Transactions on Computational and Scientific Methods | Vol. 5, No. 3, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

MS-UNet: A Transformer-Based Multi-Scale Nested Decoder Network for Medical Image Segmentation with Limited Data

Logan Whitaker[1], Ethan McAllister[2]
[1]University of Kansas, Lawrence, USA
[2]University of Kansas, Lawrence, USA
loganw88@ku.edu[1], ethanm92@ku.edu[2]

Abstract: With the rapid advancement of deep learning, neural networks have demonstrated remarkable progress in medical image segmentation, significantly enhancing the accuracy and efficiency of lesion detection and organ boundary identification. Traditional medical image segmentation relied on manually designed features, which were labor-intensive and struggled with complex image variations. The emergence of Convolutional Neural Networks (CNNs), particularly UNet and its variants, revolutionized the field by leveraging hierarchical feature extraction. More recently, inspired by breakthroughs in Natural Language Processing (NLP), Transformer-based models, such as Vision Transformer (ViT) and Swin Transformer, have been successfully applied to medical image segmentation, addressing CNNs' limitations in capturing long-range dependencies.However, the direct application of Transformer models introduces challenges, such as a semantic gap between the encoder and decoder, which can hinder segmentation performance. To address this, we propose MS-UNet, a Transformer-based multi-scale nested decoder segmentation framework that enhances feature learning and semantic communication between network modules. By designing a dense multi-scale nested decoder, MS-UNet effectively mitigates the semantic discrepancy, improving segmentation accuracy, especially in scenarios with limited training data. Experimental results on MRI and CT segmentation tasks demonstrate that MS-UNet significantly outperforms CNN-based models and other Transformer-based architectures. This study not only provides an effective solution to medical image segmentation under data-scarce conditions but also offers a novel approach for broader applications in medical imaging.

Keywords: Medical Image Segmentation, Deep Learning, Transformer, Convolutional Neural Network (CNN)

1. Introduction

With the advancement of neural networks and deep learning research, more complex network structures and training algorithms have emerged, enabling deep learning to form deeper networks and handle more complex tasks. Currently, deep learning has achieved significant results in the fields of computer vision, natural language processing, speech recognition, and reinforcement learning, especially in the field of medical image segmentation, where it has shown great potential and practical value.

For instance, the emergence of Convolutional Neural Networks (CNNs) has greatly enhanced the ability to extract image features, allowing deep learning models to better capture detailed information in medical images. Unlike the aforementioned deep learning methods, traditional medical image segmentation [1-2] methods heavily relied on manually designed features and rules, which required a significant amount of labor and time, and were difficult to adapt to the complex and variable medical images. However, deep learning, by automatically learning deep features in images, can more accurately identify key information such as lesion areas and organ boundaries, thereby achieving precise and efficient image segmentation.

With the proposal of UNet [3], CNN-based UNet network variants began to dominate the medical segmentation field. Combined with advanced network structures and training algorithms, the performance of deep learning models in medical image segmentation tasks has been significantly improved, providing strong support for medical diagnosis and treatment.

In the field of Natural Language Processing (NLP), due to the effective establishment of global connections between sequence word blocks through multi-head self-attention mechanisms, Transformers have completely changed most NLP tasks. Inspired by the success in the NLP field, researchers have begun to explore the possibility of applying Transformers to image processing. Consequently, Alexey Dosovitskiy proposed the Vision Transformer (ViT) [4], successfully applying Transformers to the field of image processing. Similarly, in the field of medical image segmentation, to make up for the shortcomings of CNNs in establishing long-distance dependency relationships, Chen Jieneng and others first proposed a Transformer-based medical image segmentation framework [5], which uses CNN feature maps as the input to the Transformer encoder. This innovative method has achieved excellent results in medical image segmentation tasks, laying a solid foundation for subsequent research.

Subsequently, Liu Ze and others further promoted the development of this field by proposing a hierarchical visual Transformer called Swin Transformer [6]. Swin Transformer ingeniously combines the advantages of Transformers and CNNs, significantly reducing computational complexity through the use of a shifted window strategy while maintaining model performance. In various tasks in the field of computer vision, Swin Transformer has achieved state-of-the-art performance, demonstrating its strong generalization capabilities and practicality.

