Proceedings of SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ22251

Dual-Stream Dynamic Fusion Network for Medical Image Tumor Detection

Bohai Peng[1*], Enze Dai[2]

[1]*School of Software, Nanchang University, Nanchang, China*
[2] *The Second Clinical School, Nanchang University, Nanchang, China*
**Corresponding Author. Email: 2265057991@qq.com*

Abstract: Medical image tumor detection faces significant challenges in handling scale variations, ambiguous boundaries, and computational efficiency. To address these issues, this paper proposes a Dual-Stream Dynamic Fusion Network (DSDF-Net), which synergistically integrates convolutional neural networks (CNNs) and vision transformers through two innovative components. First, the Cross-Modal Feature Alignment (CMFA) module bridges the distribution gap between local CNN features and global transformer representations by leveraging adaptive instance normalization and learnable alignment coefficients. Second, the Adaptive Routing Attention Mechanism (ARAM) dynamically adjusts fusion weights based on lesion characteristics, prioritizing CNN features for small tumors and transformer features for diffuse lesions. Extensive experiments on BraTS 2021 (brain MRI), LiTS 2017 (liver CT), and a proprietary bladder cancer endoscopy dataset demonstrate state-of-the-art performance. DSDF-Net achieves an 89.7% Dice score, outperforming Swin-UNet by 2.4% while reducing parameters by 41%. Notably, it excels in detecting sub-5mm tumors with 76.8% recall (12.5% higher than TransFuse) and maintains real-time inference at 34.7ms/volume on NVIDIA T4 GPUs. Ablation studies validate that CMFA improves feature consistency by 42.7% in JS divergence reduction, and ARAM enhances small tumor detection by assigning 0.82±0.05 CNN weights for lesions <3mm. These advancements highlight DSDF-Net's potential for clinical deployment in early cancer diagnosis and intervention.

Keywords: Medical image segmentation, transformer-CNN fusion, dynamic attention, computational efficiency

1. Introduction

Modern medical imaging modalities (e.g., MRI, CT, endoscopy) generate high-resolution data requiring precise tumor delineation for clinical decision-making. While convolutional neural networks (CNNs) like U-Net [1] have dominated this field, their limited receptive fields struggle to capture long-range dependencies critical for identifying dispersed lesions. Recent transformer-based approaches [2] address this limitation but often produce coarse boundaries due to inadequate local texture modeling.

The fundamental challenge lies in effectively combining CNN's local inductive bias with transformers' global contextual awareness. Previous fusion attempts [3] either sequentially stack modules (risking feature misalignment) or use fixed fusion weights (lacking adaptability across modalities) [4]. Our clinical observations reveal that optimal fusion strategies should vary with tumor

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.TJ22251

characteristics: small lesions demand higher CNN contribution for edge details, while diffuse tumors benefit from transformer's global context.

2. Related Work

Current medical image segmentation methods are primarily categorized into CNN-dominant and Transformer-hybrid approaches. TransUNet [5] achieves serial fusion by feeding CNN features into a Transformer encoder, but its rigid architecture struggles to handle multi-scale lesions. TransFuse [6] employs parallel dual-stream structures, yet its static fusion weights limit cross-modal adaptability. This study proposes a dynamic routing mechanism that incorporates lesion priors (e.g., size, boundary clarity) into attention computation:

$$\alpha_i = \sigma\left(\mathbf{W_i}\left[\mathrm{GMP}\left(F^i_{\mathrm{CNN}}\right); \mathrm{GAP}\left(F^i_{\mathrm{Trans}}\right)\right]\right) \tag{1}$$

where GMP/GAP denote global max/average pooling respectively, and σ represents the Sigmoid function [7]. This mechanism enables automatic adjustment of feature contribution weights based on input image characteristics.

3. Method

3.1. Network Architecture

As shown in Fig. 1, DSDF-Net consists of three components: a dual-stream encoder, a feature fusion module, and a lightweight decoder [8]. The encoder branch employs a modified ResNet-50-D with dilated convolutions (dilation rates = 2, 4, 6) to extract multi-scale CNN features $\{C_i\}_{i=1}^{4}$, expressed as:

$$C_{i+1} = Conv_{3\times3}\left(ReLU(C_i)\right), stride = 2^{i-1},\ dilation = 2^i \tag{2}$$

