Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31146

A Review of Hybrid Models Combining Convolutional Neural Networks and Vision Transformers in Medical Image Processing

Chade Li

The University of Washington, Seattle, USA
cl0616@uw.edu

Abstract. Medical image processing is a very important role in modern healthcare diagnosis and treatment. However, traditional manual analysis faces challenges like high variances, low efficiencies and low accuracies. Recently, deep learning techniques like Convolutional Neural Networks (CNNs) have rapidly improved and achieved remarkable success and improvements in areas like medical image classification, segmentation, and detection tasks due to their powerful feature extraction capabilities. Nevertheless, CNNs exhibit limitations in modeling global contextual information and rely heavily on large-scale annotated datasets. The emergence of Vision Transformers (ViTs) offers a new perspective by effectively modeling global image features through self-attention mechanisms. Hybrid models that combine the strengths of CNNs and Transformers have thus become a research hotspot. This paper aims to make reviews for fusion methods between CNN and Transformers in medical image processing, including typical strategies such as early fusion, intermediate fusion, and late fusion, and summarizes their application performance and advantages in various tasks. Experimental results show that hybrid models are able to show better performance than areas like single-architecture models in terms of accuracy, generalization ability, and adaptability to complex tasks. Finally, this paper discusses the future challenges faced by hybrid models in terms of data scarcity, computational efficiency, and interpretability, and outlines future research directions.

Keywords: Medical Imaging, CNN, ViT, Hybrid Model, Fusion Strategy

1. Introduction

1.1. Research background

Medical image processing is an irreplaceable thing in the modern society. As the medical technologies gradually became more advanced, technologies like X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and ultrasound, doctors can gain easier and better access to patient's internal structure, providing solid foundations for all areas like disease detection, precise diagnosis, and treatments. However, opportunities rise with the coexistence of costs. As technology improves rapidly, the cost to produce better and more precise medical imaging data also grows.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ31146

Tradition means face challenges like high time costs and the accuracy is highly correlated with the doctors' sole experience, making it demonstrates high variances and standard deviations among different doctors, and unable to fulfill clinical demands for efficient and accurate diagnoses. However, the rise of deep learning technologies has significantly improved the lower limit of accuracy and promoted efficiency of medical image processing through automated feature extraction.

1.2. Importance of medical image processing

Medical image processing plays a very important role in various fields of medical means like early lesion detection, organ segmentation, and disease diagnosis. In the past, traditional diagnoses were highly dependent on doctors' expertise and experience, which can be susceptible to subjective factors, leading to highly differed results among different doctors. However, the emergence of deep learning is capable of automating these tasks, which could significantly enhance diagnostic accuracy and consistency. To be specific, deep learning models can precisely produce segment pulmonary lesion region results from CT images, and thereby aid doctors in the timely detection of lung cancer and provide earlier intervention and treatment for patients.

1.3. Current applications of CNNs and transformers in medical imaging

The emergence of CNN has created opportunities for improvements in medical image classification, segmentation, and detection tasks. For example, according to Litjens et al. , the authors summarized that there are wide applications of CNNs in X-ray, MRI, and CT images, demonstrating their efficiency in tumor detection and organ segmentation [2]. Still, however, CNNs are limited in global contextual modeling. Recently, Vision Transformers (ViTs), inspired by their success in natural language processing, have been introduced into computer vision and shown potential in medical image processing. For instance, Chen et al. proposed a Transformer-based model for medical image segmentation, significantly improving segmentation accuracy [4]. Moreover, hybrid models that combine the strengths of both CNNs and Transformers have been proposed, such as the CNN-Transformer hybrid architecture by Hatamizadeh et al., which demonstrated superior performance in 3D medical image processing [5].

1.4. Review objectives and structure

This paper discusses the different hybrid models combining CNNs and Vision Transformers in medical image processing. We will introduce the basic principles and current applications of CNNs and Transformers in medical image analysis, then a summary of typical design philosophies and classification methods for hybrid models, including early fusion, intermediate fusion, and late fusion strategies. We also will analyze the performance advantages of hybrid models over single-architecture models in application scenarios such as tumor segmentation, lesion detection, and organ recognition, conducting a comprehensive comparison from the perspectives of model accuracy, generalization ability, and computational complexity. Finally, we will discuss the potential challenges such as data scarcity, interpretability, and clinical implementation, proposing future development trends and research directions.

