Hierarchical Feature Aggregation Network Based on Swin Transformer for Medical Image Segmentation

Hayato Iyoda[1], Yongqing Sun[2], and Xian-Hua Han[1]

[1] Graduate School of Artificial Intelligence and Science, Rikkyo University
3-34-1 Nishi-Ikebukuro, Toshima-ku, Tokyo 171-8501, Japan
`[24vr038t, hanxhua]@rikkyo.ac.jp`
[2] College of Humanities and Sciences, Nihon University
3-25-40 Sakurajosui Setagaya-Ku, Tokyo 156-8550 JAPAN

Abstract. Semantic segmentation plays a crucial role in computer-aided medical image analysis by achieving important and useful regions, which are vital for various diagnostic tasks. Recently, vision transformers (ViTs) have emerged as the leading approach in medical image segmentation, outperforming traditional convolutional neural networks (CNNs). The incorporation strategies of the ViTs for medical segmentation are dominated to leverage the widely used U-shape like architecture (U-Net) while replace the convolution blocks in both encoder and decoder paths using transformer blocks. It remains uncertain which components of the incorporated transformer block contribute most significantly to segmentation results in the medical field. This study presents a hierarchical feature aggregation method based on hierarchical Transformer features to enhance the performance of ViT-based architecture in data-constrained medical image segmentation. Specifically, our approach employs the hierarchical vision Transformer to configure the main encoder path for extracting multi-scale semantic features, and leverages several residual blocks to achieve local representation with detail spatial information. Then, we introduce a hierarchical feature aggregation module (HFAM) to serve as the decoder path for fusing multi-scale semantic features and residual spatial features. Compared with the existing transformer-based U-Net, the explored HFAM can not only effectively combine the diverse contexts but also potentially reduce the computational complexity. Experiments on 3 different medical image segmentation benchmarks have demonstrated our proposed method consistently outperformers the conventional U-Net, and various Transformer-based U-Net.

Keywords: Medical image segmentation · U-Net · Swin Transformer · Hierarchical feature aggregation

1 Introduction

Semantic segmentation is a fundamental process in computer-aided medical image analysis, serving the critical function of identifying and delineating regions of interest for various diagnostic tasks [17,20,21]. Medical segmentation, however, is often complicated due to variations in image modality, acquisition techniques, and inherent pathological or biological differences across patients. Such complexities introduce significant challenges to achieving accurate and reliable segmentation. Recently, the application of deep learning techniques has provided substantial advancements in addressing these challenges. Of particular significance is the introduction of the U-Net model [19], which has proven to be remarkably effective in medical image segmentation tasks. Thus, U-Net [19] and its numerous variants [6,7,11,13,14,16,30,32] have become the prevailing standard in many medical image segmentation tasks such as cardiac segmentation, organ segmentation, lesion segmentation and so on. The U-Net architecture is generally designed with a symmetric encoder-decoder structure with the dominated convolution components, and attempt to capture global context by creating large receptive fields with down-sampling and stacking multiple convolutional layers.

Despite the strong representational power of CNN-based U-net, they face challenges in establishing clear long-range dependencies because convolutional kernels have restricted receptive fields. This inherent limitation in the convolution operation makes it difficult to capture global semantic context [2], which is vital for dense prediction tasks such as segmentation. Motivated by the attention mechanism [22,25] in natural language processing, recent research addresses the limitations of CNNs by integrating attention into their architecture. For instance, Non-local neural networks [18] introduce a plug-and-play operator based on self-attention, allowing them to capture long-range dependencies within feature maps. However, this comes at the cost of significant memory and computational demands. Schlemper et al. [22] offer an alternative with the attention gate model, which enhances model sensitivity and prediction accuracy while introducing minimal computational overhead, making it easily adaptable to standard CNNs. In contrast, the Transformer architecture [5,25] is explicitly designed to handle long-range dependencies in sequence-to-sequence tasks, capturing relationships between any positions within a sequence. Recently, researchers have explored the application of Transformers in computer vision. The Vision Transformer (ViT) [5] was developed to tackle image recognition tasks by using 2D image patches with positional embeddings and pre-training on large datasets. ViT achieved performance comparable to CNN-based models. Furthermore, the Data-efficient Image Transformer (DeiT) [23] demonstrated that Transformers could be trained on mid-sized datasets, and its performance could be enhanced through distillation techniques. Additionally, the Swin Transformer [15], a hierarchical architecture, was later proposed as a vision backbone, achieving state-of-the-art results in image classification, object detection, and semantic segmentation. The successes of ViT, DeiT, and the Swin Transformer highlight the growing potential of Transformer models in computer vision applications.

