Hybrid CNN–Transformer Based Medical Image Segmentation for Liver and Lung CT Scans

***Abstract*—Medical image segmentation is a critical component of computer-aided diagnosis and clinical decision support, enabling precise delineation of anatomical structures from medical imaging modalities such as computed tomography (CT). Despite the CNN's advantage of learning spatial feature, it suffers the limitation of effective modeling of global contextual information, and transformer based architectures excel at capturing long range dependencies but often struggle with fine grained boundary details and incur high computationl costs. To address these complementary limitations, this work presents a faithful reimplementation and evaluation of a hybrid CNN–Transformer segmentation architecture for medical CT imaging. The proposed framework follows an encoder–decoder design that integrates convolutional feature extraction with a hierarchical transformer encoder. Multiple Mix Transformer (MiT) backbone variants are investigated to analyze the trade-off between segmentation performance and computational complexity. The model is trained from scratch and evaluated on two clinically relevant tasks: liver segmentation using the Medical Segmentation Decathlon (MSD) dataset and lung segmentation using a COVID-19 CT dataset. Experimental results demonstrate that the hybrid architecture achieves strong and consistent segmentation performance across both datasets, with intermediate-scale transformer backbones offering the most favorable balance between accuracy and efficiency. Comparative analysis against existing methods shows competitive or improved segmentation accuracy and boundary delineation. These findings indicate that hybrid CNN–Transformer architectures can be effectively adapted to medical image segmentation tasks under limited data conditions, highlighting their practical applicability for clinical CT analysis.**

I. Introduction

Medical image segmentation is a fundamental task in computer-aided diagnosis and clinical decision support systems, as it enables precise delineation of anatomical structures from medical imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI). Accurate organ segmentation plays a crucial role in applications including disease diagnosis, treatment planning, radiotherapy, and longitudinal disease monitoring. However, medical image segmentation remains challenging due to factors such as low contrast between tissues, anatomical variability across patients, complex organ boundaries, and limited availability of annotated data.

In recent years, deep learning–based approaches, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in medical image segmentation. Encoder–decoder architectures such as U-Net and its variants have become standard due to their strong capability in learning local spatial features and hierarchical representations. Despite their effectiveness, CNN-based methods are inherently limited by their localized receptive fields, which can restrict the modeling of long-range dependencies and global contextual information, an aspect that is especially important for accurately segmenting organs with complex shapes and varying scales. Transformer-based architectures, originally developed for natural language processing, have recently been introduced into computer vision tasks and have shown strong potential in capturing global feature dependencies through self-attention mechanisms. Vision Transformers and their derivatives have achieved competitive performance in dense prediction tasks, including semantic segmentation. Nevertheless, pure transformer-based models often struggle to preserve fine-grained local details and boundary information, which are critical for medical image analysis. Additionally, their high computational cost and data requirements pose challenges for training on limited medical datasets.

To address the complementary strengths and weaknesses of CNNs and transformers, hybrid CNN–Transformer architectures have gained increasing attention. These models aim to combine the strong local feature extraction capability of CNNs with the global contextual modeling ability of transformers. Several recent studies have demonstrated that such hybrid designs can achieve improved segmentation accuracy and robustness across different domains. However, many of these architectures are primarily evaluated on natural image or remote sensing datasets, and their effectiveness on medical imaging tasks, particularly across different organs and dataset scales, remains underexplored. While hybrid architectures such as SSNet [1] have shown promising results in non-medical domains, there exists a research gap in systematically reimplementing and adapting these architectures for medical image segmentation using CT datasets. In particular, there is limited empirical analysis on how different transformer backbone scales affect segmentation performance, computational complexity, and boundary accuracy in organ-specific medical tasks. Moreover, validating such architectures on multiple medical datasets with varying data sizes is essential to assess their generalization capability under realistic clinical constraints.

