Dean&Francis

ISSN 2959-6157

# The Evolution of Deep Learning in Medical Imaging

**Yuanao Ye**

School Of Mathematics and Statistic, Wuhan University of Technology, Wuhan, 430000, China
* 343920@whut.edu.cn

**Abstract:**

In the past 10 years, the evolution path of medical imaging AI in terms of model architecture is very clear: it is evolving from CNNs for end-to-end feature learning, to ResNets for alleviating deep degradation issues, to EfficientNet balancing accuracy and efficiency with compound scaling, to Transformers for global modeling with self-attention. It is systematic and clear, but it is also important to map it systematically, which will help to be clearer about technology selection and further optimize clinical translation. This paper focuses on four major models/captures CNN, ResNet, EfficientNet, and Transformer, summarizes their basic designs and innovation points.

**Keywords:** Medical imaging; Convolutional neural networks; ResNet; Transformer.

## 1. Introduction

This paper traces the systematic evolution of medical imaging AI, examining the architectural transitions from CNN to ResNet to EfficientNet to Transformer. It analyzes their design motivations and key innovations while comparing their suitability and limitations across segmentation, classification, detection, and reconstruction tasks. Additionally, it summarizes reusable practices across training strategies, data governance, and engineering deployment, emphasizing the complementary relationship between convolutional local priors and global modeling via self-attention. Building upon this foundation, the paper proposes an integrated approach combining "base models + multimodality + interpretability." This approach leverages self-supervised learning and federated learning to enhance cross-domain generalization, while advancing implementation within privacy-compliant and clinically auditable frameworks. The goal is to provide clear technical selection and implementation pathways for both research and practical applications.

## 2. CNN: Foundations of Deep Learning in Medical Imaging

Medical imaging is an important tool in clinical screening, diagnosis, treatment efficacy, and follow-up care. Deep Learning (DL) has gained significant prominence in recent years, largely driven by the expansion of big data [1]. The explicit local priors and effective parameterization of convolutional neural networks (CNNs) have made them the foundation of the deep learning age. By weight sharing, local connections and translation, CNNs build feature extraction systems. They progressively expand the effective receptive field through a cascade of 'convolution, nonlinear activation, and pooling/strided

downsampling' operations. Together with proper padding and widened convolutions, they extract more extensive contextual details with parameters almost staying the same. Goodfellow et al. [1] identified three key benefits of the CNN: equivalent representations, sparse interactions, and parameter sharing [2]. Activations like ReLU stabilize the gradient and make convergence faster, whereas separable convolutions, group convolutions, and 1x1 point convolutions lead to a large increase in parameter and computation efficiency. CNNs are typically tuned to medical data structure with either 2D or 3D convolutions, anisotropic kernels, and patch-based training to trade off resolution and memory between slice and volumetric data. This property allows CNNs to remain the state-of-the-art in 2D/3D segmentation, registration and reconstruction due to their extreme sensitivity to textual clues and structural delimitation.

However, the performance of CNNs plateaus as networks grow deeper, due to challenges such as vanishing/exploding gradients and the degradation problem, which hinder effective training and limit their ability to model very complex, long-range dependencies inherent in some medical imaging contexts.

## 3. ResNet: Residual Learning and the Optimizability of Deep Networks

ResNet uses residual units that learn the "residual function" with the identity mapping as a shortcut. Such a design substantially alleviates the degradation and vanishing gradients problem in deep networks, simplifies the optimization and improves deep semantic representation ability. The structure is shown in Fig.1.

