© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718946

Hybrid Transformer-CNN Framework for Multi-Organ Tumor Segmentation in Abdominal CT scans

DANDAGULA JAGADEESH[1], CHITLA VIGNESH[2], VADDELLI SRINIVAS RAO[3]
[1,3]*Department of Computer Science and Engineering-AIM, Aurora Deemed to be University*
[2]*Department of Computer Science and Engineering-AIM, Hyderabad*

Abstract- - Accurate segmentation of multiple abdominal organs and co-occurring tumors in CT imaging presents a significant challenge due to high inter-patient anatomical variability, low contrast boundaries, and class imbalance. This paper proposes HybridSegNet, a novel architecture that integrates a Swin Transformer encoder with a multi-scale convolutional decoder equipped with dense skip connections and a dual-branch feature fusion module. HybridSegNet is evaluated on the publicly available CHAOS dataset (liver, kidney segmentation) and the KiTS23 challenge dataset (kidney tumor segmentation). The model achieves a mean Dice Similarity Coefficient (DSC) of 0.913 on liver segmentation, 0.897 on kidney segmentation, and 0.884 on kidney tumor segmentation, outperforming leading methods including nnU-Net (DSC: 0.876) and Swin-UNet (DSC: 0.861). A lightweight model variant is also proposed for deployment on resource-constrained clinical workstations without significant performance degradation. Results demonstrate HybridSegNet's suitability for real-time clinical decision support in abdominal oncology.

Index Terms- Abdominal CT, Deep Learning, Feature Fusion, Medical Image Segmentation, Multi-Organ Segmentation, Swin Transformer, Tumor Detection

I. INTRODUCTION

Abdominal cancers, including hepatocellular carcinoma, renal cell carcinoma, and pancreatic ductal adenocarcinoma, collectively account for a substantial proportion of cancer-related deaths globally. Computed Tomography (CT) remains the primary diagnostic modality for abdominal malignancies due to its high spatial resolution and rapid acquisition times. Precise delineation of tumor boundaries in CT volumes is critical for treatment planning, surgical simulation, and longitudinal monitoring of therapeutic response.

Manual segmentation by radiologists is labor-intensive and subject to substantial inter-observer variability, particularly in abdominal regions where organs exhibit complex spatial relationships and tumors often share similar Hounsfield Unit profiles with adjacent healthy tissue. Automated segmentation methods have advanced substantially through deep learning; however, existing fully convolutional network (FCN) approaches are limited in their ability to capture long-range spatial dependencies that are essential for multi-organ boundary disambiguation.

Vision Transformers (ViTs) and their hierarchical variants such as the Swin Transformer have demonstrated exceptional global context modeling through self-attention mechanisms. However, pure Transformer architectures lack the inherent inductive biases of convolutional networks, making them data-hungry and prone to overfitting on limited medical datasets. This work proposes HybridSegNet, which strategically combines the global contextual power of Swin Transformers with the local feature extraction efficiency of multi-scale convolutional decoders. The main contributions of this paper are:

1) A hybrid encoder that pre-trains Swin Transformer blocks on a large unlabelled CT corpus via masked autoencoding before fine-tuning on segmentation tasks.
2) A Dual-Branch Feature Fusion (DBFF) module that independently processes local edge features and global contextual features before adaptive gated merging at each decoder stage.
3) Dense skip connections between non-adjacent encoder and decoder layers to prevent feature degradation across multiple upsampling stages.

4) A lightweight distilled variant, HybridSegNet-Lite, suitable for deployment on standard clinical workstations without GPU acceleration.

II. RELATED WORK

Fully convolutional networks, pioneered by Long et al. [1], laid the groundwork for semantic segmentation. The introduction of the U-Net architecture by Ronneberger et al. [2] demonstrated that encoder-decoder structures with skip connections are highly effective for biomedical segmentation tasks with limited annotations. nnU-Net [3], proposed by Isensee et al., automated the configuration of U-Net hyperparameters based on dataset fingerprints, establishing a powerful self-configuring baseline that remains competitive across numerous medical segmentation benchmarks.

Transformer-based approaches have recently entered the medical imaging domain. TransUNet [4] hybridizes a ResNet encoder with Transformer layers for global context modeling before passing features to a U-Net decoder. Swin-UNet [5] utilizes Swin Transformer blocks in both encoder and decoder paths, achieving strong performance but requiring substantial training data. UNETR [6] applies pure Transformer encoders to volumetric segmentation, demonstrating the scalability of attention mechanisms to 3D medical volumes.

Despite these advances, a persistent challenge is the efficient integration of global attention with local edge-sensitive features — particularly for multi-organ scenarios involving organs with thin, low-contrast boundaries. The proposed DBFF module directly addresses this limitation by maintaining separate processing streams for edge and context features before adaptive fusion, a design not previously explored in the abdominal multi-organ segmentation context.

III. PROPOSED ARCHITECTURE

A. Swin Transformer Encoder: HybridSegNet employs a hierarchical Swin Transformer as its encoder backbone. The encoder operates on non-overlapping 2D patches extracted from axial CT slices, with shifted window self-attention computed across four stages to progressively reduce spatial resolution while doubling channel capacity. The encoder is pre-trained using masked autoencoding on 12,000 unlabelled abdominal CT volumes sourced from publicly available archives, allowing the model to learn rich anatomical priors before fine-tuning on annotated segmentation datasets.

B. Multi-Scale Convolutional Decoder: The decoder consists of four upsampling stages, each comprising a bilinear upsampling operation followed by two 3x3 depthwise-separable convolutions with Group Normalization and GELU activations. Depthwise-separable convolutions reduce parameter count compared to standard convolutions, facilitating the lightweight variant. At each stage, dense skip connections aggregate features from all preceding encoder stages at the matching spatial resolution, preventing information loss during progressive upsampling.