Motivated by these observations, this work presents a faithful reimplementation of a hybrid CNN–Transformer segmentation architecture inspired by prior SSNet-style designs and evaluates its applicability to medical image segmentation. The model is adapted and trained from scratch for liver and lung organ segmentation using CT imaging datasets. Multiple variants of a hierarchical transformer

backbone are explored to analyze the trade-off between model complexity and segmentation performance. The study emphasizes rigorous experimental evaluation using standard overlap-based and boundary-based metrics to assess both accuracy and anatomical consistency. The primary objective of this research is to investigate whether a hybrid CNN–Transformer architecture, originally proposed for non-medical segmentation tasks, can effectively generalize to medical imaging scenarios when appropriately reimplemented and trained. By conducting a detailed comparative analysis across backbone variants and datasets, this work aims to provide insights into the suitability of such hybrid designs for medical organ segmentation under limited data conditions.

The remainder of this paper is organized as follows. Section II describes the overall methodology and architectural design of the proposed implementation. Section III presents the experimental setup and evaluation metrics. Section IV reports quantitative and qualitative results on liver and lung segmentation tasks. Finally, Section V concludes the paper and outlines directions for future work.

II. Methodology

This study adopts a quantitative research methodology to evaluate the effectiveness of a hybrid CNN–Transformer segmentation architecture for medical CT image segmentation. The primary objective of this research is to investigate whether a hybrid architecture, originally proposed for general semantic segmentation tasks, can be faithfully reimplemented and reliably adapted for medical organ segmentation, and to analyze the impact of transformer backbone scale on segmentation performance. Quantitative analysis is appropriate for this study as model performance is assessed using numerical overlap-based and boundary-based evaluation metrics, enabling objective comparison across datasets and architectural variants. The overall methodology is designed to ensure transparency, reproducibility, and fair empirical evaluation.

III. Proposed Architecture

A. Overview

The proposed segmentation framework follows a hybrid CNN–Transformer encoder–decoder architecture inspired by the SSNet design originally introduced for remote sensing semantic segmentation [1]. In this work, the architectural principles of SSNet are faithfully reimplemented and adapted for medical CT image segmentation, and evaluated on the MSD Task 03 liver dataset and a COVID-19 lung CT dataset. The network is designed to jointly leverage, local spatial feature extraction through convolutional operations, and global contextual modeling through transformer-based self-attention. An overview of the proposed architecture is illustrated in Fig. 1. The framework consists of three major components, a hybrid hierarchical encoder comprising a transformer branch and a CNN branch, cross-branch interaction modules for feature fusion and information injection, and a CNN-based decoder for progressive feature upsampling and segmentation prediction.

B. Hybrid Encoder Design

C. Transformer Branch (Global Context Encoder)

The transformer branch is based on the Mix Transformer (MiT) architecture, as employed in SegFormer, and is designed to capture long-range dependencies and global anatomical context. Such global modeling capability is essential for accurate segmentation of organs with large spatial extent and high inter-patient variability. Given an input CT slice of size $1 \times H \times W$, the image is first partitioned into overlapping patches using a convolutional patch embedding layer. The transformer encoder operates hierarchically across four stages, producing multiscale feature maps with spatial resolutions of $H/4$, $H/8$, $H/16$, and $H/32$, respectively. Each stage consists of multiple transformer blocks composed of, efficient multi-head self-attention with sequence reduction, and Mix Feed-Forward Networks (Mix-FFN), which integrate multilayer perceptrons with depthwise convolution to preserve spatial continuity. This hierarchical design enables effective modeling of global dependencies while maintaining compatibility with dense prediction tasks such as medical image segmentation.

1) CNN Branch (Local Feature Encoder): In parallel, a CNN-based encoder branch is employed to capture fine-grained local features, including organ boundaries, edges, and texture patterns that are critical in CT imaging. The CNN encoder also follows a four-stage hierarchical structure, aligned with the transformer branch in terms of spatial resolution. Each stage comprises, depthwise convolutions for efficient local feature extraction, multiscale convolutional operations to enhance receptive field diversity, and pointwise 1×1 convolutions for channel-wise interaction. This branch emphasizes local spatial continuity and boundary preservation, which are often diminished in transformer-only architectures. To effectively integrate local and global representations, the architecture incorporates two complementary interaction mechanisms between the CNN and transformer branches.