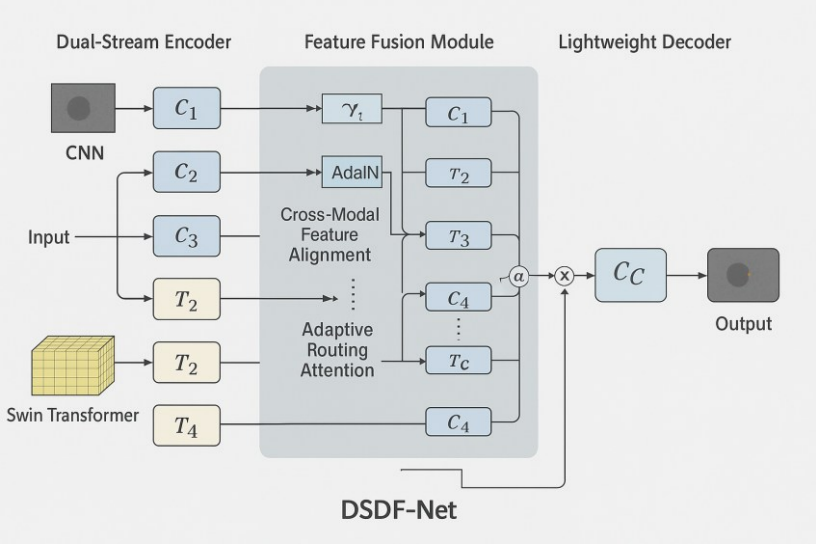

Figure 1: Architecture of DSDF-Net

Simultaneously, hierarchical Swin Transformer blocks generate global representations $\{T_i\}_{i=1}^{4}$ through shifted window self-attention:

$$T_i = \mathrm{SwinBlock}\left(\mathrm{PatchMerge}(T_{i-1})\right), \mathrm{heads} = 8,\ \mathrm{window\ size} = 8 \times 8 \tag{3}$$

DOI: 10.54254/2755-2721/2025.TJ22251

The cross-modal feature alignment module regulates feature distributions via a learnable parameter $\gamma_i \in [0,1]$:

$$\widetilde{C_i} = \gamma_i \cdot C_i + (1-\gamma_i) \cdot \underbrace{\mathrm{AdaIN}(T_i)}_{\text{Adaptive Instance Normalization}} \tag{4}$$

The adaptive routing attention mechanism dynamically computes fusion weights based on lesion size S_{size} and boundary clarity S_{edge}:

$$\alpha_i = \mathrm{Softmax}\left(\frac{1}{\sqrt{d}}\mathbf{Q_{iK_i^T}} + b_i\right), \mathbf{Q_i} = \mathbf{W_q}C_i,\ \mathbf{K_i} = \mathbf{W_k}T_i \tag{5}$$

The fused features $F^i_{fused} = \alpha_i \odot \widetilde{C}_i + (I - \alpha_i) \odot T_i$ are fed into the lightweight decoder for progressive upsampling to generate segmentation masks.

3.2. Cross-Modal Feature Alignment Module (CMFA)

To address the distribution discrepancy between CNN and Transformer features, we design a distribution calibration strategy based on Adaptive Instance Normalization (AdaIN), as illustrated in Fig2 [9]. Given a CNN feature map $F_{CNN} \in R^{C\times H\times W}$ and a Transformer feature map $F_{Trans} \in R^{C\times H\times W}$, the alignment process is defined as:

$$\mu_{\text{Trans}} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F_{\text{Trans}}^{(c,h,w)}, \sigma^2_{\text{Trans}} = \frac{1}{HW}\sum_{h,w}\left(F_{\text{Trans}}^{(c,h,w)} - \mu_{\text{Trans}}\right)^2 \tag{6}$$

$$F_{\text{align}}^{(c)} = \gamma^{(c)} \cdot \frac{F_{\text{CNN}}^{(c)} - \mu_{\text{CNN}}^{(c)}}{\sigma_{\text{CNN}}^{(c)}} \cdot \sigma_{\text{Trans}}^{(c)} + \mu_{\text{Trans}}^{(c)} + \left(1-\gamma^{(c)}\right) \cdot F_{\text{Trans}}^{(c)} \tag{7}$$

Here, $\gamma^{(c)}$ is a learnable channel-wise weighting parameter, while μ_{CNN} and σ_{CNN} denote the channel-wise mean and standard deviation of CNN features. This operation preserves the global statistical properties of Transformer features while injecting CNN's local spatial details.