C. Dual-Branch Feature Fusion (DBFF) Module: The DBFF module, inserted at each skip connection junction, splits incoming encoder features into two parallel branches: an edge branch using Sobel-inspired trainable gradient filters to extract boundary-sensitive features, and a context branch applying depthwise convolutions over a large receptive field (7x7 kernel) to capture organ-level contextual information. Outputs from both branches are concatenated and passed through a spatial gating mechanism: g = sigmoid(W_gate * [F_edge; F_ctx]) to produce a soft spatial mask that weights the contribution of each branch before the fused feature is forwarded to the decoder.

D. Training Configuration: Training uses a combined loss of Dice Loss and Cross-Entropy with class frequency weighting to address foreground-background imbalance. Mixed-precision training is performed with the AdamW optimizer (lr = 2e-4, weight decay = 1e-2) and a cosine annealing schedule over 300 epochs. Data augmentation includes random affine transforms, elastic deformations, Gaussian noise, and simulated CT beam-hardening artifacts.

© JUN 2026 | IRE Journals | Volume 9 Issue 12 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I12-1718946

All experiments are conducted on two NVIDIA A100 80GB GPUs with distributed data parallel training.

IV. EXPERIMENTAL RESULTS

HybridSegNet was evaluated on two benchmark datasets: CHAOS 2019 (Combined Healthy Abdominal Organ Segmentation, targeting liver, kidneys, and spleen in CT and MRI) and KiTS23 (Kidney and Kidney Tumor Segmentation Challenge, comprising 489 CT volumes with annotated kidneys, cysts, and tumors).

A. Quantitative Results: On the CHAOS CT subset, HybridSegNet achieved a mean DSC of 0.913 for liver, 0.897 for kidneys, and 0.905 for spleen, compared to nnU-Net (0.876, 0.862, 0.881) and Swin-UNet (0.861, 0.844, 0.868). The 95th percentile Hausdorff Distance was reduced from 7.4 mm (nnU-Net) to 3.9 mm, reflecting superior boundary precision particularly at organ peripheries.

On KiTS23, the model achieved a kidney DSC of 0.926 and kidney tumor DSC of 0.884, against nnU-Net baselines of 0.908 and 0.851 respectively. Notably, HybridSegNet demonstrated consistent performance on small tumors (< 1.5 cm diameter), a segment where existing CNN-only models frequently exhibit false negatives due to limited receptive field.

Method	Liver DSC	Kidney DSC	Spleen DSC	Tumor DSC (KiTS23)	HD95 (mm)
nnU-Net [3]	0.876	0.862	0.881	0.851	7.4
Swin-UNet [5]	0.861	0.844	0.868	0.839	8.1
TransUNet [4]	0.869	0.853	0.874	0.845	7.8
UNETR [6]	0.857	0.841	0.862	0.832	8.5
HybridSegNet (Ours)	0.913	0.897	0.905	0.884	3.9

Table I: Comparative segmentation performance on CHAOS and KiTS23 datasets. Bold indicates best results.

B. Ablation Study: Removing the DBFF module and reverting to standard skip connections reduced liver DSC by 2.9% and tumor DSC by 3.4%, confirming the module's contribution to boundary-sensitive feature integration. Replacing the Swin Transformer encoder with a standard ResNet-50 encoder decreased liver DSC by 2.1%, while removing the dense skip connections reduced tumor DSC by 1.7%. The masked autoencoding pretraining contributed an additional 1.3% improvement in tumor DSC over random initialization, underscoring the value of large-scale unlabelled CT pretraining.

C. HybridSegNet-Lite: The lightweight variant, produced via knowledge distillation from HybridSegNet, achieves 94.2% of the full model's tumor DSC while reducing inference time from 1.8 seconds to 0.6 seconds per volume slice on a standard clinical workstation (Intel Xeon, 32 GB RAM, without GPU). This demonstrates practical feasibility for real-time clinical deployment in hospitals with limited computational infrastructure.

V. CONCLUSION

This paper presented HybridSegNet, a hybrid Transformer-CNN architecture for multi-organ and tumor segmentation in abdominal CT imaging. The integration of a Swin Transformer encoder, multi-scale convolutional decoder with dense skip connections, and a novel Dual-Branch Feature Fusion module delivers state-of-the-art performance on both CHAOS and KiTS23 benchmarks. The proposed architecture demonstrates particular strength in delineating small tumors and organ boundaries — clinically critical capabilities for early cancer detection. The HybridSegNet-Lite variant further extends the applicability of the model to resource-constrained clinical settings. Future work will extend HybridSegNet to fully 3D volumetric segmentation and explore its integration with radiotherapy planning workflows.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic

Segmentation," in Proc. CVPR, 2015, pp. 3431-3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, 2015, pp. 234-241.

[3] F. Isensee, P. F. Jaeger, 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, pp. 203-211, 2021.

[4] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, 2021.

[5] H. Cao et al., "Swin-UNet: Unet-Like Pure Transformer for Medical Image Segmentation," in Proc. ECCV Workshops, 2022.

[6] A. Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation," in Proc. WACV, 2022, pp. 574-584.

[7] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. ICCV, 2021, pp. 10012-10022.

[8] A. E. Kavur et al., "CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation," Medical Image Analysis, vol. 69, p. 101950, 2021.

[9] N. Heller et al., "The KiTS23 Challenge Dataset: 500 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes," arXiv preprint arXiv:2307.01984, 2023.