Swin-UNet: A Unified Transformer–CNN Framework for Multi-Organ Medical Image Segmentation

Xiuru Li*

School of Information Science and Engineering, Lanzhou University, Lanzhou, China

Abstract. Transformer-based architectures have demonstrated significant promise in medical image segmentation due to their strong ability to model long-range contextual relationships. However, standard Vision Transformer (ViT) mod- ules used in hybrid networks such as TransUNet are limited in representing both fine-grained and coarse features effectively. To overcome this limitation, this paper introduces Swin-UNet, a hybrid framework that combines the hierarchical Swin Transformer encoder with a U-Net-inspired decoder. The encoder utilizes shifted-window self-attention for efficient local-global feature learning, while the decoder integrates residual convolutional paths and multi-scale patch embeddings for improved reconstruction and scale robust- ness. Evaluated on the Synapse multi-organ CT dataset, the model achieves competitive Dice scores and lower Hausdorff distances compared to U-Net and TransUNet, highlighting its potential as a robust and generalizable approach for medical image segmentation. These results suggest that the Swin-UNet effectively balances computational efficiency with segmentation accuracy, offering a strong foundation for future medical imaging applications.

1 Introduction

Deep learning techniques have substantially advanced the field of medical image segmentation, enabling higher precision and automation in clinical analysis. Among various architectures, Convolutional Neural Networks (CNNs) and their derivatives—including Fully Convolutional Networks (FCNs) [1] and the U-Net model [2]—serve as the foundation for most segmentation frameworks. The U-Net architecture, characterized by its symmetric encoder–decoder structure and skip connections, effectively preserves fine spatial information while integrating semantic context. Successive variants, such as UNet++ [3], further enhanced boundary delineation and multi-scale feature fusion. Despite these advancements, CNN-based approaches inherently rely on localized convolution kernels, which limit their ability to model global relationships across distant image regions.

To overcome this inherent locality restriction, numerous studies have introduced attention mechanisms into CNNs to extend their receptive fields. While such hybrid CNN–attention architectures improve contextual perception, their effectiveness remains constrained by

*Corresponding author's email: lxiuru2023@lzu.edu.cn

computational complexity and local processing scope. Transformer-based models, originally developed for sequence modelling in natural language processing, have been successfully transferred to vision tasks [4], offering superior capability in global context modelling. Building on this trend, TransUNet [5] incorporated Transformer blocks into a CNN framework, achieving end-to-end medical image segmentation with improved contextual awareness. Subsequent research, including MedT [6], observed that standard Vision Transformers often struggle to encode multi-level representations due to fixed patch sizes and single-scale attention mechanisms.

To address these challenges, this work introduces Swin-UNet, a hybrid segmentation network that employs the hierarchical Swin Transformer [7] as an encoder within a U-Net-like architecture. The Swin Transformer applies a shifted-window self-attention strategy, which facilitates both local feature learning and inter-window information exchange. Combined with multi-scale patch embeddings and residual convolutional decoding, this design achieves stable gradient propagation and enhanced feature integration. The proposed framework thus strikes a balance between global contextual reasoning and fine structural reconstruction, leading to more accurate multi-organ segmentation.

2 Method

2.1 Dataset preparation

The Synapse multi-organ CT dataset [5] was adopted for training and evaluation. It contains 30 volumetric abdominal CT cases, with 18 used for training and 12 for validation/testing. Each case was manually annotated by radiologists to label organs including the liver, kidneys, spleen, and pancreas. Each CT volume comprises around 180–200 slices at 512×512 resolution. To ensure data uniformity, the scans were resampled into isotropic voxels and normalized to the range [0, 1]. All slices were resized to 224×224 to fit the Swin Transformer input, and grayscale images were replicated into three channels. Data augmentation techniques, including random rotations, flips, elastic distortions, Gaussian noise, and contrast adjustments, were used to enhance generalization.

2.2 Model overview

Swin-UNet combines a Swin Transformer encoder and a U-Net-style decoder. The encoder employs hierarchical shifted-window multi-head self-attention (SW-MSA), enabling local and global feature learning. This paper adopts the Swin-Tiny configuration with an embedding dimension of 96, depths [2, 2, 6, 2], and heads [3, 6, 12, 24]. The decoder progressively upsamples and merges features from the encoder via skip connections, followed by convolution and normalization layers for spatial reconstruction. The final segmentation head applies a 1×1 convolution to produce class probabilities (Fig. 1).

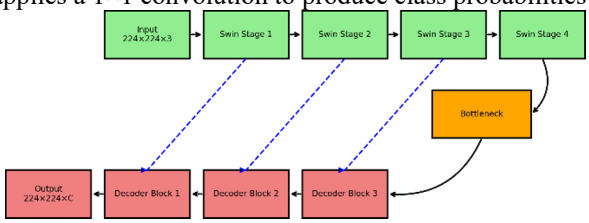

Fig. 1. Overall architecture of the proposed Swin-UNet. The encoder uses hi- erarchical shifted-window attention, and the decoder reconstructs detailed spatial features through skip connections (Picture credit: Original).

