Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30957

Visual Transformer Large Models: Basic Architecture, Structural Improvements and Applications

Xingwei Song

School of Information Engineering, Henan University of Science and Technology, Luoyang, China
3051674256@qq.com

Abstract. With the fast development of artificial intelligence, computer vision has demanded better global information capture from models. Traditional CNNs are increasingly limited in long-range dependency modeling, and Vision Transformers (ViT) have offered a new approach for advances in this field, drawing great attention to their structural optimization and practical applications. This paper uses literature analysis to focus on the structural improvement and application of large ViT models. It first analyzes ViT's core structure and self-attention mechanism based on basic architecture and theories, compares the differences between ViT and CNN in feature extraction and long-range dependency modeling, and points out ViT's strengths in global information capture as well as its high computational complexity. Then, it explores key progress in structural improvements, including multi-scale feature modeling optimization and interpretability enhancement. By designing hybrid architectures and introducing local inductive biases, the model's performance and generalization ability are effectively improved. In addition, combined with typical scenarios, it elaborates on ViT's performance in traditional visual tasks and verifies its cross-domain transfer capacity and practical value in complex situations through cases like remote sensing image fusion.

Keywords: Vision Transformer, Structural Improvement, Multi-Scale Feature, Interpretability, Cross-Domain Application

1. Introduction

In recent years, ViT has achieved remarkable breakthroughs in the field of computer vision, marking a critical shift from the long-dominant convolutional neural networks (CNNs). Endowed with robust global modeling capabilities driven by self-attention mechanisms, ViT has demonstrated performance comparable to or even superior to traditional CNNs across various core visual tasks, including image classification, object detection, and semantic segmentation. Unlike CNNs that rely on local receptive fields and struggle to capture long-range contextual dependencies effectively, ViT partitions images into fixed-size patch sequences, converts these patches into feature embeddings, and leverages self-attention to model intricate global correlations between different image regions, thereby overcoming the inherent limitations of CNNs in long-distance information interaction. Despite its prominent advantages, ViT still faces notable challenges, such as high computational complexity and heavy reliance on large-scale labeled datasets. These issues hinder its efficient

DOI: 10.54254/2755-2721/2026.TJ30957

deployment on resource-constrained edge devices and restrict its widespread practical application in scenarios with limited data. To address these bottlenecks and further unlock the potential of ViT, this paper conducts in-depth research focusing on its basic architecture, structural optimization strategies, and cross-domain application exploration. By analyzing the core mechanism of ViT, exploring feasible improvement approaches to reduce computational costs and enhance data efficiency, and investigating its adaptive performance in diverse fields, this study aims to provide valuable insights for promoting the efficient and extensive application of Vision Transformers in more practical scenarios.

2. Basic architecture and theory of vision transformer

In the development of computer vision and artificial intelligence, image processing technology has always played a crucial role. By enabling computers to analyze, process and interpret digital images, it provides a foundation for advanced visual tasks such as image classification, object detection and semantic segmentation. In recent years, with the continuous advancement of deep learning technology, CNNs and ViTs have become the two mainstream architectures in the field of image processing.

CNNs have shown excellent performance in image feature extraction and pattern recognition due to their characteristics of local perception and weight sharing. As illustrated in Figure 1, since the proposal of ResNet in 2015, CNNs have achieved breakthrough progress in the visual field. Subsequently, models such as EfficientNet and MobileNet continuously optimized network structures in May 2020, improving model performance and efficiency, which further consolidated the position of CNNs in the visual field.

In October 2020, the emergence of ViT brought new ideas for visual modeling. It abandons the convolution operation of CNNs and adopts a self-attention mechanism to directly conduct global modeling on images, demonstrating strong performance on large-scale datasets. Since then, the proposal of Swin-Transformer in March 2021 and the development of subsequent models have continuously promoted the application and improvement of Transformer architecture in the visual field.

