frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 03 October 2025
DOI 10.3389/fcomp.2025.1677905

OPEN ACCESS

EDITED BY
Zhou Zhou,
Changsha University, China

REVIEWED BY
Pengchen Liang,
Shanghai University, China
Lei Zhou,
National University of Defense Technology, China

*CORRESPONDENCE
Wei Shen
shenwei202403@126.com
Xiaowei Liu
675492062@qq.com

RECEIVED 01 August 2025
ACCEPTED 25 August 2025
PUBLISHED 03 October 2025

CITATION
Liu X, Tian J, Huang S and Shen W (2025) Enhancing medical image segmentation via complementary CNN-transformer fusion and boundary perception.
Front. Comput. Sci. 7:1677905.
doi: 10.3389/fcomp.2025.1677905

Enhancing medical image segmentation via complementary CNN-transformer fusion and boundary perception

Xiaowei Liu[1]*, Juanxiu Tian[1], Shangrong Huang[1] and Wei Shen[2]*

[1]School of Information Science and Engineering, Hunan Institute of Engineering, Xiangtan, China, [2]Coronary Care Unit, Nursing Department, The Third Xiangya Hospital of Central South University, Changsha, China

Introduction: Vision Transformers (ViTs) show promise for image recognition but struggle with medical image segmentation due to a lack of inductive biases for local structures and an inability to adapt to diverse modalities like CT, endoscopy, and dermatology. Effectively combining multi-scale features from CNNs and ViTs remains a critical, unsolved challenge.

Methods: We propose a Pyramid Feature Fusion Network (PFF-Net) that integrates hierarchical features from pre-trained CNN and Transformer backbones. Its dual-branch architecture includes: (1) a region-aware branch for global-to-local contextual understanding via pyramid fusion, and (2) a boundary-aware branch that employs orthogonal Sobel operators and low-level features to generate precise, semantic boundaries. These boundary predictions are iteratively fed back to enhance the region branch, creating a mutually reinforcing loop between segmenting anatomical regions and delineating their boundaries.

Results: PFF-Net achieved state-of-the-art performance across three clinical segmentation tasks. On polyp segmentation, PFF-Net attained a Dice score of 91.87%, surpassing the TransUNet baseline (86.96%) by 5.6% and reducing the HD95 metric from 22.25 to 11.68 (a 47.5% reduction). For spleen CT segmentation, it reached a Dice score of 95.33%, outperforming ESFPNet-S (94.92%) by 4.3% while reducing the HD95 from 6.99 to 3.35 (a 52.1% reduction). In skin lesion segmentation, our model achieved a Dice score of 90.29%, which represents a 7.3% improvement over the ESFPNet-S baseline (89.64%).

Discussion: The results validate the effectiveness of our pyramid fusion strategy and dual-branch design in bridging the domain gap between natural and medical images. The framework demonstrates strong generalization on small-scale datasets, proving its robustness and potential for accurate segmentation across highly heterogeneous medical imaging modalities.

KEYWORDS

boundary perception, CNNs, medical image segmentation, transformers, pretrained backbone

1 Introduction

With the continuous enhancement of computing power and the rapid advancement of deep learning, Convolutional Neural Networks (CNNs) have become an essential part of visual models. In the field of medical image processing, CNN-based methods have achieved significant success in recent years. These methods (Yao et al., 2024) typically use a U-shaped architecture with an encoder and decoder, which is common in medical image segmentation.

Increasing the depth of convolutional layers theoretically boosts the network's receptive field (i.e., the region of the input image that each neuron in a CNN layer responds to). By employing more convolutional layers (Zhou and Abawajy, 2025; Zhou et al., 2022b), abstract features over a larger image region can be acquired. However, factors like pooling, stride, dilated convolutions, computation limitation, and information propagation limit the effective receptive field (Ding et al., 2022). This limitation can affect the model's performance in tasks that necessitate a more global understanding of context. Therefore, better approaches that can capture long-range dependency information need to be developed for medical image segmentation analysis.

In contrast, Transformer models (Vaswani et al., 2017) have shown a distinct advantage in handling long-range dependency relationships. In recent years, a growing number of researchers have begun applying Transformers to the realm of medical image segmentation (Xia et al., 2025; Zhang et al., 2025). It's widely acknowledged that on large-scale datasets, Transformers exhibit greater adaptability compared to CNNs. However, Transformers encounter challenges, particularly due to the absence of certain inherent inductive biases present in CNNs, such as translation invariance and local correlation (Xu et al., 2025; Zhou et al., 2022a). This has prompted the development of Vision Transformers, commonly known as ViT, which necessitate more extensive data during the training phase. In contrast, CNNs extract features through localized sliding windows, enabling them to achieve remarkable results even with significantly smaller datasets. Once the convolutional kernels in CNNs are trained, their parameters remain fixed, whereas Transformers dynamically compute the correlation between pixels or features based on variations in input images.

In recent years, there has been a reciprocal exchange of ideas between Transformers and CNNs, resulting in notable advancements within their respective domains. For instance, CNNs have adopted the concept of dynamic parameters from Transformers, giving rise to innovative methodologies such as dynamic convolution and dynamic ReLU (Lou and Yu, 2025). Conversely, Transformers have benefited from insights gleaned from CNNs, employing localized self-attention computations to develop efficient models like the Swin Transformer (Liu et al., 2021). Furthermore, the emergence of the Pyramid Vision Transformer (PVT) (Kumar, 2025) is rooted in the feature pyramid concept prevalent in CNN architectures. Subsequently, Transformer-based backbone networks have predominantly retained the structural elements of local window attention and feature pyramid, thus highlighting the ongoing interaction and evolution between these two paradigms.

To reduce the data requirements of Transformer models, various methods have been proposed to apply Transformers to the field of medical image segmentation. In addition to the unit-level fusion solutions of CNNs and Transformers (Liu et al., 2023), there is also a strategy to concatenate or combine pre-trained Transformer modules with CNNs modules. In previous related research, many studies used pre-trained ViT as the backbone for feature extraction. For example, in Polyp-PVT (Qayoom et al., 2025), a Transformer-based feature pyramid structure, PVT (Kumar, 2025), was used as the backbone for feature extraction, and then the CNNs decoding module was used to fuse multi-scale features extracted by PVT. In SwinE-Net (Park and Lee, 2022), features from corresponding layers of EfficientNet (Kumar and Gunasundari, 2025) and Swin Transformer were fused by element-wise multiplication. Subsequently, these fused features, along with the segmentation mask maps predicted by EfficientNet and Swin Transformer, were input to the decoder to complete the final segmentation process. In addition, Shah et al. (2025) constructed a U-shaped structure entirely composed of Swin Transformer, known as Swin-Unet. The encoder directly used Swin Transformer as the backbone network, and both the encoder and decoder were initialized using pre-trained Swin Transformer. Furthermore, some research placed Transformers in other positions. For example, works like Transunet (Chen et al., 2024) and MedSAM (Lei et al., 2025) introduced pre-trained ViT at bottlenecks or skip connections in U-shaped structures based on CNNs for the fusion of global high-order features.