This paper will introduce the processing of medical image segmentation tasks by deep learning neural networks and compare Transformer-based network models with CNN-based network models. Experiments have proven that in MRI and CT medical image segmentation tasks, Transformer network models are superior to CNN network models. Due to the usual situation where medical images have a small number of labeled training samples, this paper proposes a novel Transformer network model for medical image segmentation tasks with few samples. Experiments have proven that this method can significantly improve the segmentation results of medical images, not only providing an effective solution to the current problem but also offering new ideas for solving other similar issues.

2. Related Work

Medical Medical image segmentation has experienced significant advancements through deep learning, particularly leveraging Convolutional Neural Networks (CNNs) and Transformer-based architectures. Traditional medical image segmentation methods relied on handcrafted features, which were time-consuming and inflexible in handling complex variations. The introduction of deep learning, especially with UNet and its variants, dramatically improved segmentation accuracy by utilizing hierarchical

feature extraction and skip connections [7]. However, CNNs have inherent limitations in capturing long-range dependencies, prompting researchers to explore Transformer-based architectures for medical imaging tasks.

The emergence of Transformer models, initially successful in Natural Language Processing (NLP), has influenced medical image analysis. The Vision Transformer (ViT) demonstrated how self-attention mechanisms could capture global dependencies effectively [8]. The hierarchical Swin Transformer further improved computational efficiency while maintaining segmentation performance [9]. These advancements encouraged researchers to apply multi-scale Transformer models to medical image classification, achieving superior results by balancing local and global dependencies [10]. Similarly, attention-enhanced UNet models have been explored to refine multi-scale semantic segmentation performance [11].

To address the issue of limited labeled medical data, hybrid CNN-Transformer models have been proposed, such as architectures for heart disease prediction using life history data [12]. Few-shot learning techniques and contrastive learning have also been introduced to enhance feature representation under data-scarce conditions [13]. Additionally, knowledge transfer strategies such as feature alignment in cross-domain knowledge extraction [14] and domain adaptation using LoRA fine-tuning have been explored to optimize large models for specialized tasks [15][16].

Beyond medical imaging, Transformer-based architectures have also been used in healthcare-related NLP tasks, such as medical text summarization with LongFormer models [17] and NLP-driven privacy-preserving solutions for medical records [18]. These studies highlight the adaptability of Transformer models across different applications. Additionally, Graph Neural Networks (GNNs) have been leveraged for hierarchical data mining, aiding in complex imbalanced data classification [19]. Other research has investigated hypergraph-based sequential visit prediction in electronic health records (EHR) analysis, showing promise in structured medical data applications [20].

Recent studies have also combined Transformer architectures with diffusion models for anomaly detection in medical images [21]. Additionally, dynamic adaptation techniques have been explored for optimizing large language models (LLMs), which may further support medical AI applications in fine-tuned segmentation tasks [22]. Time-series forecasting models integrating GNNs and Transformers have also demonstrated the effectiveness of deep learning techniques in healthcare analytics beyond static image processing [23].

In this work, we propose MS-UNet, a Transformer-based multi-scale nested decoder segmentation framework designed to mitigate the semantic gap between the encoder and decoder. By leveraging dense multi-scale nested decoders, MS-UNet enhances feature communication and learning capacity, achieving improved segmentation accuracy, particularly in scenarios with limited labeled training data. Experimental results on MRI and CT segmentation tasks validate that MS-UNet outperforms existing CNN and Transformer-based models, providing a robust solution for medical image segmentation in data-scarce conditions.

3. Background

The process of solving image segmentation problems based on deep learning mainly consists of three parts: neural network models, image preprocessing, and loss functions. Currently, the U-shaped structure based on convolutional neural networks (CNNs) is the most commonly used deep learning network model structure for image segmentation tasks. Its origin can be traced back to 2015 when Long et al.

proposed a fully convolutional network with an encoder-decoder structure, which discarded fully connected layers, used convolutional layers and deconvolution operations for upsampling, and combined skip connections to merge high and low-level feature maps, achieving pixel-level classification prediction.