To give full play to the respective advantages of CNNs and ViTs while making up for their respective shortcomings in the field of image processing, researchers have been actively exploring and proposing hybrid models that combine the two. The research and innovation of such models have been continuously deepened and widely applied in multiple research fields, thus opening a new chapter of model fusion.

Next, this paper will elaborate on the working principle of Vision Transformer and its differences from CNN.

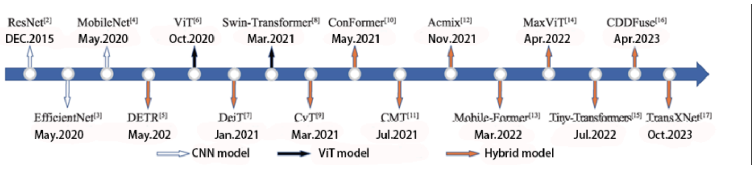

Figure 1. Typical representative works and development history of image processing models [1]

2.1 Analysis of the core structure of ViT

As shown in Figure 2, the core workflow of ViT can be divided into four key steps:

a. Image Patching and Embedding

DOI: 10.54254/2755-2721/2026.TJ30957

The original image is first evenly divided into several fixed-size image patches. Each patch is converted into a one-dimensional feature vector (Patch Embedding) through a linear projection layer, realizing the vectorized representation of local image features.

b. Introduction of Positional Encoding and [class] Token

To retain spatial position information, each Patch Embedding is superimposed with learnable Positional Encoding. Meanwhile, a learnable [class] Token is introduced to aggregate global features and participate in subsequent classification tasks.

c. Feature Extraction by Transformer Encoder

The sequence composed of Patch Embeddings, Positional Encoding and [class] Token is input into the Transformer encoder. The encoder consists of multiple identical modules, and each module includes Multi-Head Self-Attention and Multi-Layer Perceptron (MLP) sub-layers. Residual connections and layer normalization are adopted to enhance information transmission and training stability.

d. Classification and Prediction

Finally, the output of the [class] Token processed by the encoder is extracted and input into the MLP classifier to complete the prediction of image categories.

With the above design, ViT not only makes full use of the global modeling capability of Transformer, but also retains image spatial information through patch segmentation and positional encoding, thus demonstrating excellent performance on large-scale datasets.

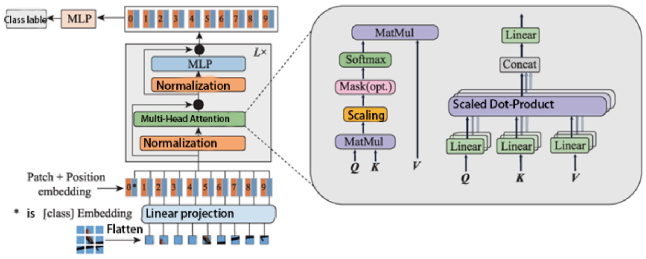

Figure 2. Overall architecture diagram of vision transformer [1]

2.2 Feature comparison between ViT and CNN

ViT and traditional CNNs exhibit significant differences in feature extraction and global modeling capabilities. ViT divides input images into equal-sized regions via an image patching strategy, transforms them into feature sequences through linear operations, and introduces positional encoding to preserve spatial topological information, thereby effectively establishing spatial correlations between image regions. In contrast, CNNs rely on local convolutional kernels for feature capture, with their receptive fields constrained by kernel size and network depth, making it difficult to directly model semantic connections between long-distance pixels.

In terms of feature extraction methods, the multi-head self-attention mechanism of ViT can adaptively integrate multi-scale visual information, breaking the limitations of fixed receptive fields in CNNs. Multiple parallel attention heads focus on different image regions in various feature subspaces, enabling simultaneous capture of diverse visual features and generation of richer, more accurate image representations. This characteristic performs particularly prominently in tasks requiring global understanding such as medical imaging; for instance, in complex lesion diagnosis, the diagnostic accuracy of ViT is significantly higher than that of traditional CNN models [2].