In medical image segmentation filed, the incorporation of Transformers into U-Net architecture has extensively explored, and led to advancements in segmentation accuracy [1,24,26,27,31], especially in tasks that require precise delineation of complex structures. By incorporating Transformer blocks into the encoder and decoder paths, these Transformer-based U-Net enhances the ability to capture both local and global features. For example, Swin-UNet [10] uses Swin Transformer blocks within the U-Net structure to improve segmentation performance on 2D medical images by modeling long-range dependencies while maintaining spatial resolution through skip connections. Swin-Unet is a fully Transformer-based U-shaped architecture, incorporating the Swin Transformer block into all components: encoder, bottleneck, decoder, and skip connections. Despite the potential performance gain, the incorporation of the Transformer blocks into all components may cause high computational cost and memory usage, and possibly brings the overfitting problem especially for data-constrained medial image analysis tasks. Moreover, the impact of incorporating Transformer blocks into various components of the network on overall performance remains underexplored, with limited research addressing how these modifications influence segmentation accuracy and computational efficiency.

To handle the above issues, this study presents a novel hierarchical feature aggregation framework, leveraging hierarchical Transformer features to enhance the performance of Vision Transformer (ViT)-based architectures for medical image segmentation in data-limited scenarios. Our methodology centers on utilizing a hierarchical Vision Transformer as the primary encoder, which facilitates the extraction of multi-scale semantic features, crucial for capturing global context. Additionally, residual blocks are integrated to preserve fine-grained local representations and detailed spatial information, addressing the need for precise segmentation boundaries. To optimize feature fusion, we introduce a Hierarchical Feature Aggregation Module (HFAM) within the decoder, which effectively merges multi-scale semantic features with the residual spatial information. Compared to existing Transformer-based U-Net variants, the proposed HFAM not only efficiently combines rich contextual information but also demonstrates potential in reducing computational complexity. Extensive experiments conducted on three benchmark datasets for medical image segmentation consistently show that our approach surpasses both conventional U-Net architectures and several Transformer-based U-Net models, highlighting its efficacy and robustness in challenging segmentation tasks.

2 Related Work

CNN-based methods: Motivated by the great success of the development of deep learning, convolutional neural networks (CNNs) have widely applied for many medical image segmentation tasks. A pivotal work was the introduction of the U-Net architecture [19], specifically designed for biomedical image segmentation. The U-shaped architecture in U-Net [19], characterized by its encoder-decoder structure with skip connections, enabled both efficient feature extraction and precise localization, making it highly effective for segmentation tasks. Due to its simplicity and strong performance, the U-Net framework has inspired numerous variants aimed at further enhancing its capabilities. Notable examples include Res-UNet [30], which incorporates residual connections to address gradient vanishing in deeper networks, Dense-UNet [14], which leverages dense connections to improve feature reuse and network efficiency, and U-Net++ [32], which refines the skip connections with nested architectures for better feature fusion. Additionally, UNet3+ [11] extends this design by introducing a full-scale skip connection mechanism, enabling a richer fusion of semantic and spatial features. U-Net and its variants have also been adapted for 3D medical image segmentation tasks, with architectures such as 3D U-Net [3] and V-Net [16] emerging to tackle volumetric data. These 3D models preserve the spatial coherence of medical images across slices, thereby improving segmentation accuracy in three-dimensional imaging modalities like CT and MRI. Overall, CNN-based methods, especially U-Net and its derivatives, have achieved remarkable success in medical image segmentation due to their powerful representation learning capabilities, adaptability to various tasks, and ability to handle both 2D and 3D medical imaging data. These advancements have significantly improved segmentation performance across a wide range of medical applications. However, all these methods generally employ the convolution layers as the dominated components, and have limited receptive fields to capture global context.