Influenced by this, Ronneberger et al. further optimized the network structure and proposed a U-shaped network structure called U-Net, which has had a profound impact on the field of medical image segmentation, making it more suitable for medical image segmentation tasks.

The uniqueness of U-Net lies in its U-shaped structure and skip connections. The U-shaped structure consists of two parts: the encoder and the decoder, resembling the English letter "U." The encoder progressively downsamples the input image through continuous convolution and pooling operations to extract key feature maps. The decoder is responsible for upsampling these feature maps back to a size close to the original image through upsampling steps and performing pixel-level fine classification on this basis. This unique U-shaped architecture endows the network with the ability to absorb both shallow details and deep semantic information, thereby significantly enhancing the accuracy of image segmentation. The structure is shown in Figure 1.

Nowadays, with the success of Transformer neural networks in the field of natural language processing and other image processing domains, to address the shortcomings of convolutional neural networks in establishing long-distance dependency relationships, TransUNet [24], while maintaining the U-Net encoder-decoder structure, introduced Transformers and convolutional neural networks to form a hybrid encoder for the first time, achieving segmentation performance beyond models that use only convolutional neural networks.

Subsequently, the U-shaped encoder-decoder model based on Swin Transformer, SwinUnet [25], has achieved even better medical image segmentation performance than TransUNet.