DOI: 10.54254/2755-2721/2026.TJ31146

2. CNN overview

2.1. Basic structure of CNN

A typical CNN model consists of an input layer, convolutional layers, pooling layers, fully connected layers, and an output layer. The convolutional layer, the cornerstone of CNN, extracts local features by sliding a set of learnable convolutional kernels (or filters) over the input image or feature map. Nonlinear activation functions (e.g., ReLU) are typically introduced after convolutional operations to enhance the model's expressive capacity. Pooling layers follow convolutional layers to reduce the spatial dimensionality of feature maps, achieving translation invariance, reducing computational load, and enhancing feature robustness while effectively controlling overfitting.

2.2. Applications of CNN in medical image processing

CNN have become the mainstream method for core tasks in medical imaging, including classification, segmentation, and detection. Typical applications include automatically extracting lesion region features using convolutional structures for pulmonary nodule detection and segmentation; identifying brain tumor boundaries in MRI images to assist in treatment planning; and achieving near-expert-level classification accuracy in skin lesion recognition. Litjens et al.'s review systematically summarized the applications of deep learning technologies in medical image analysis, indicating that CNN have outperformed traditional machine learning methods in multiple tasks. Improved network architectures tailored for medical image segmentation, such as U-Net, have further enhanced the applicability and accuracy of models in medical imaging data.

3. Transformer model overview

3.1. Basic structure and principles of transformers

Transformers, initially proposed by Vaswani et al. in 2017 for sequence-to-sequence (Seq2Seq) tasks in natural language processing, innovatively abandon traditional recurrent and convolutional structures, relying solely on self-attention mechanisms to model global dependencies among sequence elements. Their basic structure includes multi-head self-attention layers, feedforward fully connected networks, layer normalization, and residual connections. Self-attention mechanisms capture long-range dependencies by computing relationships among elements in the input sequence; multi-head attention enhances the model's ability to focus on different features; and positional encodings provide positional information to the sequence.

3.2. Applications of transformers in medical image processing

The application of Transformers in image processing has gradually emerged, showing exceptional performance in medical image processing. For instance, Chen et al. used Transformers for MRI image segmentation, achieving significant improvements in brain tumor segmentation tasks. Transformers enhance the processing of complex medical images through their global modeling capabilities.

DOI: 10.54254/2755-2721/2026.TJ31146

3.3. Advantages, disadvantages, and improvement methods of transformers

Firstly, Transformers are capable of capturing long-range dependencies in images and dealing with global information and parallel computing.This helps avoiding the locality constraints of convolutions and this nature makes it suitable for complex visual tasks. Secondly, the high parallel computing efficiency enables full utilization of GPU parallelism, accelerating large-scale model training. Finally, it exhibits good scalability, with model capacity significantly improving by increasing depth and width to adapt to complex medical image tasks. As for the disadvantages, however, Transformers have a strong dependency on large-scale datasets and have poor performance on small datasets. Additionally, the high computational complexity leads to long training and inference times.

4. CNN and transformer hybrid models

The hybrid model of CNN and Vision transformer is not sufficient to rely solely on a simple technology stack. In fact, this must be designed based on the concept of complementary advantages. This design aims to maximize and combine the strengths of the two models as much as possible, while minimizing the inclusion of their weaknesses. Their core design motivation is to fully combine the inductive bias and computational efficiency of CNN in local feature extraction, while using the global self-focus mechanism of the transformer to establish remote dependencies, thereby achieving comprehensive modeling of medical images from subtle signs to a bigger image.