Pre-trained models are essential for deep learning, particularly for transformers, as they promote network convergence and improve segmentation accuracy. Supervised pre-training on extensive datasets like ImageNet and self-supervised techniques tailored for medical imaging applications have demonstrated enhancements in the subsequent performance of machine learning models (Lei et al., 2025; Rani et al., 2024). In this work, we do not dwell on the supervised pretraining details of backbones and directly use widely adopted ones (e.g., ResNet and Transformer variants) to construct our hybrid framework.

This study aims to integrate pyramid features from Transformer and CNN backbones to address their respective limitations while leveraging their complementary strengths. The main contributions are summarized as follows:

- Feature fusion module: we propose a module that aligns and integrates pyramid features from a pre-trained ResNet with those from a Transformer backbone. To ensure feature compatibility, ResNet channels are adjusted before fusion, and a cascaded decoder progressively combines multi-level features to refine segmentation outputs.
- Boundary-aware branch: to enhance boundary representation, we design a dedicated branch that exploits multi-scale features from the pre-trained backbone. Sobel filtering is applied to capture low-level edge cues from larger-scale features, which are then integrated with hierarchical features through an edge-aware module. The resulting boundary-weighted representation strengthens predictions near object boundaries.

- Comprehensive evaluation: we validate our method on colorectal polyp, spleen CT, and skin lesion datasets. The approach consistently outperforms existing methods, with particularly strong improvements in skin lesion and polyp segmentation tasks. These results provide new evidence for the effectiveness of pre-trained Transformers in medical image segmentation.

2 Related work

2.1 Macro-level fusion strategy of CNNs and transformers

The macro-level fusion strategy seeks to integrate the output features of pre-trained CNNs and Transformers. Distinguished by its higher level of abstraction compared with unit-level fusion techniques (Liu et al., 2023), this approach harnesses existing pre-trained models to their full potential. By merging the respective output features, commonly comprising hierarchical pyramid features, this strategy enhances model performance while seamlessly integrating diverse sources of information.

Figure 1 presents various representative macro-level fusion strategies for CNNs and Transformers in U-shaped architectures. The figure serves as a comparative reference and does not depict fusion structures; skip connections are also omitted for clarity.

The basic UNet (Ronneberger et al., 2015) employs CNNs in both encoder and decoder. In contrast, models such as SwinUNet (Cao et al., 2021) and Segmenter (Strudel et al., 2021) utilize Transformers in both encoder and decoder. Although Transformers are powerful in feature extraction from large-scale image data, their use as decoders is relatively uncommon due to difficulty in recovering fine details. SegFormer (Xie et al., 2021a), for instance, replaces the decoder with a simple multilayer perceptron (MLP), achieving a balance between accuracy and efficiency.

Another strategy applies Transformers in the encoder and CNNs in the decoder, as seen in SETR (Zheng et al., 2021) and ESFPNet (Chang et al., 2023). In some designs, Transformer modules are inserted at bottlenecks to integrate high-level features, exemplified by TransBTS (Wang et al., 2021a) and TransUNet.

CTC-Net (Yuan et al., 2023) combines multi-level features from both Transformer and CNN backbones with cross-scale feature fusion. Multi-scale features are first merged from both encoders, then together with the final Transformer encoder output, are fed into the Transformer decoder for segmentation.

Other methods employ Transformers as bridges between encoder and decoder to fuse and redistribute multi-scale features, as in CoTr (Xie et al., 2021b), MISSFormer (Huang et al., 2023), and TransCeption (Azad et al., 2023). MISSFormer's encoder-decoder further integrates CNNs and Transformer modules.

In TransFuse (Zhang et al., 2021), one encoder branch uses a Transformer directly connected to a CNN decoder. Its multi-scale features are fused with those from another CNN encoder branch before being passed to the final CNN decoder. HiFormer (Heidari et al., 2023) instead incorporates Transformer-extracted features into a CNN encoder and performs cross-fusion of the highest- and lowest-level features before decoding with CNNs. Our proposed method fuses same-level multi-scale features extracted by Transformer and CNN encoders, which are then fed into a CNN decoder. The fused features are also linked to an additional CNN decoder to capture boundary information, thereby assisting the final CNN decoder in producing refined segmentation results.

Recent studies further extend macro-level fusion strategies with new designs. For example, Gao et al. (2025) proposed CoT-UNet, combining Transformer modules with U-Net for global context and local detail preservation. Yan et al. (2025) introduced MCAFETrans, which leverages multiscale convolutional attention and frequency-domain features to enhance representation and segmentation performance across imaging modalities.

2.2 Leveraging pre-trained models for medical image segmentation

Both CNN and Transformer backbone networks are commonly initialized with pre-trained weights. This initialization enables the model to possess a comprehensive understanding of various visual features and semantic information right from the beginning of training. Leveraging pre-trained weights allows the model to acquire generic feature representations from extensive datasets, thereby expediting convergence speed and enhancing overall performance. Moreover, pre-trained weights facilitate transfer learning, enabling the model to adapt more readily to new tasks or data domains, leading to improved performance, especially in tasks like medical image segmentation.

In the domain of medical image segmentation, pre-trained weights are often trained on natural images. However, natural images and medical images belong to distinct domains with significant differences in modalities. Consequently, effectively applying these pre-trained weights to the domain of medical images emerges as a critical research direction.

In recent years, there has been considerable progress in the field of medical image segmentation by leveraging pre-trained Transformer models. For instance, Park and Lee (2022) introduced SwinE-Net, which utilizes pre-trained EfficientNet and pre-trained Swin Transformer as the feature extraction backbone, resulting in significant enhancements in polyp segmentation performance. Similarly, Dong et al. (2021) employed the pre-trained Pyramid Vision Transformer (PVT) (Wang et al., 2021b) for feature extraction, combined with multiple convolutional sub-networks, achieving remarkable results in polyp segmentation tasks. Building upon the success of the Swin Transformer, Cao et al. (2021) presented Swin-Unet, a pure Transformer-based U-shaped network tailored for 2D medical image segmentation. Both the encoder and decoder components of Swin-Unet were pre-trained on ImageNet, underscoring the importance of pre-training strategies in this domain. Notably, the encoder in Swin-Unet employs a sequential Swin Transformer, while the decoder adopts an inverted Swin Transformer architecture. Furthermore, Lin et al. (2022) developed DS-TransUNet, which leverages dual-scale encoders based on the Swin Transformer architecture to extract feature representations at both high and low resolutions. This innovative approach contributes significantly to the advancement of medical image segmentation techniques.