DOI: 10.54254/2755-2721/2026.TJ30957

In addition, the modular design of ViT supports flexible adjustment of patch size or the number of encoder layers, allowing it to adapt to medical imaging data of different resolutions and types (e.g., X-ray, CT, MRI) [2]. However, constrained by fixed-size convolutional kernels, CNNs need to rely on complex feature pyramid structures or multi-scale fusion modules when processing multi-scale data.

In summary, ViT has obvious advantages in global context modeling, multi-scale feature integration and architectural flexibility, providing a new technical path for image understanding tasks.

3. Research progress on vision transformer structural improvement technologies

In the development of Vision Transformers, structural improvement technologies have always been the key driving force for enhancing their performance. With in-depth research, researchers have continuously explored and innovated in aspects such as multi-scale feature modeling and interpretability enhancement, enabling Vision Transformers to continuously optimize their performance in complex visual tasks. Below, this paper will specifically elaborate on some of these optimization methods.

3.1 Optimization of multi-scale feature modeling

In visual tasks, multi-scale features can capture both local details and global structures simultaneously, which is of great significance for improving model performance. To fully utilize information of different scales, researchers have proposed various multi-scale feature modeling methods. For example, the paper Transformer-Based Multi-scale Optimization Network for Low-Light Image Enhancement [3] presents a Transformer-based multi-scale optimized low-light image enhancement network (TMO, as shown in Figure 3). This network effectively integrates brightness, color and detail information of different scales through a multi-scale feature fusion module (MFFM) and an adaptive enhancement mechanism.

First, MFFM uses a pre-trained ResNet backbone to extract shallow features of four scales, unifies the number of channels via 1×1 convolution, and realizes scale transformation through upsampling or convolution with a stride of 2. Then, it aggregates features of all scales by dense fusion to form fused features with multi-scale receptive fields. In the global branch, a Transformer-based multi-task enhancement module (TMEM) is designed, which uses its global modeling capability to initially address issues such as insufficient brightness and color deviation. In the local multi-scale branch, an adaptive enhancement module is adopted to perform layer-by-layer enhancement from small to large scales. Combined with a two-path parallel attention mechanism (TPAM) and a local color correction structure (LCCS), it dynamically adjusts enhancement strategies and suppresses noise.

Experiments show that this method significantly improves the enhancement quality of low-light images on multiple public datasets, verifying the effectiveness of multi-scale feature modeling optimization. This method provides an effective reference for the application of Vision Transformers in complex scenarios and offers new ideas for subsequent research.

DOI: 10.54254/2755-2721/2026.TJ30957

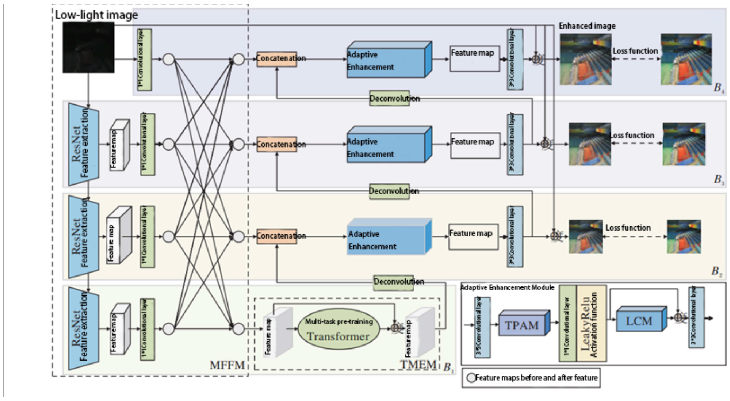

Figure 3. TMO structure diagram[3]

3.2 Enhancement and improvement of interpretability

The interpretability of deep learning models is crucial for understanding their decision-making mechanisms. According to the focus of algorithms, interpretability methods can be divided into global and local interpretability algorithms [4]. Global interpretability focuses on the overall behavior of the model on the complete dataset, explaining the conditional interactions between dependent variables and independent features; local interpretability explains specific samples or decision paths.