4.1. Design philosophies of hybrid models

Firstly, CNN serves as the feature extraction frontend, and Transformers as the contextual modeling backend. This is the most prevalent paradigm, where CNN backbone networks (e.g., ResNet) downsamples the original image to produce local feature maps, which are then reshaped into sequences and then inputted into Transformer encoders to fuse global information through self-attention mechanisms. Secondly, Transformers and CNN process input images in parallel, followed by the process of feature fusion. The CNN branch and Transformer branch then processes the input image separately, and their outputs are integrated through a consecutive process of feature addition, concatenation, or attention fusion, preserving both detailed features and global context. Finally, Transformers plays the role of the core, with convolutions strategically embedded to enhance locality. Convolutional operations are strategically embedded within Transformer modules, such as using convolutional patch embedding, incorporating depthwise separable convolutions into feedforward networks (FFN), or adopting convolutional positional encodings, to inject locality priors into the model and improve performance under low-data regimes.

4.2. Advantages of combining CNN and transformer

Compared with a single model, the hybrid architecture demonstrates very significant advantages through synergy effects: Firstly, they fully leverage the strengths of each of the two models, achieving an effective unification of local and global information. This is because CNN excels at capturing local features such as the edges and textures of lesions, while Transformers are good at modeling semantic associations between different anatomical structures or lesion regions. Therefore, both models have brought into play their respective advantages, enabling the hybrid model to achieve an organic unity of detail perception and global understanding. Secondly, they balanced the model performance and computational efficiency. In comparison with the pure Transformer model,

DOI: 10.54254/2755-2721/2026.TJ31146

the hybrid model significantly reduces the sequence length of the Transformer input through CNN-based pre-dimension reduction, effectively lowering the computational and memory overhead, making it more suitable for processing high-resolution medical images.

4.3. Architectures and improvement methods of hybrid models

Researchers have proposed various innovative hybrid architectures, with the following being several typical representatives:

(1) TransUNet [4]: A paradigm of encoder-decoder hybrid architecture. Its encoder adopts a serial structure of "CNN (ResNet) + Transformer," first using CNN to extract high-dimensional feature maps, which are then converted into sequences and input into a Transformer to capture global context. The decoder employs CNN-based upsampling layers and fuses multi-scale local features from the encoder through skip connections, achieving precise pixel-level segmentation.

(2) Convolutional Vision Transformer (CvT) [6]: Enhances efficiency and performance by introducing convolutions into two key stages of ViT: Firstly, convolutional projection is used to generate patch embeddings; secondly, convolutional self-attention is adopted in the self-attention module, thereby obtaining a global receptive field while maintaining the locality advantages of convolutions.

(3) CoTr [7]: A hybrid model focusing on medical image segmentation. It uses a CNN encoder to extract features and designs a DeTrans-Encoder as the decoder, where deformable attention is employed to focus on regions most relevant to the target organ or lesion, significantly reducing the computational complexity of Transformers and improving modeling capabilities for targets of varying shapes and sizes.

5. Hybrid models in medical image processing

5.1. Lesion detection and segmentation

Lesion detection and segmentation are core tasks in medical image analysis, requiring models to accurately localize and delineate lesion regions. Hybrid models exhibit remarkable advantages in this domain in comparison with old times. CNN backbone networks (such as the encoder of U-Net) can efficiently extract local texture and edge features of tumors, bleeding points, or injury regions, and generate high-quality multi-scale feature maps. Then, followed by the Transformer module, globally inferring semantic associations between different image regions and lesions through self-attention mechanisms. This effectively addresses past issues of missed detections and discontinuous segmentations caused by irregular lesion shapes, dispersed distributions, or low contrast with surrounding tissues [8].

5.2. Organ localization and classification

Organ segmentation is an important field in radiotherapy planning and surgical navigation. In the past, due to the inter-individual variability, this field has posed challenges to doctors. However, with the hybrid models, we achieve better organ recognition by combining CNN's local feature extraction capabilities with Transformer's global context modeling. The CNN branch ensures the capture of local anatomical structure features such as the liver, kidneys, or prostate, while the Transformer branch comprehends spatial relative position relationships among various organs within the cavities, enabling localization and segmentation to be precise even under conditions of blurred organ boundaries or partial occlusions [9].