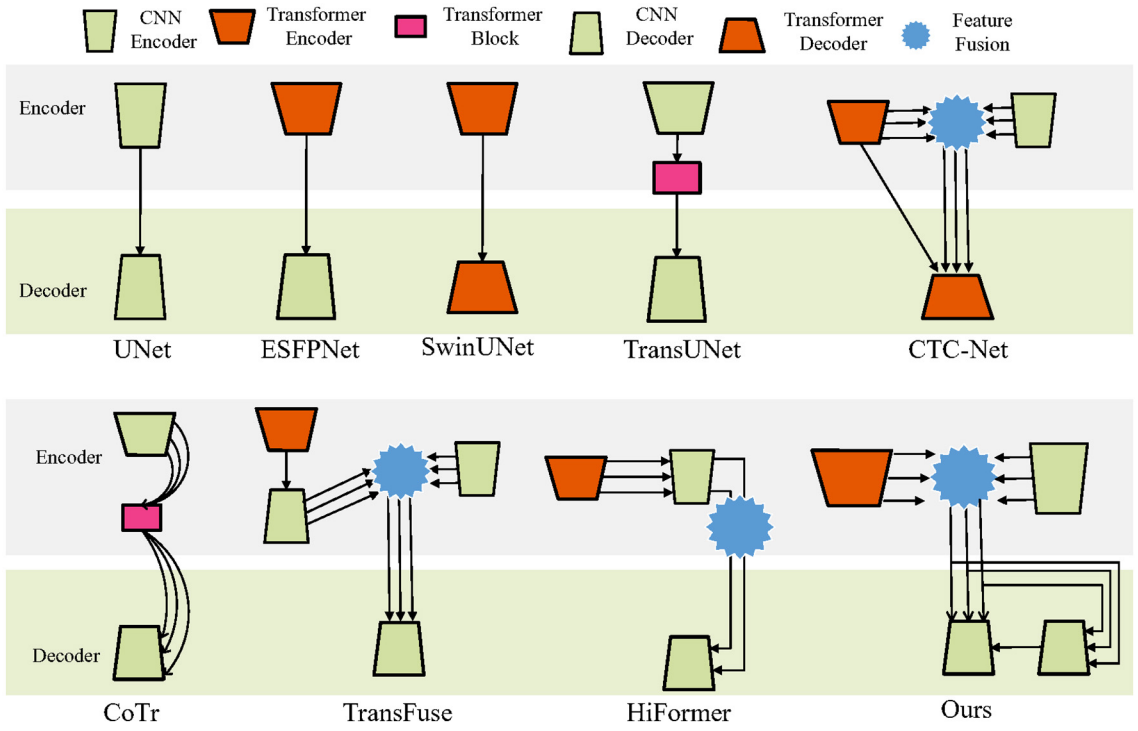

FIGURE 1
Macro-level fusion strategy of CNNs and transformers.

In addition to employing pre-trained vision transformers as feature extractors, researchers have investigated various integration strategies within medical segmentation networks. One such approach involves fusing features extracted by pre-trained vision transformers with those obtained by CNNs. Alternatively, transformers are integrated into different network components, such as the bottleneck of a U-Net or in parallel with a CNN branch.

Chen et al. (2024) pioneered the integration of transformers by incorporating a transformer block at the bottleneck of U-Net. This architectural modification enables the capture of long-distance dependencies in high-level features acquired during the segmentation process. Although this approach demonstrates the feasibility of transformer application in medical image segmentation, it requires pre-training on large datasets to achieve optimal performance. Similarly, Wang et al. (2021a) employed a similar structure, TransBTS, for multi-modal brain tumor segmentation. Zhang et al. (2021) proposed TransFuse, which comprises parallel CNN and transformer branches fused with customized modules to enhance segmentation accuracy.

Moreover, other studies, such as UNetr (Hatamizadeh et al., 2022), MISSFormer (Huang et al., 2023), and CoTr (Xie et al., 2021b), have integrated transformers within the U-Net architecture, albeit with variations in the placement of the transformer module.

While these methods share with our work the idea of exploiting complementary strengths of different architectures and optimizing for efficiency, our proposed PFF-Net differs in two key aspects: (1) it explicitly employs a dual-branch architecture with a boundary-aware branch that enhances structural precision through iterative feedback, and (2) it performs pyramid feature fusion across same-level representations from both CNN and Transformer backbones, enabling mutual reinforcement between region segmentation and boundary delineation. This design achieves superior performance across diverse modalities without relying on frequency-domain transformations, purely U-shaped collaborative structures, or domain-specific energy optimization strategies.

3 Proposed method

3.1 Overall architecture

The architecture of our method is delineated in Figure 2. This model is architectured around two encoders and two decoders. In detail, the encoders are constituted by a pre-trained ResNet34 and a pretrained Mix Transformer, tasked with the independent extraction of pyramid features. Following this, the multiscale features, as harvested by the CNN backbone, undergo channel adjustments and are subsequently amalgamated with analogous layer features from the Transformer, culminating in the formation of an enriched feature pyramid. On the decoder side, the model integrates a Region-aware Decoder and a Boundary-aware Decoder. The Boundary-aware Decoder is specifically designed to discern the perimeters of segmented entities. The insights garnered from boundary-aware processing, in concert with

the lateral output features emanating from the CNN decoder, are synergistically channeled into the Segmentation Head. This convergence plays a pivotal role in guiding and completing the segmentation task.

3.2 Detailed module design

3.2.1 Pre-trained dual encoders

The Pretrained Dual Encoder incorporates pre-trained encoders from both CNNs and Transformers. Specifically, the CNNs encoder utilizes ResNet34 (He et al., 2016), while the Transformer encoder employs Mix Transformer (Chang et al., 2023).

The structures of ResNet34 and Mix Transformer are illustrated in Figure 3. Specifically, ResNet34 (He et al., 2016) serves as an example for the CNNs encoder. Its output feature scales are 64 × 112 × 112, 64 × 56 × 56, 128 × 28 × 28, 256 × 14 × 14, and 512 × 7 × 7 (format: channels × width × height). Regarding the Transformer encoder, Mix Transformer (MiT) (Chang et al., 2023) is employed as an example, with output feature scales of 64×56×56, 128 × 28 × 28, 320 × 14 × 14, and 512 × 7 × 7.

Features with a size of 112 × 112 are extracted from ResNet34 (He et al., 2016). This step aims to enhance boundary perception and facilitate better detail restoration in the decoder. Features from the initial downsampling layer are rich in high-frequency information, akin to a common convolutional stem, which significantly differs from other layers in functionality, serving two primary purposes:

1. Learning local features, such as edges, for subsequent global/semantic feature extraction.
2. Downsampling feature maps in the initial layers reduces the workload of subsequent computations. While examining each pixel individually in the input may not be meaningful, extracting edge pixels requires rich local features encapsulating significant local spatial correlations.

To facilitate effective fusion of features extracted from ResNet34 (c_1, c_2, c_3, and c_4) with the multiscale features extracted by MiT (t_1, t_2, t_3, and t_4), a Channel Modification Module (CMM) is introduced. The primary function of this module is to adjust channels before feature fusion, thereby achieving feature transformation and alignment simultaneously. This alignment ensures that CNNs features are brought into the same space as Transformer features, enabling their effective integration despite originating from different sources. However, it is noteworthy that the design of the channel adjustment module initially aims to address disparities in feature channel numbers, enhancing compatibility between features and producing superior fusion results in subsequent fusion processes.

After the channel adjustment process, the lightweight Complementary Fusion Module (CFM), illustrated in Figure 4, developed in this study, can be employed for feature fusion. This module is responsible for combining features from both the CNN backbone and the Transformer backbone. Specifically, features from the CNN backbone undergo channel adjustment before fusion.