In addition, methods can also be divided into active and passive interpretation algorithms based on whether the model training process is changed [4]. Passive interpretation algorithms extract understandable patterns based on trained models, while active interpretation algorithms improve interpretability by modifying network structures or training processes before training.

This classification provides a theoretical basis for enhancing the interpretability of Vision Transformers and helps design more effective interpretation methods.

4. Typical application scenarios of vision transformer

4.1 Traditional visual tasks

In image classification, a classic traditional visual task, ViT demonstrates remarkable advantages in long-range dependency modeling due to its global attention mechanism. However, the original ViT is prone to overfitting when data is insufficient and has high computational complexity. To address this, researchers have proposed a multi-scale focal attention module, which dynamically adjusts the attention window size to suppress irrelevant long-distance information while retaining key local features. Experiments show that this method achieves a classification accuracy of 98.32% on the CIFAR-10 dataset, 12% higher than the baseline ViT, and significantly reduces computational resource consumption [5]. This improvement not only enhances the model's ability to capture local details but also boosts generalization performance on small datasets, providing new insights for efficient feature extraction in traditional visual tasks.

4.2 Cross-domain applications

In cross-domain applications, researchers have proposed a fusion method of dual-branch generative adversarial networks and Transformers, successfully transferring deep learning technologies in the computer vision field to remote sensing image fusion tasks [6]. This method decomposes

DOI: 10.54254/2755-2721/2026.TJ30957

multispectral and panchromatic images through guided filtering to extract global spectral information and local texture features respectively, and uses Transformers and convolutional neural networks for feature fusion [6]. Experimental results indicate that while preserving the rich spectral information of multispectral images, this method significantly improves the spatial resolution of fused images and outperforms existing methods in quantitative evaluation metrics (such as PSNR and SSIM), offering an effective cross-domain solution for remote sensing image fusion, an emerging visual task.

5. Conclusion

This paper conducts a systematic study on the basic architecture, structural improvements and applications of Vision Transformer large models. From the analysis of basic architecture to feature comparison, as well as the optimization of multi-scale feature modeling and enhancement of interpretability, it reveals the advantages and limitations of Vision Transformer in traditional visual tasks. In cross-domain applications, successful practices in tasks such as remote sensing image fusion have further verified the potential of this model. Future research can focus on enhancing model interpretability and adaptive optimization in dynamic environments to promote the in-depth application and industrial implementation of Vision Transformer in more fields.

References

[1] Guo Jialin, Zhi Min, Yin Yanjun, Ge Xiangwei. Review of Research on CNN and Visual Transformer Hybrid Models in Image Processing [J]. College of Computer Science and Technology, 2025, 19(01): 30-44.

[2] Wang Yijia. Advantages and Challenges of the Vision Transformer Model in Medical Imaging [J]. Information & Computer, 2025, 37(12): 16-19.

[3] Niu Yuzhen, LIN Xiaofeng, XU Huangbiao, LI Yuezhou, CHEN Yuzhong. Transformer-Based Multi-scale Optimization Network for Low-Light Image Enhancement, 2023, 36(06): 511-529. DOI: 10.16451/j.cnki.issn1003-6059.202306003.

[4] Yan Xin. Research on Interpretability Algorithm Based on Visual Transformer Architecture [D]. Harbin Engineering University, 2023. DOI: 10.27060/d.cnki.ghbcu.2023.001992.

[5] Bin Chen.Research on Data Efficient Vision Transformer Network [D]. Chongqing University of Technology, 2023.DOI: 10.27753/d.cnki.gcqgx.2023.002052.

[6] Ji Y.X., Kang J.Y.and Ma H.Y. Fusion of panchromatic and multispectral remote sensing images using a dualbranch generative adversarial network combined with Transformer [J]. National Remote Sensing Bulletin, 29(8): 2641-2657 [DOI:10.11834/jrs.20244047]