\times224\times3}$, a single-level discrete wavelet transforms (DWT) based on the Haar wavelet is applied for decomposition. As shown in Figure 3, the two-dimensional DWT is implemented through separable convolution along the row and column directions, resulting in four sub-bands: LL, LH, HL, and HH, each with a spatial resolution of 112 × 112. According to multi-resolution analysis theory, the majority of signal energy and global structural information is concentrated in the low-frequency LL sub-band [10]. In contrast, high-frequency sub-bands primarily capture fine-grained texture details and are more susceptible to noise [11].

Therefore, WCViT-LL adopts a minimal low-frequency retention strategy, in which only the LL sub-band is preserved while the remaining three high-frequency sub-bands are discarded. The retained three-channel LL sub-band is subsequently concatenated to form a compressed feature map of size 112 × 112 × 3.

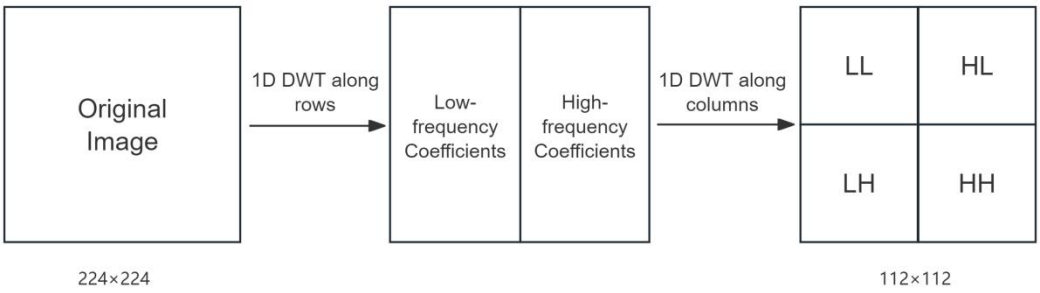

Figure 3: One-level Wavelet Decomposition

3.2.2 Sequence Serialization and Embedding

Figure 4 illustrates this step. This stage begins with the compressed 112 × 112 feature map. It is split into 49 separate 16×16 patches with no overlap. Each patch is then flattened and projected into a D-dimensional embedding vector via a linear projection layer. A learnable classification token([CLS]), is added to the front of the sequence as a classification marker. Meanwhile, learnable positional encodings are attached to every token to maintain spatial information across different patches. This process converts the compressed feature map into a sequential representation suitable for the ViT encoder.

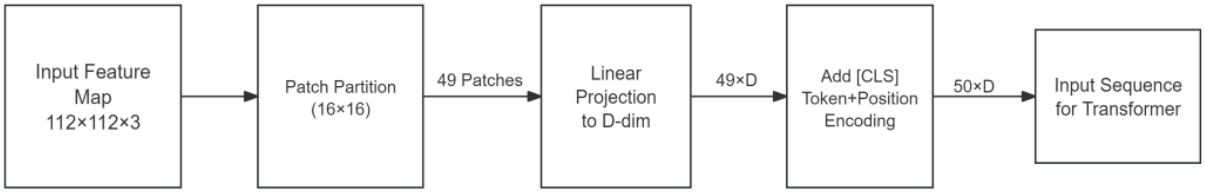

Figure 4: Patch Serialization and Embedding Process

3.2.3 ViT Encoder and Classification Head

The resulting token sequence is fed into the ViT encoder, which is composed of multiple stacked Transformer blocks. Each block consists of a multi-head self-attention (MSA) layer followed by a feed-forward network (FFN). Through the self-attention mechanism, the model is able to capture global dependencies among all image patches. After training, the output vector corresponding to the [CLS] token is passed to a multilayer perceptron (MLP) classification head, and the final prediction probabilities for the six weather categories are obtained via a Softmax function.

3.3 Model Training

The training procedure adopted the AdamW optimizer with a batch size of 16 and an initial learning rate of e^{-4}. Cross-entropy was selected as the loss function. To enhance training stability and convergence efficiency, a ReduceLROnPlateau learning rate scheduling strategy was employed. If the validation loss failed to improve for five consecutive epochs, the learning rate would be reduced by a factor of 0.1.

3.4 Model Evaluation and Computational Complexity Analysis

To comprehensively evaluate model performance, four metrics were adopted: Accuracy, Precision, Recall, and F1-score. In addition, ResNet50, ResNet101, Vision Transformer, and Swin Transformer were selected as baseline models for comparison under the same dataset and experimental settings.