3.2.2 Boundary-aware decoder

The Boundary-Aware Decoder, depicted above in Figure 2, is responsible for extracting boundary information from both high-level and low-level features. This decoder aims to filter out irrelevant boundary details, recognizing the importance of high-resolution features for accurate boundary delineation. Specifically, it utilizes the 1/2 and 1/4 high-resolution features obtained from the CNNs encoder to identify boundaries. The Sobel operator (Kanopoulos et al., 1988) is applied in both horizontal (K_x) and vertical (K_y) directions on the initial two layers (with sizes of 112 × 112 and 56 × 56) to generate gradient maps. This operation is achieved using two 3×3 fixed-parameter convolutional kernels with a stride of 1 for convolution. The definitions of these convolutional kernels, K_x and K_y, are as follows:

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{1}$$

The Sobel convolution employs fixed convolutional kernels to generate gradient maps B_0 and B_1, which may contain semantically irrelevant information. However, for enhanced precision in localizing segmented object boundaries, the Boundary-Aware Decoder also incorporates the high-level feature B_2. These features (B_0, B_1, and B_2) are subsequently fed into the Boundary Perception Head (BPH), where boundary supervision is conducted at this stage.

The detailed design of the Boundary Perception Head (BPH) is illustrated in Figure 5. Here, x_3, x_2, and x_1 correspond to B_0, B_1, and B_2 in Figure 2, respectively. Initially, features of these three scales undergo channel adjustment through a 1 × 1 convolution. Subsequently, they undergo upsampling operations, with a factor of 16 and 2 applied to x_1 and x_2 processing, respectively. These two features are then concatenated along the channel dimension. Following this, they pass through two basic convolutional modules, comprising convolution operations, batch normalization, and ReLU activation. Lastly, a 1 × 1 convolutional layer adjusts the channel number to 1, and the Sigmoid function is applied to extract boundary attention for further supervision. It is crucial to note that the Sigmoid activation function is not applied before feeding the boundary-aware results into the final segmentation head. This process can be formalized as follows:

$$x_1 = PWConv(x_1), \tag{2}$$
$$x_2 = PWConv(x_2), \tag{3}$$
$$x_3 = PWConv(x_3), \tag{4}$$
$$x_c = Concat(x_1, x_2, x_3), \tag{5}$$
$$out = \sigma(PWConv(ConvBlock(ConvBlock(x_c)))), \tag{6}$$

where $PWConv(\cdot)$ denotes point-wise convolution, specifically a 1 × 1 convolution as illustrated in Figure 5. *ConvBlock* represents a standard convolutional module. σ denotes the Sigmoid activation function.

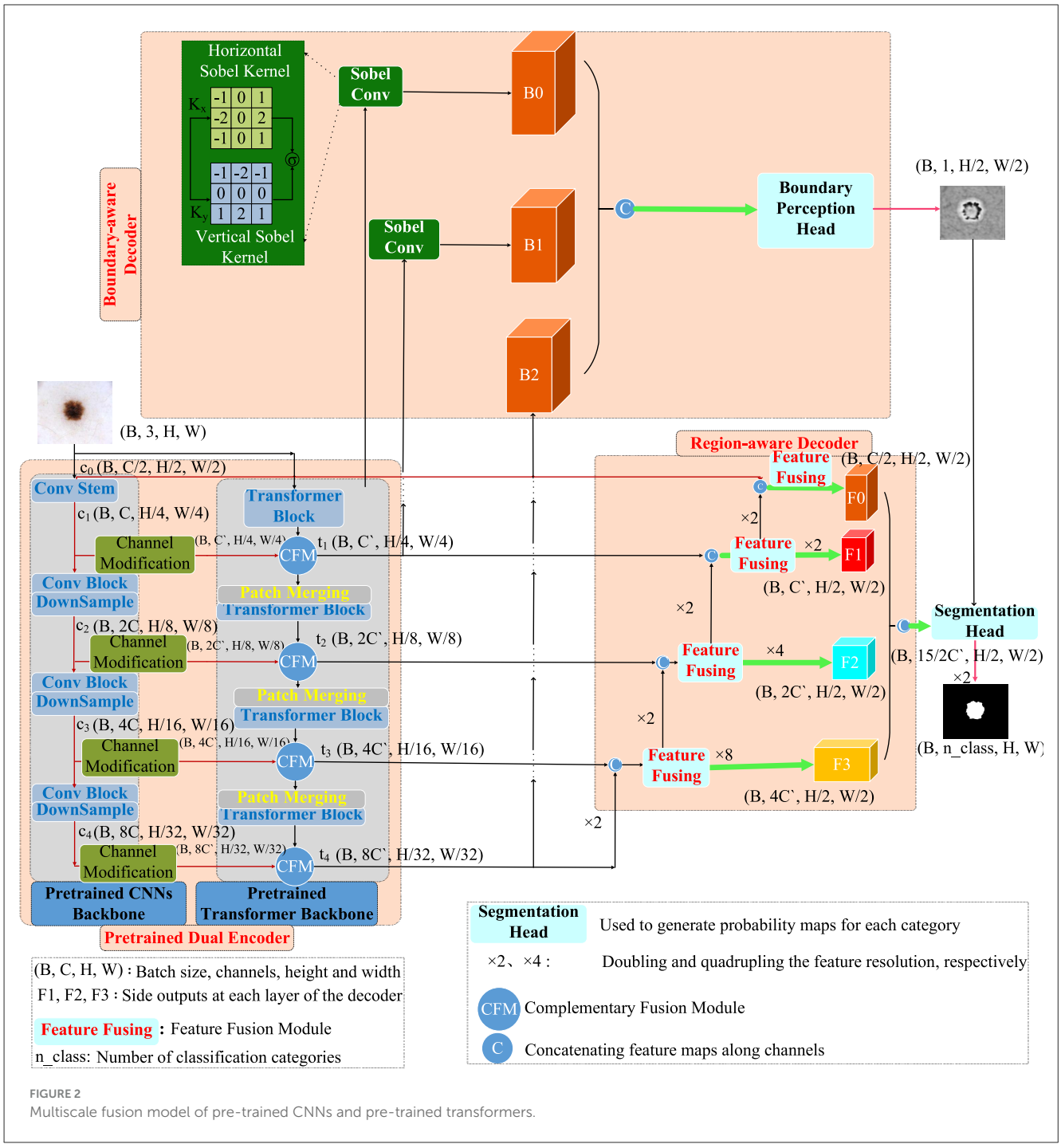

FIGURE 2
Multiscale fusion model of pre-trained CNNs and pre-trained transformers.

3.2.3 Region-aware decoder

The Region-Aware Decoder serves as the decoder component intended for performing region segmentation tasks using the pyramid features fused by the dual encoders. Illustrated in Figure 2, the structure of this decoder closely resembles that of traditional decoders found within U-shaped architectures series.

In the Region-Aware Decoder, each layer of features undergoes a 2× upsampling from bottom to top, followed by concatenation with the previous layer's features, facilitating multi-scale feature fusion. This fusion is achieved through point-wise convolution, resulting in a fused output. Subsequently, the fused result undergoes another 2× upsampling and repeats the process iteratively until the decoding process is complete.

The design of this decoder draws inspiration from classic structures such as UNet. Through the process of upsampling and feature fusion, it contributes to the restoration of high-resolution region information from features extracted by the lower-level encoders. By integrating pyramid features and feature fusion, the Region-Aware Decoder enhances its ability to perform region segmentation tasks in medical images.

