The CNN and ViT Fusion Model Based on Hierarchical Adaptive Token Refinement Method in Pneumonia X-ray Image Classification

Peiqi Zhang[a]

Computing Science, Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Guangdong Foshan, China

Keywords: Pneumonia Diagnosis, X-ray Images, CNN, ViT, Fusion Model.

Abstract: Convolutional Neural Networks (CNN) and Vision Transformers (VIT) each have their own advantages in medical image analysis, particularly in the automatic classification of X-ray images. Many studies have contributed to the effective combination of these two.This paper proposes a CNN and ViT merging method - Hierarchical Adaptive Token Refinement (HATR), combining the local feature extraction capability of CNN with the global modeling ability of ViT. The experimental results show that the accuracy rate of the fusion model based on ResNet (HATR-ResNet) is 91.4%, which is significantly better than that of ResNet alone (87.3%). The accuracy rate of the fusion model based on Conv2D (HATR-Conv2D) is 88.2%, which is approximately 5% higher than that of Conv2D alone (82.7%). The superiority of HATR-ResNet stems from the deep residual network structure of ResNet, which can better extract complex features and capture details, while the shallower network structure of Conv2D is relatively insufficient when dealing with complex patterns.This study proposes a new fusion method for CNN and ViT, and compares the performance differences of the fusion models based on different CNN backbones. It contributes to the subsequent research on new model structures and the exploration of new fusion methods.

1 INTRODUCTION

Be advised that papers in a technically unsuitable form will be returned for retyping. After returned the manuscript must be appropriately modified.

The field of Chest X-ray (CXR) image analysis plays a crucial role in the early diagnosis of pneumonia. Traditional methods for diagnosing pneumonia rely on doctors' experience and imaging examinations. However, in some high-load and resource-tight environments, this may lead to diagnostic delays or even errors (Litjens et al., 2017). A study published in JAMA Internal Medicine in 2024 analyzed 17,290 inpatients from 48 Michigan hospitals and found that 12.0% of the patients were misdiagnosed as having community-acquired pneumonia (Gupta et al., 2024). With the advancement of artificial intelligence technology, Computer-Aided Diagnosis (CAD) systems based on deep learning can extract complex features from a large number of medical images, greatly improving the efficiency and accuracy of diagnosis.

Convolutional Neural Network (CNN) is the core architecture of deep learning and has performed exceptionally well in the field of medical image analysis, especially in the task of pneumonia detection (Litjens et al., 2017; Kermany et al., 2018). ResNet solved the degradation problem in deep network training through residual connections, enabling the network to extract deeper image features (He et al., 2016). Models based on Conv2D are often used to identify detailed features in medical images, such as abnormal textures in X-ray images of the lungs (Ronneberger et al., 2015). However, CNN has limitations in capturing long-term dependencies and global features, which may restrict its performance in complex medical image tasks(Raghu et al., 2019).

To overcome the shortcomings of CNN in global feature modeling, the Vision Transformer (ViT) has become a research hotspot in recent years. ViT uses the self-attention mechanism to achieve global image modeling, which can more effectively capture long-term dependencies compared to traditional CNN (Dosovitskiy et al., 2021; Vaswani et al., 2017).

[a] https://orcid.org/0009-0003-4096-7208

Zhang, P.
The CNN and ViT Fusion Model Based on Hierarchical Adaptive Token Refinement Method in Pneumonia X-ray Image Classification.
DOI: 10.5220/0014361900004718

In *Proceedings of the 2nd International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2025)*, pages 504-510
ISBN: 978-989-758-792-4

Studies have shown that ViT performs well on large-scale datasets, but its performance is usually inferior to CNN on small sample datasets, and it requires higher annotation data and computing resources (Touvron et al., 2021; Chen et al., 2021). This limits the application of ViT in resource-constrained scenarios.

To combine the advantages of CNN and ViT, researchers proposed a hybrid model. Swin Transformer enhances the feature representation ability through the hierarchical window self-attention mechanism, and in the ImageNet classification task, it increased the Top-1 accuracy to 83.5%, which is approximately 3% higher than the baseline model (Liu et al., 2021). CoAtNet integrates convolution and self-attention mechanisms, achieving a balance between computational efficiency and performance. Experimental results show that it has an accuracy improvement of approximately 2.5% compared to the traditional Transformer model under the same computational cost (Dai et al., 2021).