From the perspective of computational efficiency, the wavelet-based compression reduces the

input image resolution from 224 × 224 to 112 × 112. When using the same patch size of 16 × 16, the sequence length of the standard ViT is 196, whereas that of WCViT-LL is reduced to 49. Based on the complexity analysis of the self-attention mechanism [12], the theoretical ratio of self-attention complexity between WCViT-LL and the standard ViT can be expressed as: $\frac{O(N_{wc}^2 D)}{O(N_{std}^2 D)} = \left(\frac{49}{196}\right)^2 = \left(\frac{1}{4}\right)^2 = \frac{1}{16}$. This indicates that the self-attention computational cost of WCViT-LL is theoretically reduced to 6.25% of that of the standard ViT, providing a theoretical basis for its potential advantage in inference efficiency.

4. RESULTS AND DISCUSSION

Table 1 summarizes the performance of all compared models on the six-class weather recognition task.

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
ResNet50	0.9091	0.9097	0.9091	0.9094
ResNet101	0.9095	0.9096	0.9095	0.9093
Vision Transformer	0.9107	0.9117	0.9107	0.9110
Swin Transformer	0.9133	0.9136	0.9133	0.9132
WCViT-LL	0.9093	0.9095	0.9093	0.9094

As demonstrated by the results, the proposed WCViT-LL model achieves recognition accuracy comparable to that of the classical ResNet series, although it remains slightly lower than the standard ViT and Swin Transformer. This phenomenon can be attributed to the wavelet compression strategy introduced at the input stage, which selectively retains only the low-frequency (LL) sub-band as the primary feature. While this strategy reduces the input data volume by 75%, it inevitably discards certain high-frequency detail information that may play an auxiliary role in distinguishing weather categories dependent on local fine-grained textures, thereby leading to a certain degree of decrease in precision.

Nevertheless, the core of meteorological recognition tasks relies heavily on large-scale structural features, such as cloud system extent and brightness distribution, which are predominantly encapsulated within the low-frequency components of the imagery. Consequently, by preserving core semantic information, WCViT-LL maintains performance metrics without significant degradation, validating the efficacy and rationality of the low-frequency compression strategy in meteorological scenarios. More importantly, WCViT-LL yields a substantial leap in computational efficiency; theoretical analysis indicates that the computational complexity of its self-attention mechanism is merely 6.25% of the standard ViT. This improvement implies that the model gains potential order-of-magnitude advantages in terms of inference speed and memory occupancy, enabling WCViT-LL to better meet the stringent requirements for real-time meteorological monitoring, edge-side deployment, or the processing of high-resolution images.

5. CONCLUSION

In this article, we propose a lightweight model called WCViT-LL, which incorporates a mechanism for compressing low frequencies using wavelets to reduce the high computational complexity of ViT in high-resolution weather image recognition. By introducing a wavelet transformation at the input stage and retaining only the subbands to approximate the low frequencies, this model effectively reduces the input dimension and the computational complexity of self-attention while preserving the essential global structural information of the weather images. According to the experimental results, it can be observed that the recognition accuracy of WCViT-LL is the same as that of popular convolutional neural network models. Despite its performance, which is worse than that of the ViT and the Swin Transformer, WCViT-LL exhibits evident benefits in terms of computational efficiency. The proposed method can be viewed as a possible and practical substitute for meteorological recognition tasks. Future research may also take into account the multiscale wavelet feature fusion and learnable wavelet basis functions to improve the capabilities of the model.

Acknowledgement

This work is supported in part by Key Research and Development Projects of Liangshan Science and Technology Program (No.23ZDYF0121), in part by the Undergraduate Teaching Research and Reform Project of Chengdu University of Information Technology (No. JYJG2025035) and in part by College Students' innovation project of Chengdu University of Information Technology（No.X202510621158）.

REFERENCES

1. Liu Y, Racah E, Correa J, Khosrowshahi A, Lavers D, Kunkel K, *et al.,* Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv preprint arXiv:160501156. 2016.
2. Shi X, Gao Z, Lausen L, Wang H, Yeung D-Y, Wong W-k, *et al.,* Deep learning for precipitation nowcasting: A benchmark and a new model. Advances in neural information processing systems. 2017;30.
3. Luo Q, Meng Y, Liu L, Zhao X, Zhou Z. Cloud classification of ground-based infrared images combining manifold and texture features.

Atmospheric Measurement Techniques. 2018;11(9):5351-61.
4. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
5. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.
6. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, *et al.,* editors. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision; 2021.
7. Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:211002178. 2021.
8. Skodras A, Christopoulos C, Ebrahimi T. The JPEG 2000 still image compression standard. IEEE Signal processing magazine. 2002;18(5):36-58.
9. Fujieda S, Takayama K, Hachisuka T. Wavelet convolutional neural networks. arXiv preprint arXiv:180508620. 2018.
10. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE transactions on pattern analysis and machine intelligence. 2002;11(7):674-93.
11. Donoho DL. De-noising by soft-thresholding. IEEE transactions on information theory. 1995;41(3):613-27.
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al.,* Attention is all you need. Advances in neural information processing systems. 2017;30.