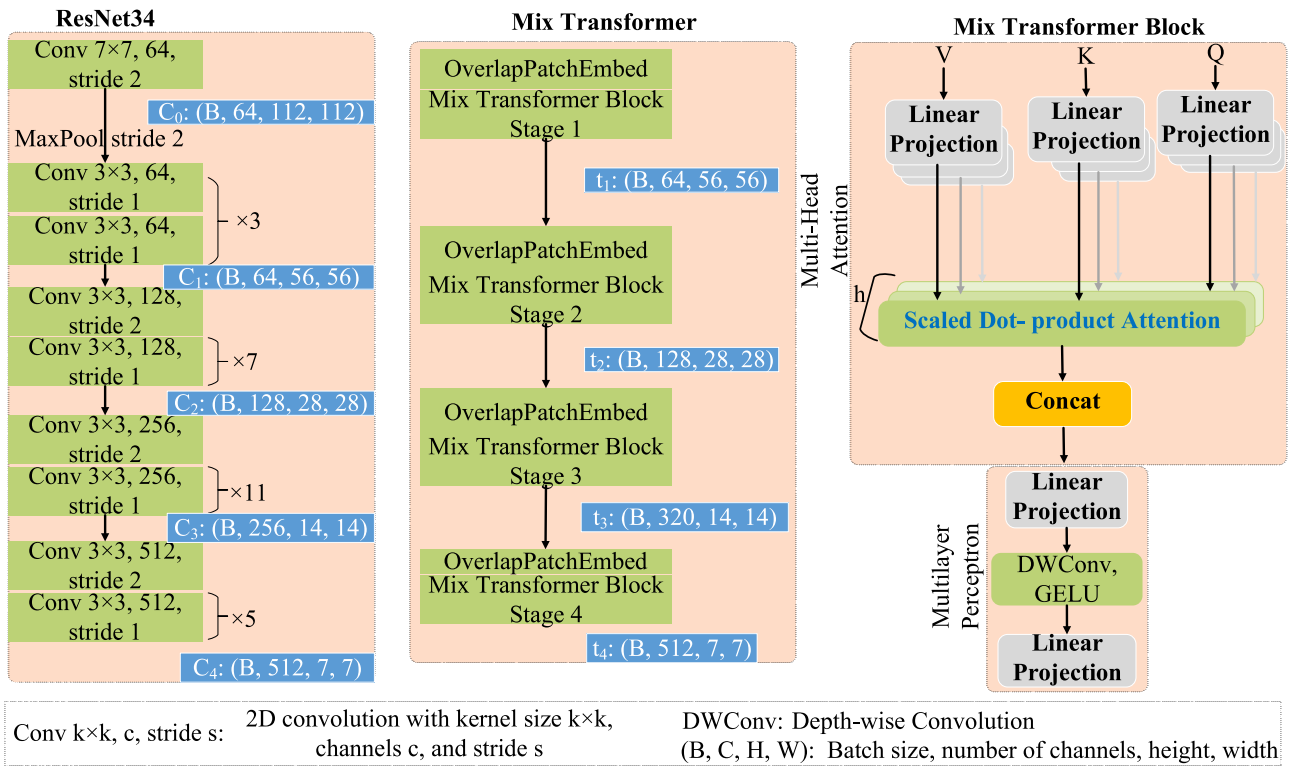

FIGURE 3
Architecture of ResNet34 and mix transformer.

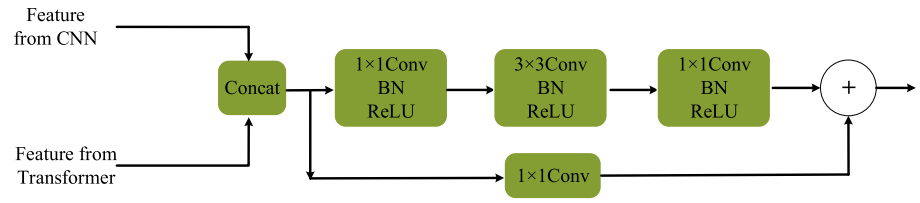

FIGURE 4
Complementary fusion module.

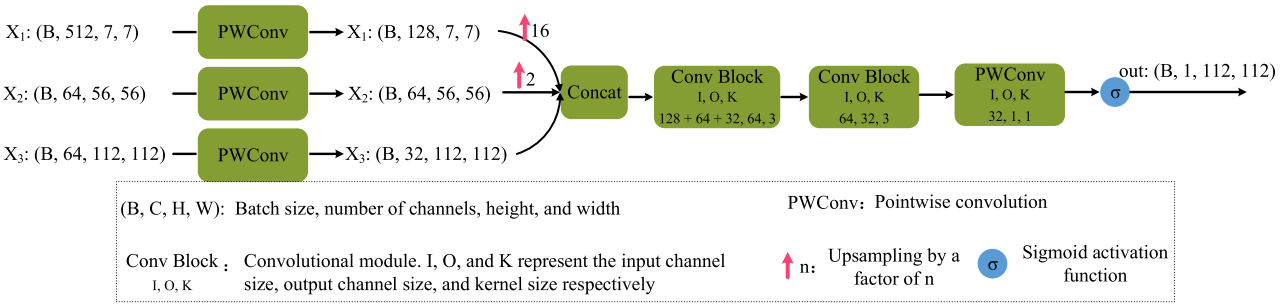

FIGURE 5
Boundary perception head.

3.2.4 Semantic segmentation head

To fully exploit the boundary-aware results, a crucial component, the final Semantic Segmentation Head (SH), is devised. The purpose of this component is to integrate multiple side outputs from the region decoder with the boundary-aware results, thereby achieving comprehensive semantic segmentation of medical images.

In this procedure, side outputs at different scales (F_0, F_1, F_2, and F_3) undergo resizing to ensure uniform dimensions. Subsequently, these adjusted features are concatenated along the channel dimension to create a comprehensive feature representation. Ultimately, these concatenated features are fed into the Semantic Segmentation Head for the final semantic segmentation process.

Furthermore, the Semantic Segmentation Head integrates the boundary-aware results to enhance the concatenated features of all side outputs. This process can be mathematically expressed as:

$$O = Conv\left(\text{Concat}(F_0, F_1, F_2, F_3) \odot (1 + BPH)\right). \quad (7)$$

Here, BPH denotes the output of the Boundary-aware Decoder, also known as the Boundary Perception Head, representing a weight matrix pertaining to the boundaries of segmented objects. The term $1 + BPH$ enhances features near the boundaries without explicitly filtering out other features. F_0, F_1, F_2, and F_3 correspond to the side outputs of the Region-aware Decoder, as depicted in Figure 2, resized to a consistent scale ($H/2 \times W/2$). The operation Concat signifies feature concatenation along the channel dimension, while $\odot$ represents element-wise multiplication (Hadamard product). Subsequently, employing a 1×1 2D convolution, known as point-wise convolution, the channel dimension is mapped to a specific number of segmentation classes. Finally, the outcome undergoes a $2\times$ upsampling to yield the final segmentation result.

