Disease Diagnosis Based on Deep Learning and Transfer Learning: Evolution from CNN to SAM

Yufei Huang

{hyf041201@outlook.com}

College of Electronic Information Science, Fujian Jiangxia University, Fuzhou, Fujian, China

Abstract. Medical image analysis is crucial for clinical diagnosis. Recent advances in deep learning, especially convolutional neural networks (CNNs), have driven progress in this field. However, challenges such as high annotation costs, small sample sizes, and cross-center data heterogeneity limit model performance. Transfer learning and pre-trained models have therefore become key solutions to these issues. This article with a review of medical image deep learning as the thread, systematically sorts out the core types and mechanisms of transfer learning, the evolution and lightweight path of classic CNN pre-training models, explores the technological development trends from CNN to Vision Transformers (ViT) and Segment Anything Model (SAM), and conducts a comprehensive analysis and discussion. Finally, it summarizes the key challenges and future research directions in the field, clarifies the technological evolution rules in this field, and provides important theoretical and practical references for building clinically applicable medical imaging intelligent systems.

Keywords: Medical image analysis; Transfer learning; Pre trained model; CNN;SAM

1 Introduction

Medical image analysis, as the core technical support for clinical disease diagnosis, treatment plan formulation, and dynamic evaluation of efficacy, occupies an irreplaceable key position in the clinical decision-making system. With the rapid development of artificial intelligence technology, algorithm systems represented by deep learning have achieved breakthrough performance improvements in a series of core analysis tasks such as medical image classification, precise lesion detection, region segmentation, and image registration, thanks to their powerful automatic feature extraction and complex data modeling capabilities. This effectively fills the gaps in efficiency, accuracy, and consistency of traditional artificial image interpretation, and opens up a new path for the intelligent development and clinical implementation of medical imaging. However, in the actual research and clinical application process, the inherent problems of medical imaging data itself have become a key bottleneck restricting the full effectiveness of deep learning models: the professional annotation of medical images highly relies on the professional knowledge and practical experience of senior clinical physicians, and completing high-quality annotation of a single case often requires a lot

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365250

of time and effort, directly leading to high annotation costs, and the size of high-quality annotation samples available for model training is generally limited; At the same time, there is a significant class imbalance in medical imaging data, with a much larger number of imaging samples for common and frequently occurring diseases than rare diseases. In addition, there are differences in equipment models, scanning parameters, diagnostic standards, etc. among different medical institutions, resulting in significant distribution heterogeneity and domain shift phenomena in cross center imaging data. These issues overlap with each other, making it difficult for deep models to fully learn generalized feature representations when training from scratch. This not only greatly reduces the efficiency of model training, but also makes it difficult for the model to maintain robust generalization ability in different clinical scenarios and data sources, which cannot meet the core requirements of clinical applications for model standardization and reliability.

In this context, transfer learning and pre-trained models, with their unique technical advantages, have become the core technological path to overcome the aforementioned data and training challenges, improve the performance of deep models, and reduce the demand for target domain data. Transfer learning transfers knowledge and feature representations obtained from training on large-scale data in the source domain to medical image analysis tasks in the target domain, effectively reducing the model's dependence on large-scale annotated data in the target domain. Pre-trained models learn universal feature extraction capabilities by completing pre-training on general large-scale datasets or medical datasets. They only need to make simple adjustments based on downstream specific medical imaging tasks to quickly adapt to application scenarios. The combination of the two can significantly improve the detection performance and clinical application robustness of the model under limited annotated data conditions, becoming the mainstream technical solution in the field of medical image deep learning. At the same time, with the continuous iteration of technology and the deepening of research, this technological system has also shown a clear trend of development and evolution. It has gradually evolved from classic convolutional neural network (CNN) pre-training models represented by AlexNet, VGG, ResNet, etc., to ViT type Transformer models with stronger global modeling capabilities, and now extends to the general basic model stage represented by SAM. The feature representation ability, general adaptability, and cross modal application potential of the model have been continuously improved.