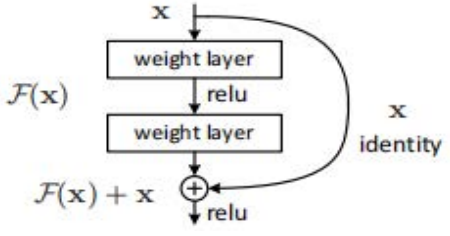


**Fig. 1 Residual learning [3]**

The generalized residual block is general to not only be used in CNNs, but also in standard fully connected layers and other feedforward layers. Replacing each of the convolutional layers within a residual block from the original ResNet with a generalized residual block leads us to a new architecture called ResNet in ResNet (RiR) [3]. In medical imaging, this mechanism brings three advantages: First, the deeper network can integrate more local-global context in cases with complicated organ backgrounds and various scales of lesions; Second, the residual architecture allows coupling with attention, grouped convolutions and deformable convolutions to form modular backbones with scalable depths; Third, transfer learning becomes more robust: pre-training on general datasets such as ImageNet makes the network easier to converge and leads to better generalization on small-sample scenarios. Its applications include multi-label classification on chest radiographs, pathological image classification and CBIR.

Common strategies include freezing shallow layers and fine-tuning upper layers to suppress over-fitting and using multi-head or multi-classifier "divide-and-conquer" strategies in multi-label classification to alleviate label coupling and imbalance. Many empirical results have shown that residual connections can substantially ease the difficulty of training deep neural networks to fit the training sample while still maintaining good generalization ability on test examples [4].

## 4. EfficientNet: Practical Feasibility of Composite Scaling

Another key direction in exploring model performance improvements is enhancing computational efficiency.For practical clinical deployment,especially under computational constraints or in time-sensitive scenarios,achieving an optimal balance between accuracy, speed, and resource consumption is essential. EfficientNet addresses this need by systematically improving the efficiency of CNN-based models.

EfficientNet is based on composite scaling as its fundamental strategy, combining the scaling of network depth, width and input resolution. It uses neural architecture search (NAS) to find cost-efficient baseline units, which makes it gain more accuracy than single-dimensional scaling on image classification tasks within a constrained computational budget. These models have been built based on an approach referred to as compound scaling where network width, depth and resolution are equally scaled with a given set of scaling coefficients[5]. EfficientNet and its adaptation versions in medical imaging have shown high scores in various activity types such as pathology image representation, ear biometric identification, prostate cancer Gleason scoring, and multi-label classification of chest X-rays. They especially fit well in situations where sample sizes are moderate, and the detection latency of inferences is important. A common strategy used in engineering practice involves pretraining at a reduced resolution and gradually increasing to the task resolution. In this method, knowledge distillation and quantization are also used to compress the model further, and the resultant architecture is a multi-task model that is

ISSN 2959-6157

represented as a shared logic rooted at task-specific heads. The design of the architecture of EfficientNet was produced using the neural architecture search method of neural network architecture creation. It maximizes the efficiency and accuracy of the concept of floating-point operations per second (FLOPS). The convolution that is used in this proposed architecture is the moveable inverted bottleneck convolution (MBConv). The researchers then expanded on this foundation network to build the EfficientNets family[6]. It is worth mentioning that scaling too many times can lead to the loss of details and edge blurring, and it can be addressed by using multi-scale training, fine-grained supervision loss, or post-processing methods.

## 5. Transformer: Global Dependencies, Hierarchical Windows, and Multimodal Potential

However,CNN-based models are inherently limited by their local receptive fields, making it difficult for them to capture long-range contextual relationships. The Transformer architecture overcomes this by employing self-attention, which directly models interactions between all input elements. This shift from local feature extraction to global relational modeling provides a fundamental advantage for understanding complex visual scenes.

Transformers are able to extract long-range dependencies with self-attention as shown in Fig.2 [7]. The quadratic complexity of computing actions would have to be added if it were applied directly to images. As a result, visual Transformers (as ViT and Swin Transformer) use segmentation, windowing, and hierarchical pyramids to control computation expenses. The capability of these transformer models is an encouragement to the entire community to research the application of transformers when completing a visual task [8].