Vision transformers: The Transformer model was originally introduced for machine translation tasks [25] and has since revolutionized the field of natural language processing (NLP). Transformer-based models have achieved state-of-the-art results across a wide range of NLP tasks [4], owing to their ability to capture long-range dependencies and model complex relationships through self-attention mechanisms. Inspired by the success of Transformers in NLP, researchers extended this architecture to the field of computer vision, leading to the development of the Vision Transformer (ViT) [5], marked a significant breakthrough in image recognition by offering an impressive balance between speed and accuracy, especially in large-scale tasks. Unlike CNN-based models, ViT relies on global self-attention, which allows it to model global context more effectively. However, a notable limitation of ViT is its reliance on large-scale datasets for pre-training. Unlike CNNs, which can be efficiently trained on smaller datasets, ViT requires extensive pre-training on datasets such as ImageNet to achieve competitive performance. To address this challenge, the Data-efficient Image Transformer (DeiT) [23] introduced several training techniques, including knowledge distillation, to enable ViT to perform well on mid-sized datasets, mitigating the need for vast amounts of data.

Building on ViT's foundations, a series of subsequent works [8, 15, 28] have further enhanced Transformer-based architectures for vision tasks. Among these, the Swin Transformer [15] stands out as a highly efficient and versatile model. The Swin Transformer introduces a hierarchical structure with shifted window-based attention, which enables the model to capture both local and global information in a computationally efficient manner. This design significantly reduces the complexity typically associated with full self-attention, making Swin Transformer scalable and suitable for high-resolution inputs. As a result, it has achieved state-of-the-art performance across various computer vision tasks. Swin Transformer's ability to balance computational efficiency with high performance makes it an influential architecture in both research and practical applications within the vision domain. Most Transformers are originally proposed as the encoder in image classification tasks to extract different image representation, and require to configure a decoder to integrate the multi-scale encoder features for dense prediction tasks such as semantic image segmentation.

Transformers for Medical Image Segmentation: The success of ViT in traditional computer vision tasks has paved the way for a paradigm shift in medical image segmentation, and the integration of the Transformer block into the U-Net achitectires have increasingly been explored [1, 24, 26, 27, 31]. Among these advancements, TransUNet [2] represents the first framework to incorporate Transformers into medical image segmentation. It leverages the strengths of both CNNs and Transformers by combining the local feature extraction capabilities of CNNs with the global context modeling power of Transformers. Additionally, Valanarasu et al. [24] proposed the Gated Axial-Attention model (MedT), specifically designed to address the challenge of limited medical image data by incorporating attention mechanisms that are computationally more efficient and less data-intensive. Cao et al. [10] proposed Swin-Unet, the first pure Transformer-based U-shaped architecture for medical image segmentation. This model replaces traditional convolutional blocks with Swin Transformer layers, allowing for hierarchical and multiscale feature extraction while maintaining the U-Net's core encoder-decoder structure. However, the naive displacement of convolutional blocks with Swin Transformer blocks in both encoder and decoder paths may lead to structural redundancy and excessive computational overhead, without fully capitalizing on the strengths of Transformer encoding capability. Moreover, the impact of integrating Transformer blocks into different paths of the U-Net for medical image segmentation remains largely underexplored. Limited research has investigated how these architectural modifications affect key performance metrics, such as segmentation accuracy and computational efficiency.