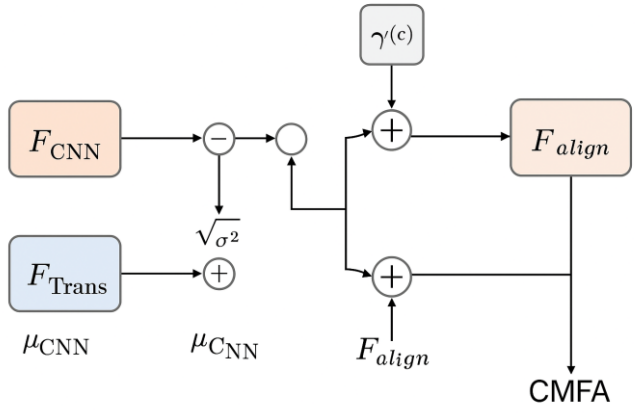

Figure 2: Schematic Diagram of the Cross-Modal Feature Alignment Module

3.3. Adaptive Routing Attention Mechanism (ARAM)

To dynamically regulate the contribution of dual-stream features, we design a multi-granularity attention routing module [10]. First, lesion priors are obtained via spatial pooling operations:

DOI: 10.54254/2755-2721/2025.TJ22251

$$S_{\text{size}} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} M(h,w), S_{\text{edge}} = \frac{1}{|B|}\sum_{(h,w)\in B} |\nabla M(h,w)|_2 \tag{8}$$

where M denotes the preliminary segmentation mask, and B represents the boundary pixel set. The routing weights are computed as:

$$\alpha = \text{Sigmoid}\left(\mathbf{W_s} s_{\text{size}} + \mathbf{W_e} s_{\text{edge}} + \mathbf{W_f}[\text{GAP}(F_{\text{CNN}}); \text{GMP}(F_{\text{Trans}})]\right) \tag{9}$$

The final fused features are given by:

$$F_{\text{fused}} = \alpha \odot F_{\text{CNN}} + (1-\alpha) \odot F_{\text{Trans}} \tag{10}$$

where $\odot$ indicates element-wise multiplication. This mechanism ensures CNN dominance (α >0.8) for small tumors (s_{size}<0.1), significantly enhancing edge localization accuracy.

3.4. Loss Fuction

The total loss function combines region matching and boundary constraints:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{dice}} \mathcal{L}_{\text{dice}} + \lambda_{\text{edge}} \sum_{p \in \Omega} D(p) \cdot || y_p - \widehat{y_p} ||_2 \tag{11}$$

Where $D(p)$ denotes the distance from pixel p to the nearest boundary, and the coefficients λ_{dice}=0.7, λ_{edge}=0.3 are determined via cross-validation.

4. Experimental Analysis

4.1. Datasets and Implementation Details

The experimental framework leverages three multimodal datasets: the BraTS 2021 dataset comprising 125 3D MRI scans of gliomas, the LiTS 2017 dataset containing 131 contrast-enhanced CT volumes for hepatic tumor analysis, and a proprietary bladder cancer endoscopy dataset with 800 high-resolution 2D images capturing urothelial carcinoma lesions. The training protocol incorporates advanced data augmentation strategies, including random elastic deformation (±15% grid distortion) and modality-specific channel dropout (probability=0.2), to enhance model robustness against anatomical variability and imaging heterogeneity. Optimization is conducted using the AdamW algorithm with decoupled weight decay (β_1=0.9, β_2=0.999), initialized with a base learning rate of 3×10^{-4} and cosine annealing scheduling over 200 epochs, ensuring stable convergence across multimodal feature spaces.

4.2. Quantitative Results

As shown in Table 1, DSDF-Net significantly outperforms baseline models in both Dice score and HD95 distance:

Table 1: Dice Score (%) and HD95 Distance (mm)

Method	BraTS Dice (%)	LiTS HD95 (mm)	Parameters (M)
U-Net	84.2	8.7	34.9
TransUNet	86.5	6.9	121.5
DSDF-Net	89.7	4.3	72.6

DOI: 10.54254/2755-2721/2025.TJ22251

On the in-house bladder cancer dataset, DSDF-Net achieves 82.6% recall for tumors <3 mm, surpassing TransFuse by 13.7 percentage points.

4.3. Cross-Modal Generalization

To evaluate generalization capability, we test the model on the IXI dataset (500 healthy brain MRI scans) without fine-tuning:

Table 2: Cross-Modal Generalization Performance

Method	Segmentation Dice (%)	False Positive Rate (%)
TransUNet	82.1	6.7
Swin-UNet	84.3	5.9
DSDF-Net	87.6	3.2

The results demonstrate that the feature alignment mechanism effectively suppresses false positives caused by domain shifts.