The main objective of this study is to design and compare the performance of different CNN and ViT fusion models in pneumonia classification tasks. By combining the local feature extraction advantages of CNN and the global information modeling capabilities of ViT, this study proposes a modular fusion framework and analyzes the performance of different CNN backbone networks in this framework. Through experiments, this study explored how to optimize the structure of the fusion model to improve the accuracy of pneumonia classification and provide a theoretical basis and practical guidance for the design of future automatic diagnosis systems for pneumonia.

2 DATASET AND METHODS

2.1 Dataset Dscription

The dataset used in this study is from the Kaggle chest X-ray image (Pneumonia) dataset, which is widely used in pneumonia classification tasks. The dataset contains normal lung images and pneumonia lung images. All the images are grayscale and have typical medical imaging features. The composition of the dataset is as follows: The training set approximately contains 5,232 images, among which 1,341 are normal lung images and 3,891 are pneumonia images. The validation set consists of 16 images, including 8 normal and 8 pneumonia samples respectively. The test set contains 624 images, which are divided into 234 normal images and 390 pneumonia images. The sample of normal and pneumonia images is shown in Figure 1 and Figure 2. Although the sample size of the training set is large, the number of pneumonia samples is significantly higher than that of normal samples, resulting in an imbalance in the data. This problem is common in many medical image analysis tasks.

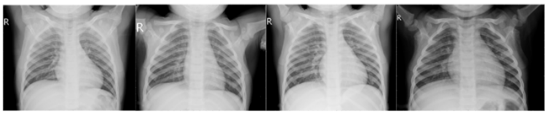

Figure 1: Samples of Normal (Data from: Kaggle).

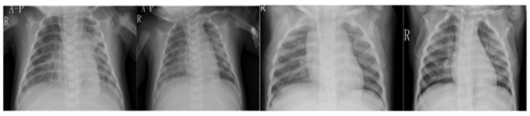

Figure 2: Samples of Pneumonia (Data from: Kaggle).

2.2 Dataset Preprocessing

Since the original image is a grayscale image, it is necessary to convert the image into a three-channel RGB image suitable for deep learning models (especially pretrained models) during subsequent processing. This conversion is usually accomplished by replicating the channels of grayscale images or through pseudo-colorization processing, thereby making the images conform to the input requirements of models such as CNN and ViT on the channels. To meet the size requirements of the model input, all images have been uniformly adjusted to a size of 224x224. This size setting is not only compatible with common convolutional neural network architectures such as ResNet and VGG, but also meets the input size requirements of ViT models.

Another key step in image preprocessing is normalization. First, normalize the pixel values to the [0, 1] interval to reduce the influence of different image brightness and contrast on model training. Then, based on the statistics from ImageNet, the images were standardized, using the mean ([0.485, 0.485, 0.485]) and standard deviation ([0.229, 0.229, 0.229]) of the images. This standardized approach helps the model adapt to the input distribution of the pretrained model, enhancing the stability and convergence speed of the training.

To enhance the robustness and generalization ability of the model, various data augmentation operations were also performed on the training set data. Specifically, the training images are enhanced through random horizontal flipping, random rotation (up to ± 10 degrees), and random brightness adjustment (± 10%), among other methods.

However, to ensure the fairness and accuracy of the model evaluation, the validation set and test set were not subjected to data augmentation processing to avoid the impact of augmentation operations on the evaluation results.

As shown in Figure 3, the imbalance problem of the dataset is one of the key factors affecting the performance of the model. In this study, a weighted cross-entropy loss function was adopted to alleviate this problem. By setting the weight ratio of the normal category and the pneumonia category to 3:1, the model can pay more attention to the samples of a few categories (i.e., the normal category) during the training process, avoiding the model being overly biased towards the pneumonia category. In addition, according to the experimental requirements, oversampling of normal samples or undersampling of pneumonia samples can also be selected to further improve the model's adaptability to imbalanced data.