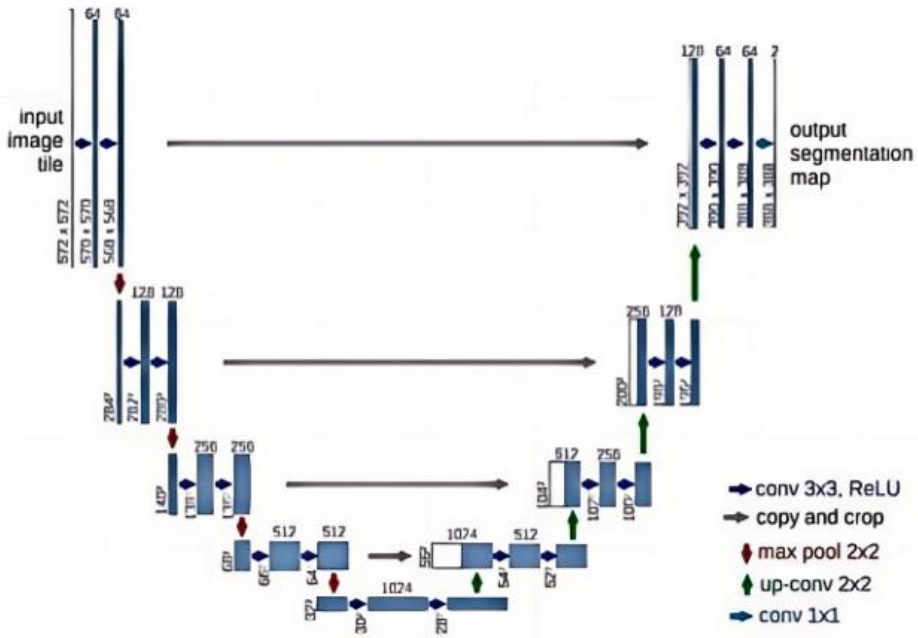


Figure 1. U-Net Network Structure

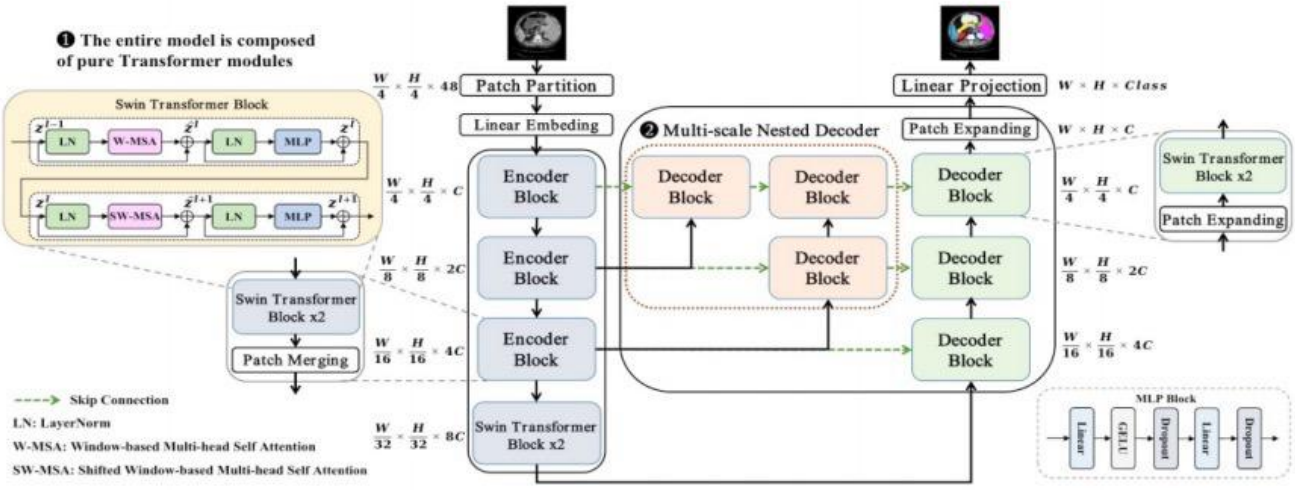


Figure 2. Overall Architecture of MS-UNet

4. Method

The overall structure of MS-UNet is depicted in Figure 2. It comprises an encoder, skip connections, and a multi-scale nested decoder, interconnected through skip connections. Like Swin-Unet, MS-UNet utilizes the Swin Transformer instead of CNN as the backbone network. Initially, MS-UNet segments the medical image into non-overlapping patches through a "Patch Partition" module and obtains raw value features by concatenating the original pixel values. Subsequently, MS-UNet projects the raw value features onto patch tokens of arbitrary dimensions using a linear embedding layer. During the encoder processing step, the transformed patch tokens are processed through a workflow composed of multiple Swin Transformer blocks and patch merging layers to generate hierarchical feature representations. In the decoder processing step, to address the semantic discrepancy between the encoder and decoder features in the feature fusion process of simple U-shaped segmentation network architectures, this paper innovatively employs a multi-scale nested decoder to independently upsample and decode features from each encoder. The skip connections concatenate the generated features with the corresponding decoder input features in the next layer.

This study posits that the novel multi-scale nested decoder can learn semantic information from the feature maps of the transformer encoder from a more dimensional perspective, resolving the semantic discrepancy issue between the encoder and decoder parts in the feature fusion process of simple U-shaped network structures. This effectively enhances the network's stability and generalization capability, enabling it to better learn the required information even when labeled data is scarce. Finally, MS-UNet outputs a resolution matching the input through an upsampled patch expansion layer and performs pixel-level segmentation prediction on the upsampled features using a linear projection layer.

4.1 Swin Transformer

Swin Transformer has achieved significant success in the field of computer vision in recent years, particularly in tasks such as image classification, object detection, and semantic segmentation. Its design combines the strengths of Convolutional Neural Networks (CNNs) and Transformer models, effectively capturing both local and global image information by introducing the multi-head self-attention module with shifted window mechanisms. Traditional Multihead Self-Attention (MSA) modules face challenges

in processing images due to high computational complexity, often making it difficult to maintain performance while achieving efficient computation. Swin Transformer addresses this issue by introducing Window-based MSA (W-MSA) and Shifted Window-based MSA (SW-MSA).

The design of the W-MSA module is inspired by the local perception characteristics of CNNs. It divides the image into non-overlapping windows and computes self-attention independently within each window. This not only reduces computational complexity but also allows the model to focus on local image features. The SW-MSA module, on the other hand, builds upon W-MSA by periodically shifting the positions of the windows, enabling information exchange between windows. This allows the model to capture global image information while maintaining its ability to extract local features.