3 Proposed Method

This study employs the Swin Transformer and simple ResBlocks to serve as dual branches of Encoder, and proposes a hierarchical feature aggregation module to server as the decoder for fusing various features learned in the Encoder. The overall framework is dubbed as hirachical feature aggregation network (HFANet), and the architecture is depicted in Fig. 1. Specifically, similar as the conventional U-Net, HFANet comprises three components: Encoder, decoder, and the skip connection bridging the interaction between the Encoder and Decoder. The Encoder path aims to incorporate Transformer blocks and convolution operation to extract both high-level semantic contexts and low-level detailed spatial structures. The first branch utilizes a Transformer architecture, initiating with window-based self-attention to model long-range dependencies and achieve multi-scale semantic contexts. Concretely, we simply adopt the Swin Transformer proposed for generic vision task [15] to server as one branch of the Encoder, where

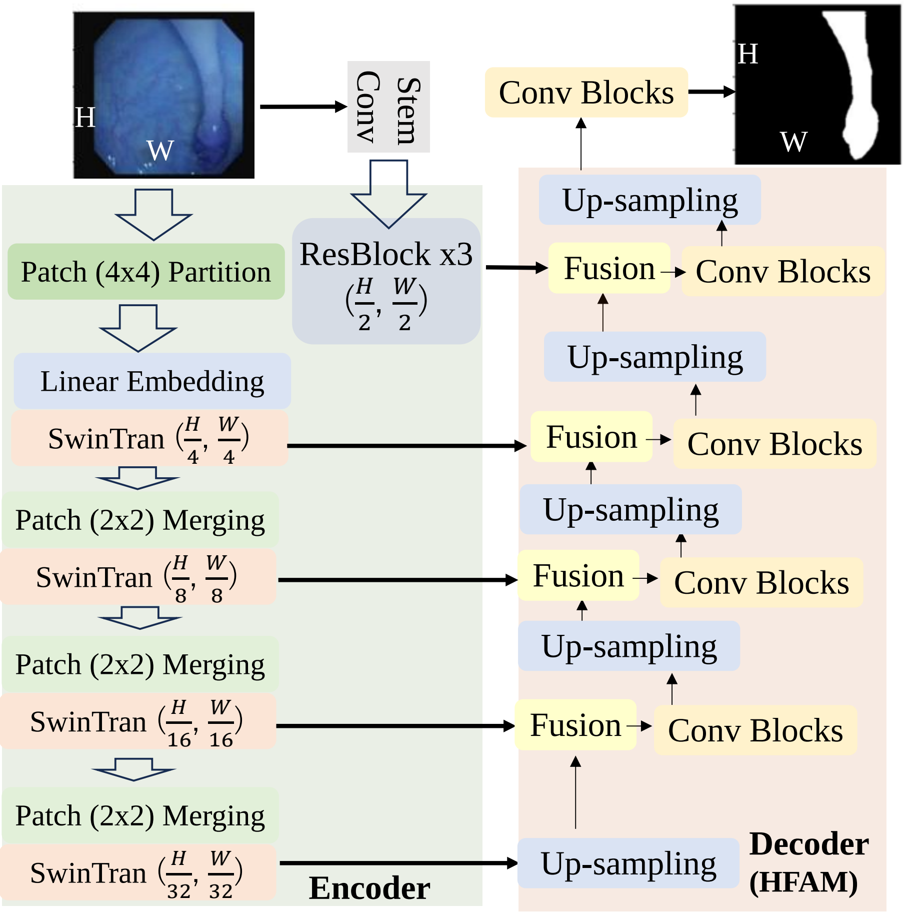

Fig. 1: Proposed hierarchical feature aggregation network based on Swin Transformer.

the extracted feature even with the highest spatial resolution is reduced to $\frac{1}{4}$ of the input resolution, causing sptail detail lost for accurate segmentation. Thus, the second path employ the simple Resblocks taken from the stem convolution and the first stage of Resblock in the ResNet50 [9] to extract detailed spatial structure, which can provide the complementary information to the Transformer branch. Then, we transfer the multi-scale contexts from the Transformer branch and the detailed spatial structure from the Resblock branch to the Decoder path via skip connections, and investigate a hierarchical feature aggregation module (HFAM) as the decoder to fuse all features for final segmentation prediction. Particularly, the HFAM utilizes the simple operations such as the up-sampling, channel concatenation and convolutions to reduce the computational complexity. Next, we will give the detailed descriptions of the Encoder and Decoder of our proposed HFANet.