2) Feature Fusion Module (FFM): At each encoder stage, a Feature Fusion Module (FFM) is employed to combine CNN features with transformer features. The primary objective of the FFM is to enhance transformer representations with fine structural details extracted by the CNN branch. The FFM integrates, spatial attention to model position-wise dependencies, channel attention to recalibrate feature importance, and pooling operations to aggregate multiscale contextual information. The resulting fused feature map is injected back into the transformer branch, improving sensitivity to fine anatomical structures and mitigating over-smoothing effects commonly observed in attention-based models.

3) Feature Injection Module (FIM): Conversely, a Feature Injection Module (FIM) is employed to inject global contextual information from the transformer branch into the CNN branch. The FIM incorporates depthwise strip

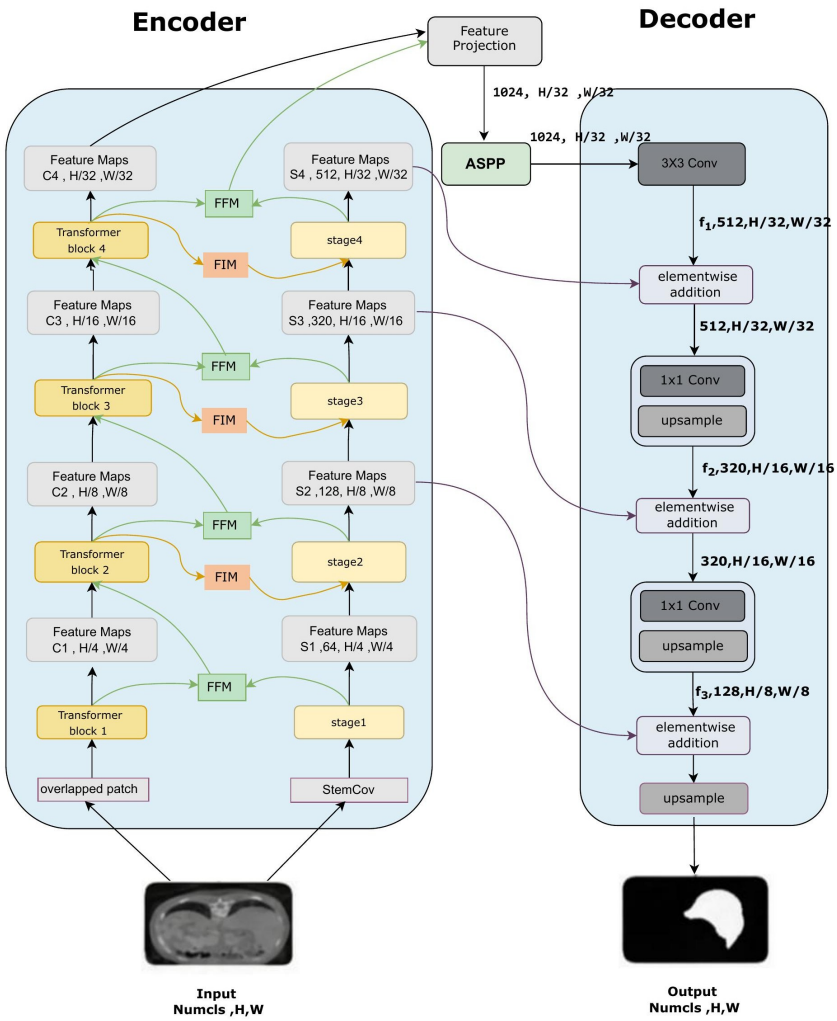

Fig. 1. Overview of the proposed hybrid CNN–Transformer architecture for medical CT image segmentation. The encoder consists of parallel transformer and CNN branches with cross-branch feature fusion and injection modules to integrate global contextual and local spatial information. The decoder progressively upsamples and refines multiscale features using skip connections and an ASPP bottleneck to produce the final binary organ segmentation mask.

convolutions to approximate large receptive fields efficiently, and squeeze-and-excitation (SE) mechanisms to adaptively reweight channel-wise responses. This process enhances the global perception capability of the CNN branch, allowing it to benefit from long-range contextual information without incurring excessive computational cost.