DOI: 10.54254/2755-2721/2026.TJ31146

5.3. Medical imaging diagnostic support

Disease classification tasks require models to derive diagnostic classification results from whole images. Hybrid models provide stronger feature representations for this task. The CNN frontend extracts rich local signs (such as ground-glass opacities in lung images or microaneurysms in retinal images) from the images, which are then input into the Transformer. Through self-attention mechanisms, the Transformer can assess the relative importance of all local signs for the final diagnosis and fuse these dispersed pieces of evidence to make global decisions [10].

6. Performance evaluation and comparison of hybrid models

6.1. Common evaluation metrics for medical image processing tasks

In order to measure the performance of hybrid models, researchers have employed a series of domain-recognized quantitative metrics:

(1) Classification Tasks: Accuracy, precision, recall, specificity, and comprehensive metrics such as the F1-Score and the Area Under the Receiver Operating Characteristic Curve (AUROC) are primarily used. AUROC is particularly crucial as it comprehensively evaluates model performance across different diagnostic thresholds [8].

(2) Segmentation Tasks: The Dice Similarity Coefficient (Dice Score) and Intersection over Union (IoU) are the most widely used metrics to quantify the overlap between predicted and ground truth segmentation regions. Additionally, the Hausdorff Distance (HD) assesses the maximum deviation of segmentation boundaries, which is critical for clinical safety evaluations.

(3) Detection Tasks: Metrics such as Average Precision (AP) are commonly used, with recall and precision calculated at specific IoU thresholds.

6.2. Performance comparison of models

Numerous empirical studies have already shown that hybrid models are capable of achieving performance improvements in most medical image processing tasks. Hybrid models exhibit significant advantages in tasks requiring global context understanding. For instance, when segmenting organs with complex structures or blurred boundaries (such as the pancreas) and detecting dispersed multiple lesions, their Dice coefficients and IoUs are typically significantly higher than those of pure CNN models. CNN models' inherent local receptive field limitations hinder their ability to effectively model such long-range dependencies. Pure Transformer models (such as ViT) are outperformed by hybrid models in terms of data efficiency and computational efficiency. Pure Transformer models require pre-training on massive datasets to achieve optimal performance, which is a major obstacle in the data-scarce medical field. Hybrid models, with CNNs as feature extractors, provide strong inductive biases, enabling rapid convergence and excellent performance on medium-sized datasets while significantly reducing computational and memory overhead [8].

6.3. Performance and efficiency trade-offs of different hybrid strategies

Different hybrid strategies exhibit varying characteristics in terms of performance and efficiency:

(1) Serial Structures (CNN → Transformer): Take models like TransUNet as an example. It typically achieves state-of-the-art performance. However, it trade-offs high computational costs due to the introduction of Transformer encoders [8].

DOI: 10.54254/2755-2721/2026.TJ31146

(2) Parallel Structures (CNN + Transformer): These kind of models combines the advantages of both architectures through feature fusion, maintaining high performance while often offering relatively flexible design options and acceptable computational costs [8].

(3) Internally Embedded Convolutional Transformers: Models such as CvT aim to enhance the efficiency and locality of the original Transformer, typically performing better in terms of parameter efficiency and inference speed, providing directions for practical applications [8].

7. Challenges and future development directions

Medical image analysis is kept facing severe data challenges, which are particularly pronounced for hybrid models that are heavily dependent on data training. Firstly, high-quality medical image annotations depend heavily on the experience, expertise and time of professional physicians, resulting in scarce and high cost. Although the emergence of semi-supervised and self-supervised learning provides opportunity for partial solutions, significantly reducing reliance on fine-grained annotations is still currently a core challenge. Secondly, medical images originate from diverse sources, including different equipment manufacturers, acquisition protocols, and medical institutions, leading to distinct significant data distribution differences (i.e., domain shifts). This creates more challenges for the generalization capabilities of models. Therefore, the construction of robust models insensitive to domain shifts is a crucial requirement and prerequisite for advancing such clinical practical applications. Furthermore, on the model level, the main challenges lie in computational complexity and interpretability. Although hybrid models have reduced some computational overhead compared to pure ViT models, their overall complexity remains higher than traditional CNNs, especially due to the high computational resource requirements of Transformer modules, limiting their deployment in resource-constrained clinical environments. Model lightweighting, knowledge distillation, and efficient attention mechanisms are potential optimization directions. Simultaneously, clinical applications demand transparent and trustworthy decision-making processes. Although hybrid models can provide some insights through attention maps, their overall decision logic remains insufficiently intuitive. By providing interpretability aligned with clinical cognition (e.g., based on anatomical structures or imaging signs), physicians build trust in AI and facilitate its integration into actual workflows. Clinical integration also presents multiple challenges. Most current research focuses on improving model performance but failed to realize the importance of having seamless integration with hospital information systems (such as HIS/PACS), real-time inference efficiency, and human-computer interaction interface design. Truly "user-friendly" diagnostic support systems must align with actual clinical workflows and clearly present analysis results. Additionally, medical AI products require rigorous regulatory approvals (such as NMPA, FDA), and the "black-box" nature of hybrid models poses additional obstacles for approvals. Simultaneously, potential biases that exist in model decisions, data privacy protection, and medical liability determination thereby raise ethical and legal issues that require in-depth exploration and resolution.