3.1 Encoder

The encoder architecture contains the main Swin Transformer branch and the complementary Resblock branch.
Transformer branch: Transformer branch includes the 4 levels of Swin Transformer blocks, and each level has two blocks. The tokenized data with C-dimensional vectors and the reduced spatial resolution $\frac{1}{4}$ of the input data is firstly processed by two successive Swin Transformer blocks to extract the first level of semantic context $\mathbf{F}_1 \in \Re^{C\times W/4\times H/4}$. These blocks facilitate representation learning while preserving both the dimensionality and spatial resolution of the features. Then, before pass $\mathbf{F}_1$ to the second level of Transformer block, a patch merging layer operates to downsample the input by a factor of 2, thereby reducing the number of tokens while simultaneously expanding the feature dimensionality to twice its original size as $\mathbf{F}_2 \in \Re^{2C\times W/8\times H/8}$. This token reduction and feature enhancement process is iterated three times throughout the encoder, progressively refining the representation at each stage. Finally, we obtain the multi-scale semantic features as $\mathbf{F} = [\mathbf{F}_1, , \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4]$, all of which will be transferred to the Decoder for aggregation. Subsequently, we present the detailed explanations of the Swin Transformer block.

Each level of the Transformer branch contains two consecutive Swin transformer blocks. In contrast to the conventional multi-head self-attention (MSA) mechanism, the Swin Transformer block [15] is designed based on a shifted window paradigm, and comprises several components: a LayerNorm (LN) layer, a multi-head self-attention mechanism, a residual connection, and a two-layer multilayer perceptron (MLP) incorporating GELU activation. The first block leverages a window-based multi-head self-attention (WMSA) mechanism, while the second block employs a shifted window-based multi-head self-attention (SWMSA) mechanism. This successive Swin Transformer blocks with the above window partitioning strategy for $l-th$ level can be formulated as:

$$\hat{\mathbf{F}}_l^1 = WMSA(LN(\mathbf{F}_l)) + \mathbf{F}_l, \quad \mathbf{F}_l^1 = MLP(LN(\hat{\mathbf{F}}_l^1)) + \hat{\mathbf{F}}_l^1. \tag{1}$$

$$\hat{\mathbf{F}}_l^2 = SWMSA(LN(\mathbf{F}_l^1)) + \mathbf{F}_l^1, \quad \mathbf{F}_l^2 = MLP(LN(\hat{\mathbf{F}}_l^2)) + \hat{\mathbf{F}}_l^2, \tag{2}$$

where , $\hat{\mathbf{F}}_l^1$ and $\mathbf{F}_l^1$ denote the results of the WMSA and MLP module in the first block while $\hat{\mathbf{F}}_l^2$ and $\mathbf{F}_l^2$ refer to the output of the SWMSA and MLP module in the second block, respectively. The self-attention mechanism in the WMSA and SWMSA modules is calculated using the following equation:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = SoftMax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \tag{3}$$

where, $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V} \in \Re^{S^2\times d}$ refer to the query, key, and value matrices. d signifies the dimensionality of the query or key while S^2 represents the number of patches within a window.
Resblock branch: Since the Transformer branch produce multi-scale features $\mathbf{F}$ even with the highest spatial resolution $W/4 \times H/4$, and thus may result in

potential loss of detailed spatial structure. This study attempts to incorporate a simple convolution-based branch for compensating the lost structures in the Transformer branch. Specifically, we take the stem block with the Maxpool operation for reducing spatial decimation and the first layer of Resblock with three Bottleneck Residual structures to serve as the Resblock branch. This branch can produce the low-level feature $\mathbf{F}_0 \in \Re^{256 \times W/2 \times H/2}$ with more detailed spatial information, which is also skip connected to the decoder path for segmentation prediction.