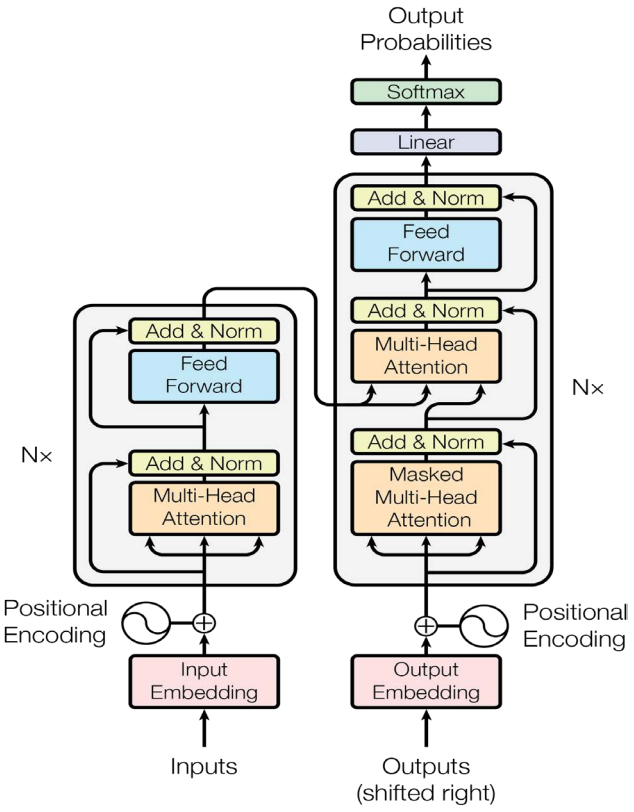


**Fig. 2 Transformer [7]**

More generally, it has three key benefits in medical imaging: First, whole-body interactions, its sensitivity to long-range structural interactions enables accurate segmentation of small organs, diffuse lesions and interactions between organs; Second, cross-layer alignment, when parallelized with CNNs, or fused at bottlenecks, it reduces the mismatch between representations through frequency-domain or attention-based bridging; Third, multimodal fusion combined with textual and structural clinical information, jointly encoding medical imaging can be used to advance semantically consistent holistic representation of medical images; Transformers began to appear as a strong competitor, capable of operating both local and long-range dependence as well as being able to parallelize training [9]. In the case of segmentation activities, TransUNet uses a hybrid design that features a CNN head and a ViT bottleneck. Transformer models with less than 50 percent information loss, like Swin-UNet, UNETR, and nnFormer, represent pure or semi-pure models directly using CNN-Transformer as the encoder and then a cascaded upsampler to provide localization accuracy [10]. Therefore, to compensate for such information loss, TransUNet employs a hybrid CNN-Transformer architecture as the encoder as well as a cascaded upsampler to enable precise localization [10]. Transformers can perform well to learn transfer learning. It is necessary to mention that they are

more punitive to the data scale and the optimization settings. The engineering practices generally mix self-supervised pretraining, mixed-precision training, and specific augmentation plans to improve performance in terms of convergence and generalization.

## 6. From CNN to Transformer: The Internal Logic of Continuous Evolution

Medical image intelligent computing takes an iterative course of problem-based innovation and structural improvement. CNNs were developed as the basis of end-to-end learning using local, connected, with weight-sharing. In a bid to eliminate gradient vanishing and degradation in deep networks, the ResNets introduced shortcuts and identity mappings. To effectively balance efficiency and accuracy requirements, EfficientNet suggested a synergistic composite scaling mechanism to optimise depth, width and resolution. Once the global dependencies were reduced to bottlenecks, the self-attention mechanism of Transformers was used to overcome the limitations of the receptive field by CNNs. Recent evidence suggests that CNNs‘ locality is an addition to global attention of the attention system by creating a hybrid architecture paradigm of convolutional backbone-attention augmentation.