Currently, the application research of transfer learning and pre-trained models in the field of medical image analysis is constantly deepening, and the technical system is continuously improving. However, there is still a lack of systematic sorting and summary of their technological evolution, application practice effectiveness, and existing core challenges in the field. Therefore, this article provides a systematic review of the research and application of transfer learning and pre-trained models in the field of medical image analysis. By clarifying the technological development path, analyzing the application characteristics of various models, summarizing the research frontiers and clinical needs, further summarizing the key challenges faced by the field's development and pointing out future research directions, it provides clear theoretical references and practical ideas for the subsequent technical research and clinical application in this field, and promotes the construction of a more robust, interpretable, and deployable medical image intelligent analysis system.

2 Conceptual Framework for Transfer Learning and Pre-Training Models

Transfer learning is a technical method that utilizes the knowledge and feature representations already learned in the source domain to empower and strengthen the learning process of the target domain. Its core function is to effectively reduce the model's dependence on large-scale annotated data in the target domain. This characteristic makes it of great application value in the field of medical image analysis, where data is scarce, and also makes transfer learning a key technical means to solve the inherent problems of medical image data.

Table 1. Genealogy of Transfer Learning and Pre-training Strategies and Applicable Scenarios for Medical Imaging.

category	representative strategy/approach	Applicable data situation	Typical tasks/modalities	reference
Instance migration	reweighting/sample selection	Small sample size and imbalanced categories	Category /Detection	[1][2]
feature migration	feature extraction /shared representation	Cross center distribution differences	multimodal classification	[1][3]
parameter migration	Pre-training + Fine-tuning/Freezing	small sample size, fast migration	Category /Segmentation	[1][4][5]
relationship migration	Relationship constraints/alignment	Heterogeneous labels or domain differences	Cross-domain task	[1][2]
self-supervised pre-training	Contrastive learning/mask modeling	Labeling scarcity, pre- training within the domain	Segmentation/Classification	[1][6]

Table 1 summarizes the five core strategies of transfer learning and pre-training, as well as their corresponding medical imaging application scenarios. Instance transfer is achieved through sample reweighting or selection, and is suitable for classification and detection tasks with small samples and imbalanced categories; Feature transfer focuses on feature extraction and shared representation to address multimodal classification problems with cross center distribution differences; Parameter transfer adopts pre training+fine-tuning or parameter freezing methods, suitable for fast transfer in small samples, focusing on classification and segmentation tasks; Relationship transfer solves cross domain tasks of heterogeneous labels or domain differences through relationship constraints and alignment; Self supervised pre training relies on methods such as contrastive learning and mask modeling, and performs outstandingly in segmentation and classification tasks that require scarce annotations and pre training within domains. The system explains that multimodal scenarios and cross domain differences make pre-training reuse an urgent need.

3 The Maturity and Application Expansion of CNN and Transfer Learning

Convolutional neural networks, centered around pre-training on the ImageNet dataset, have become the mainstream paradigm for transfer learning in the field of medical image analysis. Classic CNN models such as AlexNet, VGG, GoogLeNet, and ResNet, with their mature feature extraction capabilities, can significantly improve model performance in deeper or wider residual structure design and downstream segmentation tasks using pre-trained features [1] [7]. At the same time, in order to adapt to clinical scenarios and edge device deployment requirements, lightweight CNN models such as SqueezeNet have emerged, achieving efficient image analysis with fewer parameters and becoming an efficient choice for clinical applications [8].

Transfer learning, relying on the feature reuse ability of CNN pre-trained models, has significant advantages in small sample training scenarios of medical imaging. Dawud et al. transferred the AlexNet model pre-trained on natural images to brain CT image analysis to complete the classification task of cerebral hemorrhage, verifying the practical effectiveness of general image pre-training in small sample scenes of neuroimaging [4]. Srinivas et al. compared the performance of VGG-16, ResNet-50, and Inception-v3 in brain tumor MRI classification. The classification accuracy of all three models was significantly improved due to transfer learning, with VGG-16 showing the most stable performance [5]. Sui Xiaoyu et al. used transfer learning combined with CNN to achieve accurate detection of carotid plaques [9], while Liu Ziyan et al. used multi parameter MRI and CNN deep transfer learning to effectively predict the microsatellite status of endometrial cancer [16], fully demonstrating the application value of this technology in clinical diagnosis.