This design ensures that in completing the semantic segmentation task, not only is multi-scale information from the region decoder fully considered, but also the boundary-aware results are fully utilized, thereby enhancing the performance of semantic segmentation for medical images.

3.3 Loss functions

Our approach implements supervision in two areas. The first is the boundary-aware decoder, which perceives the boundaries of segmented objects. The second is the final region segmentation, specifically at the output of the semantic segmentation head after the region-aware decoder.

3.3.1 Boundary-aware loss

Compared to other regions, boundaries occupy a smaller proportion. Therefore, we employ a weighted cross-entropy loss function:

$$L_{\text{boundary}} = -\frac{1}{N}\sum_{i=1}^{N}\left(w_b B_g^{(i)} \log(B_p^{(i)}) + w_{\neg b}(1 - B_g^{(i)}) \log(1 - B_p^{(i)})\right). \quad (8)$$

Here, N represents the total number of samples within a batch. $B_g^{(i)}$ denotes the actual boundary of the i-th sample, while $B_p^{(i)}$ refers to the predicted boundary. Moreover, w_b and $w_{\neg b}$ signify the weights allocated to positive and negative samples, respectively. The determination of w_b and $w_{\neg b}$ follows the outlined procedure:

$$w_b = \frac{N_{\neg b}}{N_b},\ w_{\neg b} = \frac{N_b}{N_{\neg b}}, \quad (9)$$

where N_b represents the number of positive samples, and $N_{\neg b}$ is the number of negative samples. This method employs class frequency-weighting, where weights are determined computed on the number of samples in each class. If the number of samples in a class is high, its weight can be set to a smaller value; conversely, if the number of samples in a class is low, its weight can be set to a larger value. This adjustment of weights effectively balances the influence of different classes during training.

3.3.2 Region segmentation loss

In the semantic segmentation head, located to the right of the region-aware decoder, a combination of the commonly used Dice loss and cross-entropy loss is employed. Specifically, it is formulated as follows:

$$L = L_{Dice} + L_{CE}, \quad (10)$$

where L_{Dice} and L_{CE} refer to the Dice loss and cross-entropy loss between the model's predicted output and the ground truth, respectively. It is noteworthy that, in the experiments, the criterion for selecting the optimal model weight is based solely on the Dice score. That is, the weight with the highest Dice score during the validation phase is selected for saving.

4 Experiments

This section commences with an overview of the datasets employed in the experiments, followed by an elaborate description of the experimental setup and evaluation criteria. Subsequently, performance comparisons are made with other prevalent models, and ablation experiments are conducted on various components of the model. Finally, corresponding conclusions are drawn based on the experimental findings.

4.1 Datasets

4.1.1 ISIC 2018

This experiment revolves around Task 1 of the ISIC 2018 Challenge [1], which primarily deals with segmenting skin lesion regions. The challenge comprises three tasks: lesion segmentation, lesion attribute detection, and disease classification. ISIC 2018 dataset comprises RGB images of diverse sizes, which have been resized to 224×224 for experimental convenience. We exclusively utilize 2,589 training images from the Challenge, employing 5-fold cross-validation in our experiments.

1 https://challenge2018.isic-archive.com

4.1.2 Spleen segmentation CT

Furthermore, we participate in the Medical Segmentation Decathlon Challenge [2], specifically in Task 9, which involves spleen segmentation. This subset comprises a total of 61 portal venous phase CT scans, with 41 scans allocated for training and 20 for testing. Each comparison model is trained on the 2D slices of these CT scans in the transverse plane, also known as the axial view. The raw images' scale of the 2D slices is downscaled to 224×224. However, the spacing parameter remains unchanged during the scaling process, and the original spacing values provided by the CT scans are used in distance calculations.

4.1.3 Polyp segmentation

For the polyp segmentation task, several public datasets are utilized, including Kvasir SEG (Jha et al., 2020), CVC-ClinicDB (Bernal et al., 2015), CVC-ColonDB (Tajbakhsh et al., 2015), CVC-T (Vázquez et al., 2017), and ETIS (Silva et al., 2014). These datasets are used to validate the effectiveness of the proposed model. Specifically:

- The Kvasir SEG dataset comprises 1,000 polyp images along with corresponding ground truth masks annotated by endoscopy experts. Images in the Kvasir SEG dataset have resolutions ranging from 332×487 to $1,920 \times 1,072$ pixels.
- The CVC-ClinicDB dataset, also known as CVC-612, provides 612 publicly available images from 25 colonoscopy videos at a resolution of 384×288 pixels.
- The CVC-ColonDB dataset contains 380 images extracted from 15 colonoscopy videos at a resolution of 574×500 pixels.
- The ETIS dataset comprises 196 images extracted from 34 colonoscopy videos at a resolution of $1,225 \times 996$ pixels.
- The CVC-T dataset, a subset of EndoScene, includes 60 images from 44 colonoscopy sequences involving 36 patients at a resolution of 574×500 pixels.

For ease of training and testing, the images from these datasets are resized to 224×224 due to resolution discrepancies. It is important to note that, since these images are from different datasets, a stratified sampling method is employed during the 5-fold cross-validation to maintain proportional distribution of samples across datasets.

4.2 Experimental setup and evaluation criteria

4.2.1 Experimental setup

The experimental setup, including both hardware and software, is detailed in Table 1.

During the training phase, a uniform batch size of 16 was upheld, utilizing the AdamW optimizer with an initial learning rate of 1×10^{-4}, subject to exponential decay over successive iterations. Noteworthy is that the hybrid Transformer model underwent 250 epochs of training, while the CNN-based model converged within 150 epochs. In contrast, SwinUNet was trained from scratch for a total of 400 epochs.

4.2.2 Evaluation criteria

In our experimental analysis, seven metrics are employed to evaluate the efficacy of the comparative methodologies, namely Dice coefficient, 95th percentile Hausdorff distance (HD95), Jaccard similarity coefficient, precision, accuracy, specificity, and sensitivity. It is noteworthy that due to the nature of the input data, which consists of either 2D images or 2D slices extracted from 3D images, the Average Surface Distance (ASD) metric is not included in the evaluation framework of this investigation.

4.3 Comparison experiments with other state-of-the-art methods

4.3.1 Comparative methods

The experimental setup mirrors that of Liu et al. (2023), including the hardware platform. The compared methods are categorized into two main groups: those based on CNNs and those based on Transformers, or hybrid models combining both. Among CNN-based methods are U-Net (Ronneberger et al., 2015), U-Net++ (Zhou et al., 2018), ResUNet, Attention U-Net (Oktay et al., 2018), DAUNet (Fu et al., 2019), and R2B (Liu et al., 2022), each offering distinct architectural enhancements. Conversely, Transformer-based methods include TransUNet (Chen et al., 2024), SwinUNet (Cao et al., 2021), UTNet V2 (Gao et al., 2022), ViTAE V2 (Zhang et al., 2023), MISSFormer (Huang et al., 2023), PVT-CASCADE (Rahman and Marculescu, 2023), ESFPNet (Chang et al., 2023), HiFormer (Heidari et al., 2023), and PyramidalConv (Liu et al., 2023), each integrating Transformer structures in diverse ways, such as replacing or enhancing traditional CNN components or fusing them with Transformer units.