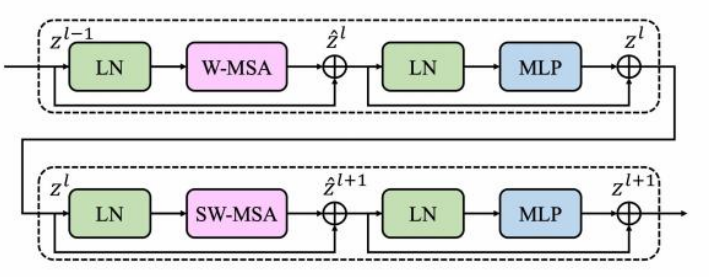


Figure 3. Swin Transformer Structure

Each layer of Swin Transformer consists of two distinct Swin Transformer modules, each with a relatively simple and efficient structure, namely Window-based MSA (W-MSA) and Shifted Window-based MSA (SW-MSA). Each Swin Transformer module also includes a LayerNorm (LN) layer, an MLP module with a GELU activation function, and residual connections. As shown in Figure 3, the computation of Swin Transformer composed of two consecutive Swin Transformer modules can be represented as:

$$\widehat{z}^l = W - MSA(LN(z^{l-1})) + z^{l-1} \qquad (2\text{-}1)$$

$$z^l = MLP\left(LN\left(\widehat{z}^l\right)\right) + \widehat{z}^l \qquad (2\text{-}2)$$

$$\widehat{z}^{l+1} = SW - MSA\left(LN\left(z^l\right)\right) + z^l \qquad (2\text{-}3)$$

$$z^{l+1} = MLP\left(LN\left(\widehat{z}^{l+1}\right)\right) + \widehat{z}^{l+1} \qquad (2\text{-}4)$$

Additionally, the computation of self-attention is as follows:

$$\text{Attention}\ (Q, K, V) = \text{SoftMax}\ \left(\frac{QK^T}{\sqrt{d}} + B\right)V \qquad (2\text{-}5)$$

In summary, Swin Transformer achieves efficient and accurate image processing by combining the local perception characteristics of CNNs with the global modeling capabilities of Transformers. Its unique shifted window design allows the model to capture both local and global image information fully while maintaining computational efficiency. This feature has made Swin-Transformer exhibit powerful performance in tasks such as image segmentation and object detection.

4.2 Encoder-Decoder Module

The encoder-decoder module is a crucial component in the network structure, with the encoder branch being a hierarchical network structure centered around Swin Transformer blocks. Within each Swin Transformer block, the patch merging layer is responsible for performing the critical down-sampling operation. This process not only effectively reduces the dimensionality of the input data but also retains key information in the image by increasing the feature dimension. This design strategy significantly expands the model's receptive field, enabling it to more comprehensively capture details and contextual information in the image. The Swin Transformer module plays an essential role in the encoder branch, allowing it to conduct in-depth representation learning from the input medical images, extracting intrinsic features and patterns from the images. This capability makes the Swin Transformer particularly suitable for processing complex medical image data, providing robust support for subsequent image processing tasks.

Complementing the encoder branch is the decoder branch, which is also based on the Swin Transformer backbone, supplemented by upsampling modules and Patch Expanding layers, together forming a complete network architecture. In traditional U-shaped models, the encoder and decoder modules are typically connected through simple skip connections. However, in this study, an innovative multi-scale nested decoder branch is proposed to replace the decoder branch in traditional U-shaped networks. With this design, the decoder network can effectively learn more information from the more complex and richer Swin Transformer encoder output characteristics. The multi-scale nested blocks perform upsampling on the decoding features from each encoder and use skip connections to concatenate the generated features with the corresponding decoder in the next layer. This connection method not only promotes the effective fusion of multi-scale features from the encoder with upsampled features but also strengthens the feature communication and sharing between adjacent decoders.

Overall, the multi-scale nested decoder branch optimizes the decoding process in traditional U-shaped networks, allowing the network to draw rich feature information from more complex Swin Transformer encoder outputs. This design not only enhances the model's representational capacity but also provides robust support for handling challenging medical image processing tasks.
The core of the encoder branch is a hierarchical network architecture, meticulously constructed from Swin Transformer modules. In each Swin Transformer module, the patch merging layer plays an indispensable role. This layer is responsible for the down-sampling operation, a step that enhances the representational capacity of features while reducing data dimensions. Through this design, the model can effectively capture key information in the image and significantly expand its perceptual range.