4 Disease Diagnosis Based on Transformer and Transfer Learning

Transformer, with self attention mechanism as its core, was initially applied in the field of natural language processing. It has powerful global context capture and long-range dependency modeling capabilities. After being transferred to medical image analysis, it has become an important technical supplement to CNN and has also promoted the further development of transfer learning in this field.

Compared with CNN, Transformer breaks through the limitations of local receptive fields and can achieve global feature extraction without relying on stacked network layers. It has more advantages in capturing the correlation features between lesions and surrounding tissues in medical images, and has stronger robustness in complex disease imaging diagnosis. CNN is better at extracting detailed features such as local textures.

In the application of disease diagnosis, Usman et al. analyzed the transfer learning ability of ViT in chest X-ray images and confirmed that its feature transfer and global modeling performance were superior to traditional CNN [10]. Che et al. constructed a transfer learning model based on pre-trained ViT to achieve accurate prediction of lymph node metastasis in CT images of lung adenocarcinoma [11]. Springenberg et al. compared CNN and ViT in pathological section analysis, verifying the advantages of ViT in performance and robustness

[12], providing an important reference for the application of the Transformer in medical imaging diagnosis.

5 SAM and Basic Models: Introduction of Universal Segmentation Capability

SAM, as an emerging universal segmentation foundation model, marks the shift of transfer learning from model parameter reuse to universal capability reuse. Its powerful universal segmentation ability, combined with transfer learning, provides a new technological path for medical image disease diagnosis, demonstrating significant potential for precise lesion segmentation and recognition[1].

Compared with traditional CNN, although CNN is good at extracting local texture features, it is limited by the local receptive field and has weaker general segmentation ability; SAM breaks through this limitation by achieving global scale object segmentation without the need for hierarchical stacking, and can directly adapt to various medical image segmentation tasks, becoming an important advancement and supplement of CNN in the field of medical image segmentation.

At present, the integration of SAM and transfer learning has been a key exploration in tumor diagnosis and has become a research hotspot in this field. Asha and Bhavanishankar introduced SAM into CT image analysis of lung nodules, combined with transfer learning to optimize the nodule segmentation process, effectively improving the segmentation accuracy of lung nodules and providing an efficient intelligent method for early screening of lung cancer [13]. Das et al. proposed the AdaViT adaptive pre-training framework based on SAM, which enables compatible input of multimodal 3D medical images, and combines transfer learning to complete feature adaptation, further expanding the application boundaries of the basic model in multimodal disease diagnosis [14]. Overall, SAM combined with transfer learning has brought new breakthroughs in medical disease diagnosis, but its adaptability and clinical reliability in medical scenarios still need further verification.

6 Comparative Analysis And Discussion

The local receptive field limitation of CNN has limitations in long-range dependency modeling, which promotes the exploration of Transformer like methods in medical imaging. Wang et al. pointed out that Transformer is more likely to capture global context through self attention mechanism and demonstrate potential in multimodal scenes [1].

In specific research, Usman et al. systematically analyzed the transfer learning ability of ViT in chest X-ray images and found that ViT is superior to traditional CNN in feature transfer and global modeling [10]. Chen et al. constructed a deep transfer learning model based on pre-trained ViT for predicting lymph node metastasis in lung adenocarcinoma CT images, verifying the feasibility and competitiveness of ViT [11]. Springenberg et al. compared the performance, robustness, and interpretability of CNN and ViT in pathological sections, providing a benchmark reference for the transition from CNN to Transformer [12]. Overall, ViT demonstrates global modeling and robustness advantages in complex medical imaging scenarios, but computational cost and data requirements remain limiting factors [1] [12].