2.3 Training details

Training of the proposed network was implemented using PyTorch with GPU acceleration provided by NVIDIA hardware. The optimization process extended across 150 epochs, involving approximately 30,000 parameter updates. Each GPU processed 24 CT slices per batch. Model weights were updated using stochastic gradient descent, incorporating a momentum of 0.9 and a decay rate of 1×10^{-4}. The learning rate, initialized at 0.01, followed a polynomial decay schedule to ensure smooth convergence.

To enhance both regional segmentation accuracy and edge localization, a hybrid objective function was employed, combining Dice and cross-entropy losses [8, 9]. Quantitative evaluation utilized the Dice Similarity Coefficient (DSC) [10] and the 95th percentile Hausdorff Distance (HD95) as complementary metrics to assess overlap quality and boundary precision. The evolution of training loss and validation performance was continuously monitored using TensorBoard to facilitate performance tracking and stability analysis.

3 Results and discussion

3.1 Quantitative evaluation

The proposed Swin-UNet was quantitatively evaluated on the Synapse dataset using Dice and HD95 metrics. Compared with U-Net, TransUNet, and MedT, the model achieved a competitive Dice score of 0.802 and a substantially improved HD95 of 24.69 mm. As shown in Table 1, Swin-UNet demonstrates superior boundary precision while maintaining comparable overlap performance, confirming the effectiveness of its hierarchical attention mechanism and multi-scale feature integration strategy. These results indicate that Swin-UNet successfully balances segmentation accuracy and structural consistency, making it a robust alternative for multi-organ medical image segmentation tasks.

Table 1. Quantitative comparison on the Synapse dataset. The proposed model improves boundary precision while maintaining strong overall segmentation accuracy.

Method	Mean Dice (%)	Mean HD95 (mm)
U-Net	0.812	32.12
TransUNet	0.795	27.54
MedT	0.798	26.81
win-UNet (The proposed model)	**0.802**	**24.69**

3.2 Qualitative evaluation

Visual inspection of segmentation maps indicates that Swin-UNet produces smooth and coherent organ contours, even in small or low-contrast regions such as the pancreas. The model demonstrates strong robustness to noise and intensity variations while maintaining consistent segmentation across adjacent slices. Minor misclassifications observed at ambiguous tissue boundaries suggest opportunities for further enhancement, such as integrating boundary-aware attention modules or refining post-processing strategies to improve structural delineation accuracy.

3.3 Analysis and discussion

The combination of hierarchical self-attention and convolutional decoding enables the network to balance fine-grained reconstruction with broad contextual reasoning. Compared

to CNN-only or Transformer-only architectures, Swin-UNet achieves a more favorable trade-off among segmentation accuracy, computational efficiency, and generalization capability. The hybrid loss formulation contributes to stable optimization and consistent performance across multiple organ classes. However, the computational requirements remain relatively high for real-time clinical applications. Future work could focus on optimizing model efficiency through attention pruning, lightweight transformer designs, or domain adaptation techniques to facilitate faster inference and improved adaptability in practical deployment scenarios.

4 Conclusion

This paper presented Swin-UNet, a Transformer–CNN hybrid architecture for multi-organ CT segmentation. The network integrates hierarchical attention from the Swin Transformer with residual convolutional decoding to enhance both global context understanding and local detail reconstruction. Experiments on the Synapse dataset demonstrated that the proposed method achieves reliable segmentation accuracy with strong boundary alignment and robustness across organs. In future work, lightweight variants and self-supervised learning strategies will be investigated to enhance model adaptability, computational efficiency, and clinical applicability in real-world medical imaging environments.

References

1. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440 (2015)
2. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241 (Springer, 2015)
3. Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested U-Net architecture for medical image segmentation, *International Workshop on Deep Learning in Medical Image Analysis*, 3–11 (Springer International Publishing, Cham, 2018)
4. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16×16 words: Transformers for image recognition at scale, *International Conference on Learning Representations (ICLR)* (2021)
5. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, Y. Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, *arXiv preprint* arXiv:2102.04306 (2021)
6. Q. Qi, L. Lin, R. Zhang, C. Xue, MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis, *IEEE Access* **10**, 28750–28759 (2022)
7. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows, *IEEE International Conference on Computer Vision (ICCV)*, 10012–10022 (2021)
8. R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, Y. Pan, Rethinking dice loss for medical image segmentation, *IEEE International Conference on Data Mining (ICDM)*, 851–860 (IEEE, 2020)

9. A. Mao, M. Mohri, Y. Zhong, Cross-entropy loss functions: Theoretical analysis and applications, *International Conference on Machine Learning (ICML)*, 23803–23828 (PMLR, 2023)
10. P.L. Yeap, Y.M. Wong, A.L.K. Ong, J.K.L. Tuan, E.P.P. Pang, S.Y. Park, H.Q. Tan, Predicting dice similarity coefficient of deformably registered contours using Siamese neural network, *Phys. Med. Biol.* **68**(15), 155016 (2023)