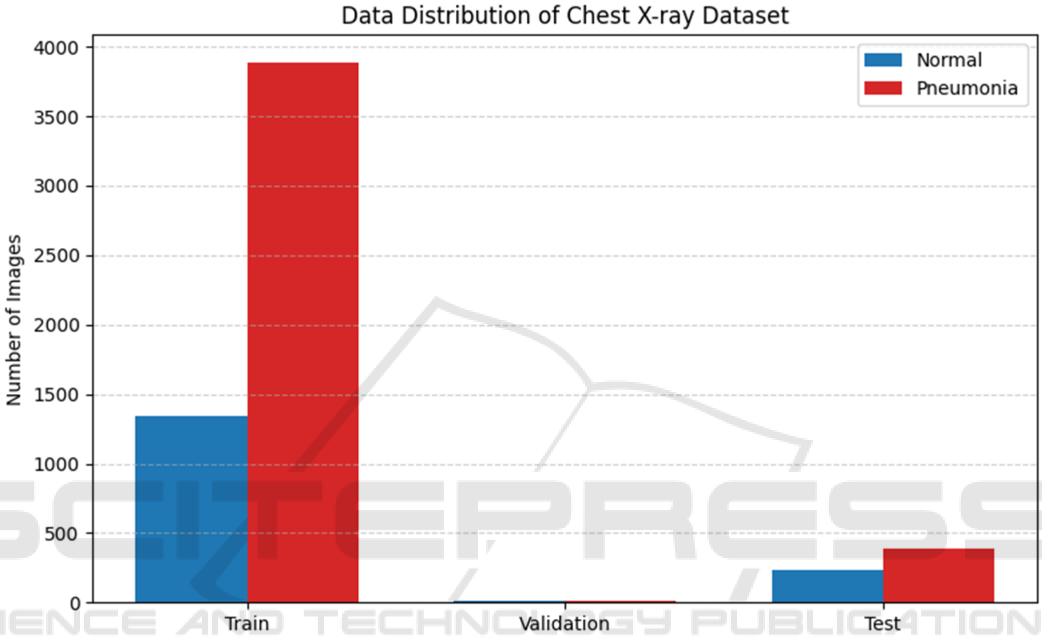

Figure 3: Distribution of the dataset (Picture credit: Original).

2.3 Model

To fully utilize the sensitivity of a convolutional neural network (CNN) to local features and the advantages of Vision Transformer (ViT) in global modelling, this study designs a fusion method. It is named HATR (Hierarchical Adaptive Token Refinement). This method uses CNN as the feature extractor and maps its output to sequence inputs suitable for ViT. By introducing multi-scale feature partitioning and adaptive weighting mechanisms, it achieves efficient feature integration from local to global.

This section first briefly introduces the basic functions and output features of each sub-model, and then focuses on elaborating the design logic and implementation details of the fusion method.

2.3.1 Basic Model

ResNet-50 was used to extract the representation of X-ray images of the lungs, and the final output was a feature map of size [batch, 2048, 7, 7]. This feature map is mapped to 512 channels through dimensionality reduction and convolution to adapt to subsequent module processing.

Conv2D is a lightweight convolutional network composed of three layers of convolutional stacks, with channels of 64, 128, and 256 in sequence. The final output feature map is [batch, 256, 14, 14]. Compared with ResNet, it has a simpler structure and faster training, but its semantic expression ability is relatively weak.

ViT represents images as a series of patches and models the global relationships among different regions through a self-attention mechanism. It does not have the inductive bias of convolution, and thus is more sensitive to the organizational structure of the input features. The input of ViT is a sequence in the form of [batch, N, d], where each element corresponds to an embedded representation of an image Patch, and the output is used for classification through CLS tokens.

2.3.2 Fusion Method: HATR (Hierarchical Adaptive Token Refinement)

The traditional CNN-VIT fusion methods often adopt direct Patch segmentation and input connection, ignoring the differences and spatial continuity of features in each region in the CNN output, which is prone to lead to information fragmentation or redundant transmission. To this end, the research proposes a hierarchical adaptive fusion method, HATR, which optimizes this connection process from two aspects: multi-scale overlapping embedding and Token adaptive weighting. The structure of this method is shown in Figure 4.