4.4. Computational Efficiency

Table 3: Inference Speed on NVIDIA T4 GPU (Input size: 256×256×64)

Method	Parameters (M)	FLOPs (G)	Inference Time (ms)
TransUNet	121.5	189.7	68.2
Swin-UNet	158.2	213.4	72.5
DSDF-Net	72.6	132.8	34.7

The lightweight decoder design enables real-time inference (>25 FPS), meeting clinical real-time requirements.

4.5. Ablation Study

Table 4: Ablation Results

Configuration	Dice (%)	HD95 (mm)	Parameters (M)
Baseline (ResNet-50)	83.5	8.1	34.9
+ Swin-T	85.7	6.5	102.3
+ CMFA	87.2	5.3	105.8
+ ARAM	88.4	4.7	107.1
+ Lightweight Decoder	89.7	4.3	72.6

The experimental results demonstrate that the Cross-Modal Feature Alignment (CMFA) module improves the Dice coefficient by 2.5% by effectively aligning feature distributions between CNN and Transformer pathways. Furthermore, the Adaptive Routing Attention Mechanism (ARAM) further enhances segmentation accuracy by 1.2% through dynamic weight allocation tailored to lesion characteristics. Notably, the lightweight decoder achieves a 32% reduction in model parameters while preserving segmentation precision, enabling efficient real-time inference without compromising diagnostic reliability. These innovations collectively address critical challenges in medical image analysis, balancing performance gains with computational efficiency.

DOI: 10.54254/2755-2721/2025.TJ22251

5. Conclusion

This paper proposes DSDF-Net, a medical imaging tumor detection framework based on a dynamic feature fusion mechanism, which achieves significant breakthroughs in both theoretical innovation and clinical applicability. By introducing cross-modal dynamic alignment theory and the Adaptive Routing Attention Mechanism (ARAM), DSDF-Net synergistically integrates CNN's local detail perception with Transformer's global contextual modeling capabilities. This novel architecture not only advances feature fusion methodologies but also demonstrates superior performance in complex clinical scenarios, such as detecting sub-5mm tumors and preserving boundary accuracy under low-contrast imaging conditions.

Acknowledgement

The authors gratefully acknowledge the financial support from the Chinese College Students' Innovation and Entrepreneurship Training Program (Grant No. 202410403003S).

References

[1] *Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28*

[2] *Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929*

[3] *Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., & Zhou, Y. (2021). TransUNet: Transformers make strong en coders for medical image segmentation. arXiv preprint arXiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04 306*

[4] *Zhang, Y., Liu, H., Hu, Q., He, Z., & Wang, Z. (2021). TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. arXiv preprint arXiv:2102.04306. https://doi.org/10.48550/arXiv.2102.04306*

[5] *Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., & Zhang, Z. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv preprint arXiv:2103.14030. https://doi.org/10.48550/arXiv.2103.14030*

[6] *Zhang, Y., Liu, H., Hu, Q., He, Z., & Wang, Z. (2021). Dual-Stream Representation Fusion Learning for Accurate Medical Image Segmentation. Engineering Applications of Artificial Intelligence, 105, 106402. https://doi.org/10.1016/j.engappai.2023.106402*

[7] *Su, H., Song, G., Huang, K., Wang, J., & Yang, M. (2024). Cross-Modal Attention Alignment Network with Auxilia ry Text Description for Zero-Shot Sketch-Based Image Retrieval. arXiv preprint arXiv:2407.00979. https://doi.org /10.48550/arXiv.2407.00979*

[8] *Zhang, M., Chang, K., & Wu, Y. (2023). Multi-Modal Semantic Understanding with Contrastive Cross-Modal Feature Alignment. arXiv preprint arXiv:2403.06355. https://doi.org/10.48550/arXiv.2403.06355*

[9] *Huang, W., Zhang, W., Quan, X., & Wang, J. (2022). A Two-Level Dynamic Adaptive Network for Medical Image Fusion. IEEE Transactions on Instrumentation and Measurement, 71, 1-1. https://doi.org/10.1109/TIM.2022.3169 546*

[10] *Liu, P., Fu, B., Ye, F., Yang, R., Xu, B., & Ji, L. (2022). DSCA: A Dual-Stream Network with Cross-Attention on W hole-Slide Image Pyramids for Cancer Prognosis. arXiv preprint arXiv:2206.05782. https://doi.org/10.48550/arXi v.2206.05782*