These characteristics make the Swin Transformer module particularly suitable for in-depth learning of the complex and subtle features of medical images, laying a solid foundation for subsequent image processing tasks.

5. Experiment

5.1 Dataset

The Synapse Lower Abdominal Multi-Organ CT Segmentation Dataset is a valuable resource for medical image analysis and computer-assisted intervention research. It originated from the Medical Image Computing and Computer Assisted Intervention (MICCAI) Multi-Atlas Abdominal Labeling Challenge held in 2015. Under the supervision of an Institutional Review Board (IRB), the organizers

randomly selected 50 abdominal CT scan images from an ongoing rectal cancer chemotherapy trial and a retrospective study of inguinal hernias.

This dataset is particularly useful for developing and evaluating deep learning models for multi-organ segmentation in the lower abdomen from CT scans, which is crucial for various clinical applications such as surgical planning, disease diagnosis, and treatment assessment.

5.2 Experiment Results Analysis

In this paper, three representative models are selected as the baseline models for the experiment: UNet, which is based solely on the Convolutional Neural Network (CNN) architecture, TransUNet, a hybrid model that combines CNN with Transformer networks, and SwinUnet, which is based solely on the Transformer network architecture. To comprehensively evaluate the performance of MS-UNet, this paper selects three typical U-shaped models from open sources as benchmark models for comparison. These benchmark models have certain representativeness and superiority in image segmentation tasks.

Table 1: Performance of experiments

Method	DSC↑(%)	HD↓(mm)
R50 U-Net	74.68	36.87
R50 Att-UNet	75.57	36.97
Att-UNet	77.77	36.02
R50 ViT	71.29	32.87
MT-UNet	78.59	26.59
UNet	77.22±0.41	29.31±0.85
TransUNet	77.07±0.34	31.41±1.27
SwinUnet	78.45±0.49	25.69±1.28
MS-UNet	79.97±0.20	21.95±2.11

In the more complex Synapse dataset, to further highlight the superiority of the baseline models and MS-UNet, this paper also includes segmentation results from five models from the articles TransUnet and MT-UNet [26] as comparative data. These additional selected models have achieved significant results in their respective fields and possess a certain level of competitiveness.

Table 1. presents the average DSC (Dice Similarity Coefficient) and average HD (Hausdorff Distance) metrics of different methods on the Synapse multi-organ CT dataset.

After a comprehensive analysis of the experimental results on the Synapse multi-organ CT dataset, it is evident that the proposed MS-UNet model has achieved significant performance improvements in medical image segmentation tasks. Specifically, as shown in Table 1, on the Synapse multi-organ CT dataset, MS-UNet achieved the best results in both average DSC and average HD metrics compared to other baseline models. Notably, when compared with the SwinUnet model, MS-UNet improved the DSC metric by 1.21% and reduced the HD metric by 6.51mm, demonstrating MS-UNet's exceptional performance in handling complex medical images.

6. Conclusion

The simplicity of feature transfer in this structure may result in a semantic gap between the encoder and decoder, especially in Transformer models where their powerful global information capture capabilities could exacerbate this semantic discrepancy, thereby affecting segmentation performance. To address this limitation, this paper innovatively proposes a Transformer-based multi-scale nested decoder segmentation network framework—MS-UNet. This framework enhances semantic communication between modules by designing a dense multi-scale nested decoder, thereby improving the model's feature learning ability and network performance. Experimental comparative results have convincingly demonstrated the effectiveness of the MS-UNet design, particularly under conditions of extremely limited training data, where MS-UNet has achieved significant improvements in segmentation outcomes compared to other advanced models in the U-Net series. This innovation provides a new solution for the field of medical image segmentation, offering particular value in scenarios where data is scarce.

References

[1] Xiao H, Li L, Liu Q, et al. Transformers in medical image segmentation: A review. Biomedical Signal Processing and Control, 2023, 84: 104791.

[2] Qureshi I, Yan J, Abbas Q, et al. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. Information Fusion, 2023, 90: 316-352.