3.2 Decoder

To effectively integrate the multi-scale features extracted by the encoder, the decoder implements a hierarchical fusion mechanism, dubbed as hierarchical feature aggregation module (HFAM), that replaces the computationally intensive Swin Transformer blocks with more efficient convolutional layers. This design choice not only enhances computational efficiency but also maintains the structural integrity of the feature representations. The HFAM involves an iterative process, where lower-resolution feature maps generated by the encoder are progressively upsampled by a factor of 2× and subsequently refined through convolutional operations. After each upsampling step, these refined features are fused with their corresponding encoder-derived feature maps that possess matching spatial dimensions. This hierarchical fusion strategy is executed recursively, enabling the gradual reconstruction of features at progressively higher resolutions. The process continues until the reconstructed feature map matches the input image's spatial resolution, thereby ensuring fidelity between the output and the original input in terms of both scale and detail. Specifically, given the lowest spatial resolution feature $\mathbf{F}_4$ extracted in the Encoder, we first employ a simple up-sampling operation and a point-wise convolution to double the spatial size and half the channel number, respectively. Then after concatenating with the feature $\mathbf{F}_3$ of the up one level, we further adopt two convolution layer to refine the fused feature. The above process can be formulated as:

$$\bar{\mathbf{F}}_3 = f_{conv}([\mathbf{F}_3, f_{up-Pw}(\mathbf{F}_4)]) \tag{4}$$

Then, $\bar{\mathbf{F}}_3$ follows the similar procedure as $\mathbf{F}_4$ for aggregating with $\mathbf{F}_2$. Finally, we achieve the fused feature $\bar{\mathbf{F}}_0$ by hierarchically aggregating all Encoder features, which is further up-sampled to the spatial resolution for producing segmentation output using a convolution block as the prediction head.

4 Experiments

4.1 Datasets

To assess the performance of the proposed HFANet, we conducted a series of experiments utilizing three publicly available datasets, each representing diverse tasks and imaging modalities. The datasets employed in the evaluation BUSI [3],

Table 1: Comparison with the state-of-the-art models.

Models	ClinicDB	BUSI	GLaS
U-Net [19]	90.66	72.27	87.99
MultiResUNet [12]	88.20	72.43	88.34
Swin-Unet [10]	90.69	76.06	86.45
UCTransnet [26]	**92.57**	75.56	87.17
SMESwin-Unet [29]	89.62	73.94	83.72
HFANet (Ours)	92.39	**84.33**	**91.70**

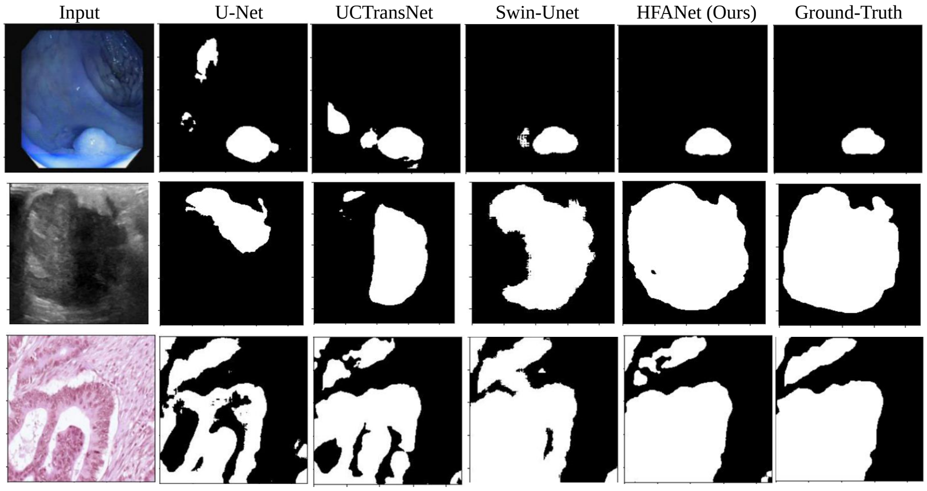

Fig. 2: Comparative qualitative results

which includes 437 benign and 210 malignant breast ultrasound images, similar to those used in [13]; CVC-ClinicDB [4], a colonoscopy dataset with 612 images; and GlaS [20], for gland segmentation, consisting of 85 training and 80 test images. To ensure consistency, all images and their corresponding segmentation masks were resized to 224×224 pixels. For the GlaS dataset, we adhered to the predefined test split to evaluate the model. In contrast, for the remaining datasets, we randomly allocated 20% of the images for testing purposes. The remaining data was divided into 60% for training and 20% for validation.