D. CNN-Based Decoder

A CNN-based decoder is employed to progressively recover spatial resolution and generate the final segmentation output. The decoder operates in a top-down manner and consists of, bilinear upsampling followed by convolutional refinement, skip connections from corresponding encoder stages to preserve low-level spatial details, and multiscale feature aggregation to improve boundary delineation. To further enhance multiscale contextual understanding, an Atrous Spatial Pyramid Pooling (ASPP) module is applied at the bottleneck layer. The ASPP module aggregates features using parallel dilated convolutions with varying dilation rates, enabling robust segmentation of organs with diverse sizes and shapes. The final segmentation output is produced using a 1×1 convolution followed by a sigmoid activation function, yielding a binary segmentation mask for liver or lung regions.

E. Architectural Adaptation for Medical CT Segmentation

No structural modifications were introduced to the core SSNet architecture. Instead, adaptation to medical CT segmentation is achieved through, replacing RGB remote sensing inputs with single-channel CT slices, reformulating the task as binary organ segmentation, and training the model from scratch on medical datasets without pretrained weights. This design choice ensures that observed performance gains are attributable to the architectural formulation itself rather than domain-specific tuning or reliance on pretrained representations. segmentation.

F. Datasets and Data Preparation

Experiments are conducted on two publicly available medical CT imaging datasets selected for their clinical relevance and availability of expert-annotated ground truth masks. Liver segmentation experiments utilize the Medical Segmentation Decathlon (MSD) Task 03 dataset [20], which consists of contrast-enhanced abdominal CT volumes with corresponding liver annotations. Lung segmentation experiments are performed on a COVID-19 CT dataset [21] containing annotated lung regions. All volumetric CT scans are converted into 2D axial slices to facilitate efficient training and evaluation. Each slice is resized to a spatial resolution of 256×256 pixels and processed as a single-channel grayscale image. Intensity normalization is applied to reduce inter-scan variability caused by differences in imaging protocols. Segmentation is formulated as a binary classification task, distinguishing the target organ from the background. For both datasets, data splitting is performed at the *volume level* using a 70% / 15% / 15% ratio for training, validation, and testing, respectively. This ensures that all slices derived from a given CT volume belong exclusively to a single split, thereby preventing data leakage and enabling reliable and unbiased performance evaluation.

G. Training Strategy

The proposed model is trained from scratch without using pretrained weights to ensure that all learned representations are derived exclusively from medical imaging data. Training is conducted on a GPU-enabled system using mini-batch gradient-based optimization. The network is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4}. A combined Dice loss and Binary Cross-Entropy (Dice+BCE) loss function is employed to address class imbalance and improve both region overlap and boundary accuracy. Batch size is selected based on dataset size and GPU memory constraints. Training is performed for a maximum of 50 epochs, with early stopping applied based on validation performance. If no improvement in validation Dice score is observed for five consecutive epochs (patience = 5), training is terminated to prevent overfitting. The model achieving the best validation performance is retained for final evaluation.

H. Evaluation Metrics and Analysis

Segmentation performance is quantitatively evaluated using standard overlap-based and boundary-based metrics. The Dice Similarity Coefficient (DSC) is used to assess region overlap between predicted and ground truth masks, while the 95th percentile Hausdorff Distance (HD95) is employed to measure boundary accuracy. Additional region-based metrics, including precision, recall, and mean Intersection over Union (mIoU), are also reported to provide a comprehensive evaluation of segmentation quality. These metrics enable consistent comparison across backbone variants and datasets.

IV. Results

This section presents the quantitative results obtained from evaluating the reimplemented hybrid CNN–Transformer segmentation model on medical CT imaging datasets. The results are reported in relation to the study objective of assessing segmentation performance across different transformer backbone variants and datasets using standard overlap-based and boundary-based metrics.

A. Experimental Setup Overview

The model is evaluated on two segmentation tasks: liver segmentation using the Medical Segmentation Decathlon (MSD) Task 03 dataset and lung segmentation using a COVID-19 CT dataset. Performance is measured on held-out test data following volume-level dataset splits. All results are reported using Dice Similarity Coefficient (DSC), 95th percentile Hausdorff Distance (HD95), precision, recall, and mean Intersection over Union (mIoU).