Attention mechanisms of Image segmentation are benchmarked to the U-Net as well as the 3D extensions that are advantageous to provide global consistency when there are blurred boundaries or for smaller objects. To classify and retrieve using limited resources, the ResNet/ EfficientNet series has better performance to efficiency trades. Multi-head classifiers can be used in multi-label classification in order to alleviate feature competition. Multi-scale integration multi-instance learning is needed in a pathological whole-slide analysis. Object delineation is full-fledged in 2D imaging. In complex backgrounds and in 3D data and situations with a strong long-range dependency, cross-scale attention enhances performance. Reconstruction tasks are still mainly controlled by CNNs, in which the presence of attention in critical stages boosts structural fidelity and detail restoration. Models used in clinical translation need to be interpretable, robust, and engineerable. An explainability framework must include intrinsic (concentrating on visualizations of attention, class activation maps) and after-the-fact (SHAP/LIME) procedures, and must be aligned on the medical semantic level. The quantification of uncertainty should be able to distinguish between the random and the cognitive uncertainty, which are assessed and approximated by deep ensembles and Monte Carlo Dropout. Pruning-quantization-distillation optimization is necessary to deploy the model efficiently, and engines such as ONNX/TensorRT can be exploited. The edge computing should employ lightweight network and memory reuse tactics to have a millisecond response but be sensitive to small lesions. Integration of systems should be done in accordance with standards like DICOM and HL7/FHIR and a complete cycle quality monitoring mechanism and drift detection should be provided.

This paper suggests that future studies will build on three main directions: constructing foundational models that specifically relate to medical imaging, achieving unified model-multi-task adaptation with large-scale self-supervised pre-training and parameter-efficient fine-tuning, multi-level, multi-modal fusion methodology creating a unified representation of both imaging and non-imaging data, and considerate integration of trustworthy AI principles (explainability, fairness, protection of personal data) across the research and development cycle as a whole. These technological advancements include the use of deformable attention, anatomy-based graph networks, and combined frequency spatial domain modeling, which are prone to making significant breakthroughs in arduous tasks like 3D image processing and tiny image organ segmentation, which advance the technologies to clinical-grade productivity.

## 7. Conclusion

The paper presents a survey of the ongoing medical imaging AI development in the context of deep learning technology, and it can be seen that the common logic inherent in this field is a problem-solution formulation: the development of end-to-end representations by convolutional local priors, the recovery of deep trainability by residual learning, the maximization of accuracy-efficiency tradeoffs by composite scaling, and the long-range dependencies and global context, in turn, by self-attention. In tasks such as segmentation, classification, retrieval, detection, and reconstruction, experience suggests that the best practice is neither based on a single architecture but rather leads to synergy between structure, training, and engineering constructing cost-effective building blocks using lightweight convictions and composite scaling, improving representations using residual and attention modules, improving cross-domain generalization using self-supervised and federation learning, and bending clinical reliability and conformability through interpretation and uncertainty management. Although the obtained results have shown that the DL algorithms have high diagnostic accuracy in medical imaging, it is currently hard to tell whether they are clinically acceptable or applicable. The combined strategy of the integrated models of foundational models plus multimodal capabilities plus interpretability will lead

ISSN 2959-6157

to the transition of the leading laboratory measures to auditable, replicable, and scalable clinical-grade productivity in the future.

## References

[1] Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning, vol. 1. Cambridge: MIT Press, 2016.

[2] Alzubaidi L, Zhang J, Humaidi A J, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 2021, 8(1): 53.

[3] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770–778.

[4] Tan M, Le Q V. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML), 2019: 6105–6114.

[5] Azim M A, Jahan M U, Almotiri S H, et al. DU-Net: A dual-encoder U-Net architecture for scalable medical image segmentation. Journal of Healthcare Engineering, 2024, Article ID 3583612:,pages 1–20.

[6] Chowdhury M E H, Rahman T, Khandakar A, et al. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access, 2020, 8: 132665–132676.

[7] Ghafoor S S M, Munir K, Al Farraj O, et al. A comprehensive CNN approach for Alzheimer's disease classification using MRI scans. IEEE Access, 2022, 10: 42613–42626.

[8] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. Advances in Neural Information Processing Systems (NeurIPS), 2021, 34: 1–25.

[9] Kirillov A, Mintun E, Ravi N, et al. Segment anything. arXiv:2304.02643, 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[10] d' Ascoli S, Touvron H, Leavitt M L, et al. ConViT: Improving vision transformers with soft convolutional inductive biases. arXiv:2103.10697, 2021.[Online]. Available: https://arxiv.org/abs/2103.10697