4.2 Implementation Details

The HFANet was implemented using PyTorch framework. To enhance data variability and improve model generalization, several data augmentation techniques, including horizontal and vertical flips as well as random rotations, were applied to the training dataset. Training was conducted on an Nvidia Geforce RTX 4090 GPU with 24GB of memory, and the model's Encoder parameters were initial-

Table 2: Ablation Study.

Encoder2	Decoder	ClinicDB	BUSI	GLaS
ResNet34	Swin	91.99	79.99	91.12
ResNet50	Swin	92.01	81.71	91.68
×	HFAM	92.15	82.79	**91.74**
ResNet50	HFAM	**92.39**	**84.33**	91.70

ized using pre-trained weights from ImageNet, leveraging transfer learning to accelerate convergence and improve performance, while the Decoder parameters were randomly initialized. The model optimization was performed using the Adam optimizer, initialized with a learning rate of 10^{-5}, and dynamically adjusted throughout training using a cosine annealing scheduler. The models were trained over a total of 1000 epochs with the training process incorporating an early stopping mechanism. Specifically, an early stopping patience of 100 epochs was employed, meaning that training was halted if no improvement in performance was observed over 100 consecutive epochs, thus preventing overfitting. To optimize the model, we minimized a hybrid loss function that combined cross-entropy loss and Dice loss, a strategy designed to balance pixel-wise classification accuracy with segmentation overlap quality.

4.3 Comparisons with State-of-the-Art Methods

We conducted a comprehensive evaluation of The HFANet by comparing it against five representative models from U-Net-based architectures: U-Net [19], MultiResUNet [12], Swin-Unet [10], UCTransnet [26], and SMESwin-Unet [29]. These models were selected to represent key variants within the UNet family. The compared results in terms of Dice score, are summarized in Table 1, which presents the performance across various test datasets. It can be observed from Table 1 that our proposed model outperforms competing methods, achieving the best segmentation accuracy for the BUSI and GLaS datasets, and the second rank for the ClinicDB dataset in terms of the Dice Similarity Coefficient (DSC). The improvement in DSC compared to existing methods, such as U-Net [19] and Swin-Unet [10], is marginal for the ClinicDB dataset while our model demonstrates a substantial gain for BUSI and GLaS datasets. Concretely, our approach achieves an improvement of approximately 12% over U-Net and 8% over Swin-Unet for the BUSI dataset. Finaly, we provide the visulizations of the segmentation results with several representative models including U-Net, UCTarnsNet and Swin-Unet in Fig. 2, and have demonstrated that our proposed HFANet achieves much better segmentations.

4.4 Ablation study

We conducted an ablation study to systematically evaluate the impact of the proposed Encoder and Decoder components on all three datasets. Aa introduced above that our HFANet contains two Encoders: Swin Transformer branch and ResBlock branch eaxtected from the stem and first layers of the pretrained ResNet with The ImageNet dataset (Denoted as Encoder2), and one Decoder with the proposed HFAM. To verify the effectiveness of the HFAM for feature aggregation, we also employed the symmetric Swin Transformer blocks by replacing the patch merging with upsampling operation as the Decoder component. For Encoder, we removed the Resblock branch or utilized the layer from the ResNet34 and ResNet50, respectively. The compared results are manifested in Table 2, and manifested that the proposed HFAM and the incorporation the Resblock branch can improve the segmentation performance.

5 Conclusions

This paper introduced a novel approach for improving the performance of ViT-based architectures in medical image segmentation, especially when data is limited. The core of our method lies in a hierarchical feature aggregation strategy built upon hierarchical Transformer features. Our framework employed the hierarchical Vision Transformer as the primary encoder to capture multi-scale semantic information, while residual blocks are incorporated to preserve fine spatial details and local representations. To combine these features effectively, we proposed a hierarchical feature aggregation module (HFAM) that acts as the decoder, seamlessly merging the multi-scale semantic and spatial features. Compared to traditional Transformer-based U-Net models, the HFAM not only enhances context fusion but also holds the potential to reduce computational demands. Extensive experiments across three medical image segmentation datasets showed that our method consistently surpasses both standard U-Net and other Transformer-based U-Net models in performance.