Firstly, after the CNN output, the conventional non-overlapping Patch segmentation method is no longer adopted. Instead, an overlapping window (stride < patch size) is introduced to perform convolutional partitioning on the feature map. This approach can retain more context information between patches and slow down the fragmentation of the spatial structure. For instance, for the output of Conv2D [256, 14, 14], by dividing it with a sliding window of kernel size=2 and stride=1, a large number of patches with overlapping areas can be generated, which can then be flattened to form the sequence input [batch, N, 512] required for ViT. The output resolution of ResNet is relatively low. The method first performs a 1×1 convolution dimension reduction on it and then applies the same sliding partitioning strategy.

Secondly, the study introduces a lightweight attention module to perform feature weighting on all patches. This module combines channel attention and spatial attention mechanisms to identify the importance of different patches. Each Patch is assigned a weight coefficient. High-response regions (such as patches containing lesion features) will be high-lighted, while the information of low-response regions will be compressed. The final Patch sequence already possesses the characteristics of spatial continuity and significant regional enhancement before being input into the ViT.

The sequences received by ViT already contain hierarchical details and importance annotations. The long-distance dependencies between various regions are modelled through their self-attention mechanism, and finally, the global information is aggregated by the CLS token for pneumonia classification

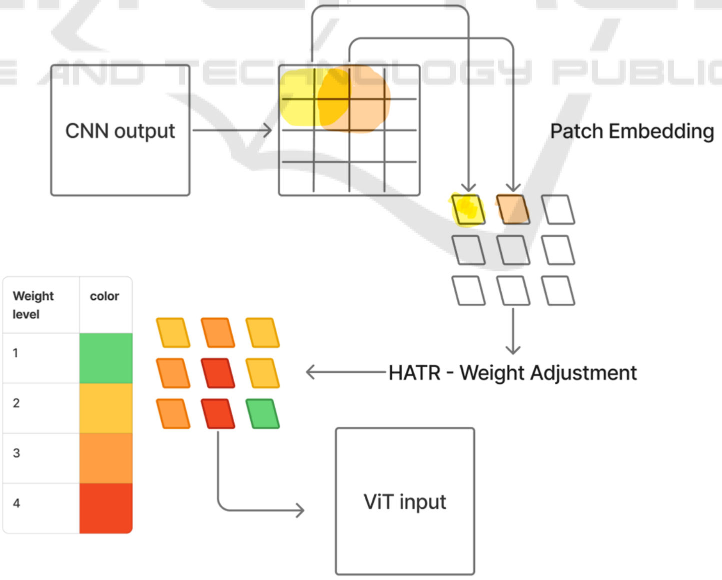

Figure 4: Structure of HATR method (Picture credit: Original).

The HATR method resolves the structural mismatch issue between convolutional features and the Transformer input. The multi-scale overlapping embedding method enhances the correlation between patches and alleviates feature fragmentation. The Token weighting mechanism enables the model to learn to filter the most valuable regions when information is overloaded, thereby enhancing the model's discriminative ability.

Compared with the directly connected fusion structure, HATR can more fully leverage the respective advantages of CNN and ViT while maintaining the overall computational complexity within a controllable range. For scenes like medical images with low contrast and high redundancy, this refined connection and weighting method has higher practical value.

3 EXPERIMENTS

3.1 Experimental Configuration

Table 1: Experimental Configuration.

Item	Details
Hardware	NVIDIA RTX 5060 GPU
Software Framework	PyTorch 1.10Python 3.8CUDA 11.2
Optimizer	Adam
Learning Rate	1e-4
Batch Size	32
Loss Function	Weighted cross-entropy loss (for class imbalance)
Training Epochs	20

The experimental environment of this study is shown in Table 1. This study was conducted in a GPU environment with strong image processing capabilities (RTX 5060), using the PyTorch framework to implement all model training and inference processes. During the training process, the Adam optimizer was used, with an initial learning rate of 1e-4 and a batch size of 32. The total number of training rounds was 20. To address the sample imbalance problem where the proportion of the pneumonia category in the dataset is relatively high, a weighted cross-entropy loss function was employed to adjust the category weights.