Table 2. Technological Evolution and Comparison of CNN → ViT → SAM.

Architecture phase	representative model	main advantages	Typical tasks	Main limitations	reference
Classic CNN	AlexNet/VGG/ResNet	Strong local features and stable training	Classification/Detection	Insufficient long-range dependency modeling	[1][7]
Lightweight CNN	SqueezeNet	Few parameters, deployment friendly	Classification	Limited expression ability	[8]
Vision Transformer	ViT/Swin	Global modeling, cross modal potential	Classification/segmentation	High computational and data costs	[10][11][12]
foundation model	SAM/AdaViT	Universal segmentation ability, pre-training transfer	Segmentation/Multimodal	Medical adaptation to be validated	[1][13][14]

With the continuous accumulation of medical imaging data, self supervised learning has become a key direction for improving pre training representation and robustness as shown in Table 2. Jiang et al. demonstrated in CT segmentation of lung tumors that self supervised pre training can significantly improve the robustness of the model under different imaging conditions [15]. Almatta et al. proposed a multi-scale self-monitoring framework, combined with CNN and ViT, to effectively alleviate the problem of labeling scarcity and improve the grading performance of diabetes retinopathy, reflecting the trend of transition from cross domain migration to intra domain pre-training [6]. These studies indicate that self supervised learning has the potential to become an important pathway for constructing medical imaging basic models and provide more robust initialization and transfer capabilities for downstream tasks.

Clinical implementation requires models to be transparent, robust, and accountable. The interpretability review summarizes multiple types of explanatory methods and emphasizes the importance of evaluating the quality of interpretation. In practice, Grad CAM visualization has also been used for medical model interpretation and reliability enhancement [16]. Therefore, the performance improvement of transfer learning and pre-trained models should be promoted in conjunction with the improvement of interpretability.

Although significant progress has been made in the research of transfer learning and pre-training models in the field of medical image analysis, various methods still face many common challenges in practical clinical applications. At the data level, the problems of limited sample size and high annotation costs always exist. The distribution heterogeneity and domain shift phenomenon of cross center data are significant, which makes it difficult for the model's generalization ability to meet clinical standardization requirements; In terms of computing power and efficiency, advanced models such as ViT and SAM have high computational costs, complex network architectures, and large parameter scales, which limit their deployment and practical application in clinical edge devices; At the level of evaluation and credibility, there is a lack of unified evaluation standards in the field, and the interpretability and reliability verification of various models are still insufficient, making it difficult to meet the core requirements of clinical transparency and accountability for models; At the level of multimodal fusion, collaborative modeling of multi-source medical images is still in the exploratory stage and has not yet formed a mature technical framework.

In response to the core challenges of current field development, future research needs to focus on three core directions: firstly, to construct high-quality and standardized medical imaging datasets, establish a unified model evaluation system, alleviate sample scarcity and cross center domain offset problems through domain adaptation, data expansion, and other methods, and provide a reliable foundation for model training; Secondly, focus on the research and development of lightweight and efficient training technologies for models, reduce the computational costs of advanced models such as Transformers and SAMs, and enhance their deployment feasibility in clinical edge devices; Thirdly, deeply integrate model performance optimization with interpretability and credibility enhancement, improve the interpretation methods and reliability verification system of medical AI models, promote transfer learning and pre training models to better adapt to clinical application needs, and achieve deep integration of technology and clinical practice. Future research needs to strike a balance between high-quality data and standardized evaluation, lightweight models and efficient training, and trustworthy AI and clinical usability.