References

1. Ailiang, L., Xu, J., Jinxing, L., Guangming, L.: Contrans: Improving transformer with convolutional attention for medical image segmentation. Medical Image Computing and Computer Assisted Intervention (MICCAI) pp. 297–307 (2022) 2, 5
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Y.Wang, Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. CoRR (2021) 5
3. Cicek, O., Abdulkadir, A., Lienkamp, S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention (MICCAI) **9901**, 424–432 (2016) 4
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies **1** (2019) 4
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021) 2, 4

6. F, K.S.P.J.M.H.K.I., PF, J.: Nnu-net: a self-con
guring method for deep learning-based biomedical image segmentation. Nat Methods **18**(2), 203–211 (2021) 2
7. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE Transactions on Medical Imaging **38**(10), 2281–2292 (2019) 2
8. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. CoRR (2021) 4
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015) 6
10. Hu, C., Wang, Y., Joy, C., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. ECCV Computer Vision Workshop pp. 205–218 (2023) 3, 5, 9, 10
11. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., J.Wu: Unet 3+: A full-scale connected unet for medical image segmentation. ICASSP pp. 1055–1059 (2020) 2, 4
12. Ibtehaz, N., Rahman, M.S.: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Networks **121** (2020) 9, 10
13. Jin, Q., Meng, Z., Sun, C., Cui, H., Su, R.: Ra-unet: A hybrid deep attentionaware network to extract liver and tumor in ct scans. Frontiers in Bioengineering and Biotechnology **8**, 1471 (2020) 2
14. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE Transactions on Medical Imaging **37**(12), 2663–2674 (2018) 2, 4
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. CoRR (2021) 2, 4, 5, 7
16. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. Fourth International Conference on 3D Vision (3DV) pp. 565–571 (2016) 2, 4
17. Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M., Tomaszewski, J.: Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. 5th IEEE International Symposium in Biomedical Imaging: From Nano to Macro pp. 284–287 (2008) 2
18. abd R. Girshick, X.W., Gupta, A., He, K.: Non-local neural networks. IEEE conference on computer vision and pattern recognition pp. 7794–7803 (2018) 2
19. Ronneberger, O., P.Fischer, Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI) **9351**(3), 234–241 (2015) 2, 3, 9, 10
20. Rouhi, R., Jafari, M., Kasaei, S., Keshavarzian, P.: Benign and malignant breast tumors classi
cation based on region growing and cnn segmentation. Expert Systems with Applications **42**(3), 990–1002 (2015) 2
21. Schindelin, J., Rueden, C.T., Hiner, M.C., Eliceiri, K.W.: The imagej ecosystem: an open platform for biomedical image analysis. Molecular reproduction and development **82**, 518–529 (2015) 2
22. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis **53**, 197–207 (2019) 2

23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. CoRR (2020) 2, 4
24. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. CoRR (2021) 2, 5
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., u. Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017) 2, 4
26. Wang, H., Cao, P., Wang, J., Zaiane, O.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. AAAI Conference on Artificial Intelligence **36**(3), 2441–2449 (2022) 2, 5, 9, 10
27. Wang, W., Chen, C., Ding, M., Li, J., Yu, H., Zha, S.: Transbts: Multimodal brain tumor segmentation using transformer. CoRR (2021) 2, 5
28. Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. CoRR (2021) 4
29. Wang, Z., Min, X., Shi, F., Jin, R., Nawrin, S., Yu, I., Nagatomi, R.: Smeswin unet: Merging cnn and transformer for medical image segmentation. Medical Image Computing and Computer Assisted Intervention (MICCAI) pp. 517–526 (2022) 9, 10
30. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. 9th International Conference on Information Technology in Medicine and Education (ITME) pp. 327–331 (2018) 2, 3
31. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. CoRR (2021) 2, 5
32. Zhou, Z., Siddiquee, M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. Springer Verlag pp. 3–11 (2018) 2, 4