4.3.2 Quantitative results

For the Spleen CT dataset, the mean and variance of performance evaluation metrics are assessed for the entire 3D image rather than individual slices. Conversely, for the ISIC 2018 and polyp segmentation, performance evaluation is conducted based on each independent medical image. This differs from the evaluation method employed in UTNet V2, where the statistics are calculated for each fold in 5-fold cross-validation, i.e., averaging and calculating the standard deviation of the 5-fold performance.

4.3.2.1 Evaluation on the polyp dataset

We combined five public polyp datasets into a single cohort for evaluation. The performance of each model was assessed via 5-fold cross-validation, and the averaged results are presented in Table 2.

The results on the polyp dataset (Table 2) reveal several critical findings. First, CNN-based models such as U-Net, U-Net++, and ResUNet deliver solid and consistent performance, with R2B achieving the best Dice (87.29%) among CNNs by explicitly modeling boundary information. Nevertheless, CNNs still fall short of capturing long-range dependencies, which limits their overall

2 https://decathlon-10.grand-challenge.org/

TABLE 1 Software and hardware for experiments.

Software	Operiting system	CUDA version	Python version	Pytorch version
	Ubuntu 18.04	11.8	3.8	2.1.0
Hardware	CPU	RAM	GPU	VRAM
	E5-2678 v3	64GB DDR4	1*NVIDIA Tesla V100	32G

segmentation accuracy compared to Transformer-based or hybrid approaches.

Second, pretrained Transformer variants, including PVT-CASCADE*, HiFormer-B*, and ESFPNet-S*, demonstrate substantial advantages, with Dice scores above 89% and markedly lower HD95 values. This highlights the importance of pretraining on large-scale natural image datasets, whose feature distributions are relatively close to polyp images, thereby providing a strong initialization for Transformer-based encoders.

Third, our proposed PFF-Net* achieves the best results across all evaluation metrics, with Dice (91.87%), Jaccard (86.58%), and HD95 (11.68) outperforming all baselines by a clear margin. Its superior boundary perception, evidenced by both low HD95 and balanced Sensitivity (93.39%) and Precision (91.93%), confirms the effectiveness of our pyramid fusion mechanism and boundary-aware branch in integrating global Transformer context with fine-grained CNN features.

Finally, some models underperform despite pretraining. For instance, SwinUNet shows significantly lower Dice (80.14%), likely due to suboptimal decoder initialization and the reversed Swin Transformer design, which lacks detailed boundary modeling. Similarly, MISSFormer and UTNet V2 struggle to generalize well, indicating that not all hybrid or Transformer-based models are universally suited for polyp segmentation. In contrast, PFF-Net demonstrates consistent and superior performance across all metrics, establishing its robustness and reliability for clinical polyp detection and segmentation tasks.

4.3.2.2 Evaluation on the ISIC 2018 dataset

Similarly, 5-fold cross-validation was performed on the ISIC 2018 dataset, and the experimental results are summarized in Table 3.

The experimental results on the ISIC 2018 dataset reveal several critical insights, some of which are consistent with observations on polyp datasets. First, notable performance disparities exist among different architectural families. CNN-based models, such as U-Net and its variants, demonstrate robust and consistent performance, establishing strong baselines. In contrast, the pure Transformer model (SwinUNet) exhibits significantly inferior performance (Dice: 83.13%; HD95: 32.64), highlighting the challenges pure Transformers face in learning effective representations from limited medical data without the inductive biases inherent in CNNs.

Conversely, hybrid architectures that integrate CNNs and Transformers, such as TransUNet and our proposed PFF-Net, dominate the top rankings. This underscores the effectiveness of combining local fine-grained features with global contextual dependencies in medical image segmentation. It is worth noting that models marked with * (e.g. TransUNet*, ESFPNet-S*) show substantial improvements over their base versions, suggesting that pre-training on large-scale datasets is a powerful strategy to mitigate overfitting and fully leverage the potential of Transformer-based models.

Among all models, our PFF-Net achieves state-of-the-art performance, ranking first in Dice (90.29%), HD95 (11.02), Jaccard (83.75%), Sensitivity (92.39%), and Accuracy (96.40%). The exceptionally low HD95 value, along with its small standard deviation, indicates that PFF-Net not only provides accurate segmentation on average but also exhibits superior robustness when handling challenging cases with ambiguous boundaries or irregular shapes. Furthermore, PFF-Net strikes an excellent balance between Sensitivity and Precision (92.39% and 90.92%, respectively), demonstrating its capability to effectively reduce both false negatives and false positives. This comprehensive advantage is attributed to our novel core technique, a Pyramid Fusion mechanism integrating pre-trained Transformer and CNN backbones with boundary awareness, which enables precise boundary delineation and effective multi-scale feature fusion.

Although specialized models such as PyramidalConv excel in Specificity (96.69%) and ESFPNet-S* leads in Precision (91.01%), their performance in other metrics is not as uniformly outstanding as that of PFF-Net. This consistent excellence across all evaluation metrics solidifies PFF-Net's position as a superior and reliable solution for skin lesion segmentation tasks.

4.3.3 Evaluation on the spleen segmentation CT dataset

Finally, 5-fold cross-validation was conducted on the CT dataset for spleen segmentation. The primary challenge of this task lies in the relatively low proportion of spleen presence, with most slices lacking spleen structures. Nearly all methods exhibited relatively poor validation results in the initial 20 epochs of training. Notably, SwinUNet performed inadequately on this task and is therefore omitted from Table 4. However, except for SwinUNet, the performance of all other methods is included in Table 4.

The experimental results on the spleen CT dataset provide several important insights distinct from those observed in natural-like modalities such as polyps and skin lesions. First, while CNN-based models (e.g., U-Net, U-Net++, ResUNet) establish strong baselines with Dice scores above 92%, they primarily benefit from inductive biases well-suited for intensity-based CT images. Among these, R2B stands out with the best Dice (95.45%) and lowest HD95 (3.62), confirming the effectiveness of boundary modeling in organ segmentation.

Second, Transformer-based and hybrid models show more varied performance in this task. Unlike in polyp or skin lesion datasets, models pretrained on natural images (e.g., PVT-CASCADE*, HiFormer-B*) do not dominate, reflecting the modality gap between single-channel CT scans and RGB natural

TABLE 2 Performance comparison on the Polyp dataset.