B. Liver Segmentation Results

Quantitative liver segmentation results obtained using three different Mix Transformer (MiT) backbone variants are summarized below.

TABLE I
LIVER SEGMENTATION PERFORMANCE USING DIFFERENT MIT BACKBONE VARIANTS ON THE MSD TASK 03 DATASET.

Backbone	GFlOPs	Params (M)	DSC↑	HD95↓	Precision↑	Recall↑
MiT-B0	3.73	20.0	95.94	7.54	97.69	96.44
MiT-B3	13.0	96.0	**96.03**	**7.49**	**97.83**	**96.49**
MiT-B5	20.0	133.99	95.13	10.50	97.08	96.26

C. Comparison with Existing Methods

This subsection presents a quantitative comparison between the proposed hybrid CNN–Transformer architecture and previously published methods on liver and lung segmentation benchmarks. Comparisons are reported using Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) as available from the respective studies.

TABLE II
PERFORMANCE COMPARISON WITH EXISTING METHODS ON THE MSD LIVER DATASET.

Model	Dice (%) ↑	HD95 (mm) ↓
FF Swin-Unet [2]	94.42	15.94
vMixer [3]	94.89	10.26
UNetFormer [4]	95.73	7.68
nnUNet [5]	95.75	7.94
TransUNet [6]	92.66	–
SwinUNet [7]	94.17	–
FSS ULivR [8]	94.78	–
Swin UNETR [9]	95.35	–
DiNTS [10]	95.35	–
EffiDec 3D [11]	93.68	–
VNet with Attention Gate [12]	95.54	–
SSNet (Ours)	**96.03**	**7.49**

TABLE III
PERFORMANCE COMPARISON WITH EXISTING METHODS ON THE COVID-19 LUNG DATASET.

Model	Dice (%) ↑	HD95 (mm) ↓
U-Net [13]	89.00	–
UNet++ [14]	98.30	–
MultiResUNet [15]	89.88	–
QAPNet [16]	81.63	–
3D U-Net [17]	95.60	–
nnUNet [18]	87.90	–
LungQuant2 [19]	96.01	–
SSNet (Ours)	**96.03**	**3.65**

As shown in Tables II and III, the proposed SSNet-based implementation achieves competitive or superior Dice scores compared to existing methods on both liver and lung segmentation benchmarks. Notably, the proposed approach attains the highest reported Dice score on the MSD Liver dataset while maintaining improved boundary accuracy as measured by HD95.

D. Qualitative Segmentation Results

Representative qualitative segmentation results for liver and lung organs are shown in Fig. 2 and 3.

E. Summary of Results

Overall, the results demonstrate consistent segmentation performance across liver and lung datasets using the evaluated hybrid CNN–Transformer architecture. Quantitative metrics indicate stable overlap accuracy and boundary delineation across experiments, with performance variations observed across transformer backbone scales and datasets. These findings provide the basis for further analysis and discussion in the subsequent section.

V. DISCUSSION

This work evaluates a reimplemented SSNet-style CNN–Transformer hybrid architecture for medical image segmentation under limited data conditions. The proposed model demonstrates strong and consistent performance across both liver and lung datasets, indicating effective generalization across different anatomical structures. Among the evaluated transformer backbones, MiT-B3 achieves the best balance between segmentation accuracy and computational efficiency. It attains the highest Dice scores and lowest HD95 values for liver segmentation while maintaining stable performance on lung segmentation. This suggests that intermediate-capacity transformer backbones are sufficient for capturing global context without degrading boundary precision.

The hybrid design leverages the complementary strengths of convolutional and transformer-based components, enabling effective learning of both local structural details and long-range dependencies. Improved boundary accuracy on the MSD Liver dataset further confirms the suitability of this architecture for anatomically sensitive segmentation tasks. Finally, the results indicate that increasing model capacity does not necessarily lead to improved performance. Larger backbones such as MiT-B5 introduce higher computational cost without clear accuracy gains, highlighting the importance of balancing model complexity with dataset scale in medical image segmentation.