[3] Huang H, Lin L, Tong R, et al. Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 1055-1059.

[4] Yin H, Vahdat A, Alvarez J M, et al. A-vit: Adaptive tokens for efficient vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10809-10818.

[5] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.

[6] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[7] Li, X., Lu, Q., Li, Y., Li, M., & Qi, Y. (2025). Optimized Unet with Attention Mechanism for Multi-Scale Semantic Segmentation. arXiv preprint arXiv:2502.03813.

[8] Hu, J., Xiang, Y., Lin, Y., Du, J., Zhang, H., & Liu, H. (2025). Multi-Scale Transformer Architecture for Accurate Medical Image Classification. arXiv preprint arXiv:2502.06243.

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

[10] Wang, J. (2024). Multivariate Time Series Forecasting and Classification via GNN and Transformer Models. Journal of Computer Technology and Software, 3(9).

[11] Qi, Y., Lu, Q., Dou, S., Sun, X., Li, M., & Li, Y. (2025). Graph Neural Network-Driven Hierarchical Mining for Complex Imbalanced Data. arXiv preprint arXiv:2502.03803.

[12] Hao, R., Xiang, Y., Du, J., He, Q., Hu, J., & Xu, T. (2025). A Hybrid CNN-Transformer Model for Heart Disease Prediction Using Life History Data. arXiv preprint arXiv:2503.02124.

[13] Hu, J., An, T., Yu, Z., Du, J., & Luo, Y. (2025). Contrastive Learning for Cold Start Recommendation with Adaptive Feature Fusion. arXiv preprint arXiv:2502.03664.

[14] Li, P. (2024). Improved Transformer for Cross-Domain Knowledge Extraction with Feature Alignment. Journal of Computer Science and Software Applications, 5(2).

[15] Liao, X., Wang, C., Zhou, S., Hu, J., Zheng, H., & Gao, J. (2025). Dynamic Adaptation of LoRA Fine-Tuning for Efficient and Task-Specific Optimization of Large Language Models. arXiv preprint arXiv:2501.14859.

[16] Liao, X., Zhu, B., He, J., Liu, G., Zheng, H., & Gao, J. (2025). A Fine-Tuning Approach for T5 Using Knowledge Graphs to Address Complex Tasks. arXiv preprint arXiv:2502.16484.
[17] Sun, D., He, J., Zhang, H., Qi, Z., Zheng, H., & Wang, X. (2025). A LongFormer-Based Framework for Accurate and Efficient Medical Text Summarization. arXiv preprint arXiv:2503.06888.
[18] Zhu, Z., Zhang, Y., Yuan, J., Yang, W., Wu, L., & Chen, Z. NLP-Driven Privacy Solutions for Medical Records Using Transformer Architecture.
[19] Gao, J., Lyu, S., Liu, G., Zhu, B., Zheng, H., & Liao, X. (2025). A Hybrid Model for Few-Shot Text Classification Using Transfer and Meta-Learning. arXiv preprint arXiv:2502.09086.
[20] Gao, Z., Mei, T., Zheng, Z., Cheng, X., Wang, Q., & Yang, W. (2024, September). Multi-Channel Hypergraph-Enhanced Sequential Visit Prediction. In: 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS) (pp. 421-425). IEEE.
[21] Wang, X. (2025). Data Mining Framework Leveraging Stable Diffusion: A Unified Approach for Classification and Anomaly Detection. Journal of Computer Technology and Software, 4(1).
[22] Wu, L., Gao, J., Liao, X., Zheng, H., Hu, J., & Bao, R. Adaptive Attention and Feature Embedding for Enhanced Entity Extraction Using an Improved Bert Model.
[23] Deng, Y. (2025). A hybrid network congestion prediction method integrating association rules and LSTM for enhanced spatiotemporal forecasting. Transactions on Computational and Scientific Methods, 5(2).
[24] GB/T 7714 Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306, 2021.
[25] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 205-218.
[26] Jha A, Kumar A, Pande S, et al. Mt-unet: a novel u-net based multi-task architecture for visual scene understanding. In: 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 2191-2195.