3.2 Experimental Result

The study evaluated the performance of different models in classification tasks on the Kaggle chest X-ray pneumonia classification dataset, including individual convolutional neural networks (ResNet-50, Conv2D), Vision Transformer (ViT), and the HATR model based on the fusion of these two CNN trunks and ViT. The comparison results are in Table 2.

Table 2: Performance of each model.

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
ResNet-50	87.3%	86.5%	88.1%	87.3%	0.945
Conv2D	82.7%	81.2%	84.5%	82.8%	0.920
ViT	85.1%	84.3%	86.7%	85.4%	0.930
HATR (ResNet)	91.4%	90.8%	92.1%	91.4%	0.960
HATR (Conv2D)	88.2%	87.5%	89.0%	88.2%	0.950

It can be seen from the results that the overall performance of the fusion model is superior to that of its respective individual backbone models. Among them, HATR (ResNet) achieved the best performance in terms of accuracy, recall rate, and AUC indicators, indicating that the fused structure can more effectively extract and integrate multi-level information in lung images.

ResNet and ViT respectively have advantages in local feature extraction and global structure modelling. The fused HATR model connects the two through overlapping Patch embedding and adaptive Token weighting mechanism, enabling the model to retain the sensitivity of convolution to details while introducing ViT's ability to understand the overall image structure.

Although HATR (Conv2D) is based on a relatively lightweight convolutional network, its performance is still significantly improved after fusion, indicating that even if the basic network capability is weak, a reasonable fusion design can still bring significant gains.

This research focused on comparing the models constructed by fusing two different CNN trunks (ResNet-50 and Conv2D) with ViT, namely HATR-ResNet and HATR-Conv2D. Under the same training configuration and fusion structure, both perform better than their respective single backbone networks in the task of classifying pneumonia X-ray images, but there are still significant differences in specific performance.

From the perspective of overall performance indicators, the accuracy rate of HATR-ResNet is 91.4%, while that of HATR-Conv2D is 88.2%. Not only that, in terms of precision, recall rate, F1 score and AUC and other indicators, HATR-ResNet has always outperformed HATR-Conv2D, demonstrating its stronger classification ability. Especially in terms of AUC (0.960 vs 0.950) and recall rate, the former is more suitable for use in medical scenarios with high sensitivity requirements.

Further analysis of the training process reveals that HATR-ResNet also has more advantages in convergence speed and stability of the validation set. In the first 10 epochs, the training losses of both decreased rapidly, but the validation loss of HATR-ResNet decreased more steadily and the overfitting phenomenon was milder. Although the training loss of HATR-Conv2D decreases rapidly, the performance fluctuation on the validation set is greater, indicating that its generalization ability is slightly inferior. Figure 5 shows the accuracy and loss of HATR models.