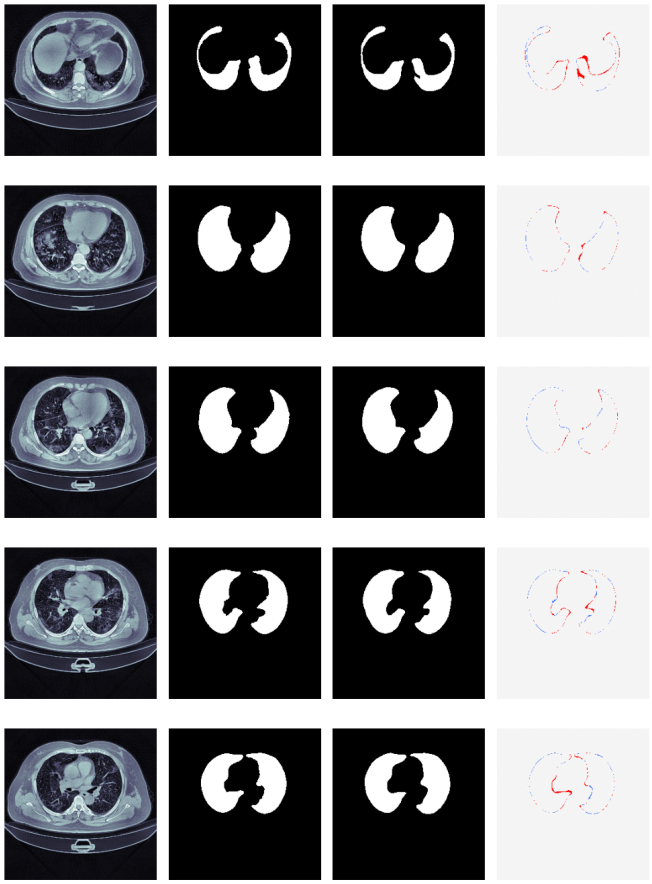

Fig. 2. Qualitative Segmentation results for Lungs. The first column represents input CT images, second column represent Groun truth and third column represent the Predicted Segmentation Mask and the last column represent the difference map highlighting the difference between the GT and predicted.

VI. CONCLUSION

This work investigated the applicability of a hybrid CNN–Transformer segmentation architecture for medical CT image analysis through a faithful reimplementation and empirical evaluation on liver and lung segmentation tasks. By adapting the architecture to medical imaging data and training it from scratch, the study assessed whether such hybrid designs, originally proposed for non-medical domains, can generalize effectively to organ segmentation problems under realistic data constraints. Experimental results demonstrate that the proposed implementation achieves strong and consistent segmentation performance across both datasets. The analysis of multiple MixVision Transformer backbone variants highlights a clear trade-off between model complexity and segmentation accuracy, with intermediate-scale backbones offering an effective balance between computational efficiency