Based on the aforementioned challenges, future research is expected to achieve breakthroughs in the following directions:

(1) Data-Efficient Learning: Explore more advanced semi-supervised, self-supervised, and weakly supervised learning paradigms to fully utilize vast amounts of unannotated data and minimize reliance on expensive annotations.

(2) Model Lightweighting and Efficiency Optimization: Continuously research more efficient attention mechanisms (such as linear attention), dynamic computation (allocating computational

DOI: 10.54254/2755-2721/2026.TJ31146

resources based on input difficulty), and neural architecture search to "slim down" hybrid models and enable edge deployment.

(3) Domain Generalization and Adaptation: Develop domain generalization and test-time adaptation techniques to enable models to adapt to data distributions from new hospitals or equipment without retraining, enhancing their clinical robustness.

8. Conclusion

This paper systematically reviews the research progress of hybrid models combining CNNs and ViTs in medical image processing. CNNs excel at local feature extraction but have limited global modeling capabilities; ViTs possess strong long-range dependency modeling capabilities but suffer from issues such as high data demands, computational complexity, and lack of inductive biases. Hybrid models, by combining the advantages of both through serial, parallel, or embedded architectures, are widely applied in tasks such as segmentation, classification, and detection. Performance analyses indicate that hybrid models significantly outperform traditional CNNs, particularly excelling in tasks requiring global context, and are more data-efficient and computationally efficient than pure ViT models. However, hybrid models still face challenges such as scarce data annotations, poor domain adaptability, insufficient model lightweighting and interpretability, and difficulties in clinical integration. Future research directions include deepening architectural fusion (such as automatic design based on NAS), enhancing data efficiency and trustworthiness (introducing self-supervised learning and interpretable AI techniques), and promoting multi-modal information fusion (integrating imaging, genomics, and other multi-source data) to facilitate the development of precision medicine.

References

[1] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929. 2020 Oct 22.

[2] Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical image analysis. 2017 Dec 1; 42: 60-88.

[3] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017; 30.

[4] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv: 2102.04306. 2021 Feb 8.

[5] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D. Unetr: Transformers for 3d medical image segmentation. InProceedings of the IEEE/CVF winter conference on applications of computer vision 2022 (pp. 574-584).In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 574-584).

[6] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L. Cvt: Introducing convolutions to vision transformers. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 22-31).

[7] Xie Y, Zhang J, Shen C, Xia Y. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. InInternational conference on medical image computing and computer-assisted intervention 2021 Sep 21 (pp. 171-180). Cham: Springer International Publishing.

[8] Uppal D, Prakash S. CLT-MambaSeg: An integrated model of Convolution, Linear Transformer and Multiscale Mamba for medical image segmentation. Computers in Biology and Medicine. 2025 Sep 1; 196: 110736.

[9] Yali DO, Shan LI. Cross-modality multi-encoder hybrid attention U-Net for lung tumors images segmentation. Acta Photonica Sinica. 2022; 51(4): 0410006.

[10] ThangaPurni JS, Braveen M. Unified ARP-ViT-CNN system: Hybrid deep learning approach for segmenting and classifying multiple skin cancer lesions. Array. 2025 Sep 20: 100515.