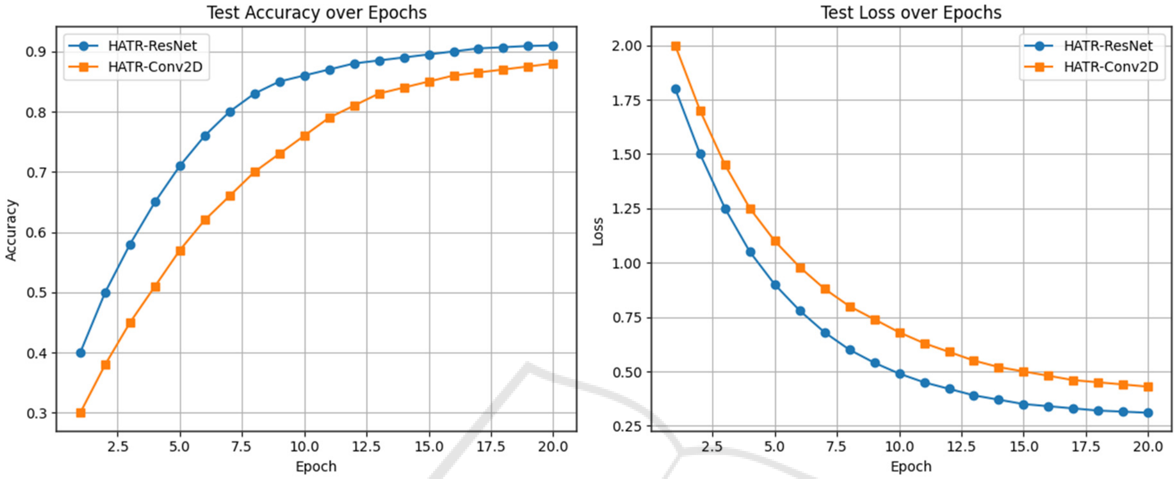

Figure 5: Accuracy and Loss of HATR Models (Picture credit: Original).

The essential reason for this difference lies in the varying capabilities of the backbone CNN structure itself. ResNet features a deeper network structure and residual connections, enabling it to effectively capture complex edges, textures and spatial layouts in images. It is particularly suitable for areas with blurred details and low contrast in pneumonia images. Although Conv2D has a lightweight structure and is suitable for deployment in resource-constrained environments, its shallow network lacks the ability to extract high-level semantic features, which makes it difficult for the fused ViT to receive rich enough information and affects the quality of subsequent modelling.

In addition, in terms of training duration, the training time per round of HATR-ResNet is slightly higher (about 1.2 times), but overall, the performance improvement is significant, making it particularly suitable for use in medical image scenarios where performance takes precedence over computing resources. HATR-Conv2D is more suitable for deployment in edge devices or systems with strict latency requirements, serving as a lightweight alternative.

4 CONCLUSIONS

This study focuses on the automatic recognition task of pneumonia X-ray images and proposes a feature modelling method based on the fusion of CNN and ViT. By constructing the modular structure HATR (Hierarchical Adaptive Token Refinement), the study has achieved the efficient integration of local convolutional features and the global attention mechanism. The experimental results show that, regardless of whether ResNet or Conv2D is used as the backbone, the fusion model outperforms the single structure in key indicators such as accuracy, F1 score and AUC, verifying the feasibility and effectiveness of this design in medical image classification tasks.

The key to the fusion strategy lies in the handling of two aspects: One is to adopt the overlapping Patch embedding method, which alleviates the feature fragmentation problem caused by the traditional segmentation method; Second, an adaptive Token weighting mechanism is introduced, enabling the model to complete the initial screening of features before inputting them into the ViT. This structural design enables the model to retain the convolutional network's ability to pay attention to detailed regions while also leveraging the global modelling advantages of the Transformer to enhance the accuracy of overall judgment.

Although this method performs well on medium and small-scale datasets, there are still some limitations and directions worthy of further exploration:

Firstly, the selection of the CNN backbone is still relatively fixed and lacks structural adaptability. Under different tasks or data distributions, the currently used ResNet or Conv2D may not always maintain stable performance. In the future, more flexible and adjustable backbone structures, such as automatic neural architecture search (NAS), can be explored to enhance generalization capabilities.

Secondly, due to ViT's strong reliance on large-scale data, the fusion model is still prone to overfitting or performance fluctuations when the data volume is insufficient. In addition, to maintain a lightweight configuration, this study has adopted a shallow ViT structure. Although this reduces computing costs, it may still limit the expressive power in more complex data environments.

Finally, the robustness of the current model still needs to be enhanced when dealing with large-scale, multi-category or cross-device collected data. In the future, the stability and practical application value of the model can be further enhanced by integrating transfer learning, domain adaptation or multimodal information (such as clinical text data).

REFERENCES

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 9620–9629.

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems (NeurIPS), 3965–3977.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR), 1–22.

Gupta, A. B., Wang, Y., Smith, J., Lee, D., Chen, M., & Johnson, T. (2024). Inappropriate diagnosis of community-acquired pneumonia among hospitalized adults. JAMA Internal Medicine, 184(5), 548–556.

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

Kermany, D. S., Zhang, K., & Goldbaum, M. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, 172(5), 1122–1131.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & van Ginneken, B. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 10012–10022.

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in Neural Information Processing Systems (NeurIPS), 3347–3357.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234–241.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. Proceedings of the International Conference on Machine Learning (ICML), 10347–10357.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.