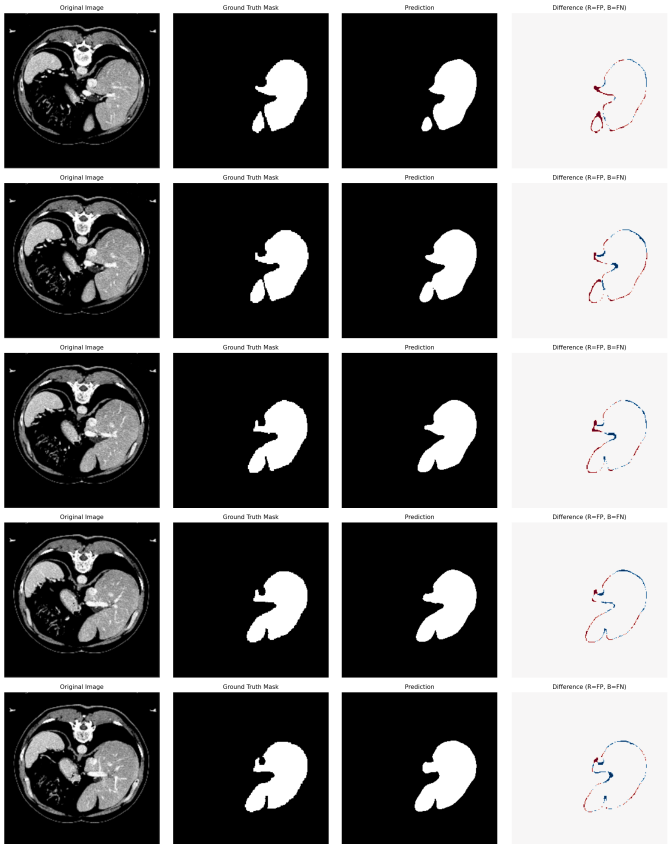

Fig. 3. Qualitative Segmentation results for Liver. The first column represents input CT images, second column represent Groun truth and third column represent the Predicted Segmentation Mask and the last column represent the difference map highlighting the difference between the GT and predicted.

and performance. Comparative evaluation against existing methods further indicates that the proposed approach attains competitive, and in several cases superior, overlap accuracy and boundary precision, underscoring the benefit of integrating convolutional inductive biases with transformer-based global context modeling.

Looking forward, this work opens several promising research directions. A key extension involves scaling the proposed framework to multi-organ and multi-class segmentation tasks, which would better reflect real-world clinical scenarios. Additional future efforts may include volumetric 3D modeling, evaluation across a broader range of anatomical structures, and the incorporation of pretrained or self-supervised representations to further improve robustness and generalization. Overall, this study provides empirical evidence that hybrid CNN–Transformer architectures constitute a practical and effective solution for medical image segmentation and establishes a solid foundation for future advancements in clinical image analysis.

REFERENCES

[1] Min Yao, Yaozu Zhang, Guofeng Liu, and Dongdong Pang "SSNet: A Hybrid CNN–Transformer Architecture for Semantic Segmentation," *IEEE Access*, 2024.

[2] M. A. Ahmed et al., "FFSwinUnet: Feature Fusion Swin Transformer for Medical Image Segmentation," *Computer Methods and Programs in Biomedicine*, vol. 245, 2024, Article 107652.

[3] Y. Zhang, H. Liu, Q. Hu, and Y. Yang, "VMixer: Vision MLP-Mixer for Medical Image Segmentation," *arXiv preprint arXiv:2204.09811*, 2022.

[4] H. Wang, Y. Chen, Y. Zhang, and W. Liu, "UNetFormer: A UNet-like Transformer for Efficient Medical Image Segmentation," *arXiv preprint arXiv:2109.07164*, 2021.

[5] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[6] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[7] Z. Cao, Y. Xu, J. Xu, T. Liu, X. Li, and Z. Zhang, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," in *Proc. ECCV Workshops*, 2022.

[8] Y. Zhang et al., "Few-Shot Liver Segmentation with Global and Local Information Fusion," *Medical Image Analysis*, vol. 74, 2021.

[9] Y. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *Proc. MICCAI*, 2022.

[10] Y. Tang, Y. Yang, and L. Wang, "DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation," in *Proc. CVPR*, 2022.

[11] M. Shakeri et al., "Efficient Encoder–Decoder Networks for 3D Medical Image Segmentation," *Medical Image Analysis*, vol. 76, 2022.

[12] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *Proc. 3DV*, 2016.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. MICCAI*, 2015.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis*, Springer, 2018.

[15] I. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

[16] M. Zhou et al., "QAPNet: Quantization-Aware Pyramid Network for COVID-19 Lung Segmentation," *Pattern Recognition*, 2021.

[17] A. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Proc. MICCAI*, 2016.

[18] K. H. Weygand, B. H. Menze, F. Isensee, and K. H. Maier-Hein, "nnU-Net for Medical Image Segmentation: State-of-the-Art AutoML," *arXiv preprint arXiv:2004.12537v1*, 2020.

[19] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task Deep Learning Based CT Imaging Analysis for COVID-19: Classification and Segmentation," *arXiv preprint arXiv:2105.02566*, 2021.

[20] Medical Segmentation Decathlon, "Task 03: Liver Tumor Segmentation," Available: https://medicaldecathlon.com/

[21] S. Ma et al., "COVID-19 CT Lung and Infection Segmentation Dataset," Zenodo, 2020.