7 Conclusion

Transfer learning and pre-trained models have become the core technical support for breaking through the bottleneck of medical image analysis data and improving model performance, promoting the rapid evolution of this field from traditional manual analysis to intelligent deep learning analysis. Starting from the evolution of technology, this article systematically reviews the development process from classical CNN to ViT, and then to SAM basic models, clarifying the technical characteristics, advantages, and applicable scenarios of models at different stages. Classic CNN lays the foundation for transfer learning applications with its stable local feature extraction ability, while lightweight CNN further adapts to clinical edge deployment; ViT breaks through the limitations of long-range dependency modeling with its self attention mechanism, enhancing the potential for global feature capture and cross modal applications; SAM promotes the upgrade of transfer learning from parameter reuse to general capability reuse, providing a new technological path for medical image segmentation. At the same time, the development of self supervised learning has provided important directions for alleviating the scarcity of medical image annotation and constructing basic representations in the field. The research on interpretability and trustworthy AI has also built a key bridge between technology development and clinical implementation. At present, this field still faces core challenges such as data scarcity, high computing power costs, inconsistent evaluation standards, and immature multimodal fusion. Future research needs to collaborate on high-quality data construction, lightweight model training, and trustworthy AI development to achieve a balance between technical performance and clinical usability. Only by continuously promoting the development of models towards more robust, interpretable, and deployable directions can transfer learning and pre-trained models truly empower clinical diagnosis, build medical imaging intelligent systems that meet clinical needs, and provide stronger technical support for the development of precision medicine.

References

[1] Wang, J., Wang, S., Zhang, Y.: Deep learning on medical image analysis. CAAI Transactions on Intelligence Technology. Vol. 10, No. 1, pp. 1-35 (2025)

[2] Zhuang, F., Qi, Z., Duan, K., et al.: A comprehensive survey on transfer learning. Proceedings of the IEEE. Vol. 109, No. 1, pp. 43-76 (2020)

[3] Yu, X., Wang, J., Hong, Q.Q., et al.: Transfer learning for medical images analyses: A survey. Neurocomputing. Vol. 489, pp. 230-254 (2022)

[4] Dawud, A.M., Yurtkan, K., Oztoprak, H.: Application of deep learning in neuroradiology: Brain haemorrhage classification using transfer learning. Computational Intelligence and Neuroscience. (2019)

[5] Srinivas, C., S, N.K., Zakariah, M., et al.: Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images. Journal of Healthcare Engineering. (2022)

[6] Almattar, W., Anwar, S., Al-Azani, S., et al.: Multi-scale self-supervised learning for deep knowledge transfer in diabetic retinopathy grading. Scientific Reports. (2025)

[7] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM. Vol. 60, No. 6, pp. 84-90 (2017)

[8] Iandola, F.N., Han, S., Moskewicz, M.W., et al.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv. (2016)

[9] Sui, X., Han, J., Cao, Y., et al.: Research on multi-network carotid artery image classification and detection based on transfer learning and convolutional neural networks. Chinese Journal of Biomedical Engineering. Vol. 43, No. 3, pp. 267-277 (2024)

[10] Usman, M., Zia, T., Tariq, S.A.: Analyzing transfer learning of vision transformers for interpreting chest radiography. Journal of Digital Imaging. (2022)

[11] Chen, C., Luo, Y., Hou, Q., et al.: A vision transformer-based deep transfer learning nomogram for predicting lymph node metastasis in lung adenocarcinoma. Medical Physics. (2024)

[12] Springenberg, M., Frommholz, A., Wenzel, M., et al.: From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology. Medical Image Analysis. (2023)

[13] Asha, V., Bhavanishankar, K.: Advanced lung nodule segmentation and classification for early detection of lung cancer using SAM and transfer learning. arXiv. (2024)

[14] Das, B.K., Zhao, G., Liu, H., et al.: AdaViT: Adaptive vision transformer for flexible pretrain and finetune with variable 3D medical image modalities. arXiv. (2025)

[15] Jiang, J., Rangnekar, A., Veeraraghavan, H.: Self-supervised learning improves robustness of deep learning lung tumor segmentation models to CT imaging differences. Medical Physics. (2024)

[16] Liu, Z., Ding, Y., Bai, G.: Deep transfer learning features based on multi-parameter magnetic resonance imaging for predicting microsatellite status in endometrial cancer. Chinese Journal of Medical Imaging. Vol. 33, No. 5, pp. 546-552 (2025).