Arch.	Model	AVG						
		Dice%↑	HD95↓	Jaccard%↑	Specificity%↑	Sensitivity%↑	Precision%↑	Accuracy%↑
CNNs-based model	U-Net	86.33 ± 0.20	31.08 ± 66.74	79.71 ± 0.22	97.20 ± 0.12	88.19 ± 0.21	87.28 ± 0.20	96.11 ± 0.13
	U-Net++	86.94 ± 0.20	31.90 ± 70.91	80.54 ± 0.21	97.10 ± 0.13	88.21 ± 0.20	88.46 ± 0.20	95.93 ± 0.14
	ResUNet	86.46 ± 0.20	29.16 ± 74.01	80.01 ± 0.22	96.56 ± 0.14	88.54 ± 0.21	87.47 ± 0.20	95.57 ± 0.15
	Attention U-Net	86.07 ± 0.20	33.06 ± 71.19	79.29 ± 0.22	96.85 ± 0.13	88.29 ± 0.20	87.03 ± 0.20	95.83 ± 0.13
	DAUNet	86.10 ± 0.20	28.91 ± 57.82	79.25 ± 0.22	97.43 ± 0.10	88.31 ± 0.20	87.36 ± 0.20	96.31 ± 0.11
	R2B	87.29 ± 0.19	23.61 ± 64.45	80.89 ± 0.21	97.11 ± 0.13	88.86 ± 0.20	88.37 ± 0.19	96.23 ± 0.13
Transformer/Hybrid	TransUNet	86.00 ± 0.20	27.76 ± 65.97	79.21 ± 0.22	96.99 ± 0.13	88.06 ± 0.20	87.16 ± 0.21	95.93 ± 0.13
	TransUNet*	86.96 ± 0.19	22.25 ± 52.21	80.34 ± 0.21	97.70 ± 0.10	89.20 ± 0.20	87.82 ± 0.19	96.75 ± 0.10
	SwinUNet*	80.14 ± 0.25	35.17 ± 65.83	72.07 ± 0.26	96.49 ± 0.13	83.22 ± 0.25	81.61 ± 0.25	95.10 ± 0.13
	UTNet V2	85.89 ± 0.21	27.63 ± 68.63	79.21 ± 0.22	96.85 ± 0.13	87.61 ± 0.21	87.43 ± 0.21	95.78 ± 0.14
	ViTAE V2	86.18 ± 0.20	24.50 ± 61.88	79.46 ± 0.22	97.18 ± 0.12	88.00 ± 0.21	87.36 ± 0.20	96.17 ± 0.12
	PVT-CASCADE*	90.37 ± 0.15	14.52 ± 39.40	84.60 ± 0.17	98.35 ± 0.08	91.89 ± 0.15	90.75 ± 0.15	97.72 ± 0.08
	MISSFormer	84.09 ± 0.21	27.61 ± 55.83	76.70 ± 0.23	97.24 ± 0.11	86.45 ± 0.22	85.30 ± 0.22	96.15 ± 0.11
	HiFormer-B	82.60 ± 0.23	33.05 ± 79.19	75.07 ± 0.24	95.64 ± 0.16	85.08 ± 0.24	84.04 ± 0.24	94.60 ± 0.16
	HiFormer-B*	89.91 ± 0.15	15.30 ± 38.51	84.03 ± 0.17	98.42 ± 0.07	91.48 ± 0.16	90.41 ± 0.15	97.71 ± 0.07
	ESFPNet-S	85.82 ± 0.20	22.93 ± 53.27	78.91 ± 0.22	97.59 ± 0.10	87.83 ± 0.21	87.05 ± 0.20	96.63 ± 0.10
	ESFPNet-S*	90.32 ± 0.15	15.47 ± 42.57	84.60 ± 0.17	98.34 ± 0.08	91.88 ± 0.15	90.72 ± 0.15	97.71 ± 0.08
	PyramidalConv	88.86 ± 0.18	19.51 ± 51.52	83.04 ± 0.20	97.96 ± 0.10	90.43 ± 0.18	89.92 ± 0.17	97.09 ± 0.10
	PFF-Net*	**91.87 ± 0.12**	**11.68 ± 27.07**	**86.58 ± 0.15**	**98.78 ± 0.05**	**93.39 ± 0.12**	**91.93 ± 0.13**	**98.26 ± 0.05**

Each entry in the table is presented as the mean ± standard deviation of the corresponding metric. The best result in each column is highlighted in bold black.

* indicates that the model utilized pre-trained weights on the ImageNet21K dataset or pre-trained weights published in relevant literature. The bold values represent the best or second-best performance, facilitating the comparison of the methods corresponding to the top-performing results.

10.3389/fcomp.2025.1677905

TABLE 3 Performance comparison on the ISIC 2018 dataset.

Arch.	Model	AVG						
		Dice%↑	HD95↓	Jaccard%↑	Specificity%↑	Sensitivity%↑	Precision%↑	Accuracy%↑
CNNs-based model	U-Net	87.88 ± 0.14	17.07 ± 22.47	80.43 ± 0.17	95.97 ± 0.08	91.31 ± 0.13	88.32 ± 0.16	95.21 ± 0.07
	U-Net++	88.14 ± 0.13	17.13 ± 24.62	80.73 ± 0.16	96.00 ± 0.08	91.58 ± 0.13	88.40 ± 0.16	95.38 ± 0.07
	ResUNet	88.29 ± 0.14	15.12 ± 20.68	81.10 ± 0.17	95.79 ± 0.10	91.28 ± 0.13	89.06 ± 0.16	95.19 ± 0.08
	Attention U-Net	87.89 ± 0.13	16.27 ± 21.23	80.31 ± 0.16	96.11 ± 0.08	91.40 ± 0.14	88.18 ± 0.15	95.22 ± 0.07
	DAUNet	87.93 ± 0.15	17.03 ± 26.23	80.73 ± 0.18	95.97 ± 0.09	91.31 ± 0.13	88.49 ± 0.17	95.06 ± 0.08
	R2B	88.95 ± 0.13	13.16 ± 17.27	81.94 ± 0.16	95.71 ± 0.11	91.42 ± 0.13	90.03 ± 0.15	95.47 ± 0.08
Transformer/Hybrid	TransUNet	88.81 ± 0.13	14.16 ± 19.07	81.80 ± 0.16	96.28 ± 0.09	91.11 ± 0.14	90.02 ± 0.15	95.63 ± 0.07
	TransUNet*	89.41 ± 0.12	12.37 ± 15.65	82.51 ± 0.15	96.10 ± 0.10	91.41 ± 0.14	90.62 ± 0.13	95.95 ± 0.07
	SwinUNet	83.13 ± 0.19	32.64 ± 42.11	74.41 ± 0.21	95.71 ± 0.09	86.34 ± 0.20	84.96 ± 0.19	93.86 ± 0.10
	SwinUNet*	87.62 ± 0.14	18.39 ± 25.99	80.09 ± 0.17	96.24 ± 0.09	90.02 ± 0.15	89.05 ± 0.16	95.19 ± 0.08
	ViTAE V2	88.82 ± 0.13	13.41 ± 17.91	81.80 ± 0.16	96.08 ± 0.09	91.32 ± 0.14	89.92 ± 0.15	95.57 ± 0.08
	UTNet V2	88.14 ± 0.15	14.81 ± 20.29	81.05 ± 0.18	95.59 ± 0.11	91.58 ± 0.13	88.83 ± 0.17	95.20 ± 0.09
	MTUNet	87.67 ± 0.14	16.23 ± 28.40	80.18 ± 0.17	95.21 ± 0.11	91.66 ± 0.15	87.67 ± 0.16	94.95 ± 0.09
	MISSFormer	87.17 ± 0.15	17.52 ± 23.17	79.69 ± 0.18	95.89 ± 0.09	90.63 ± 0.15	88.26 ± 0.16	95.08 ± 0.08
	PVT-CASCADE*	87.01 ± 0.15	15.11 ± 20.44	79.36 ± 0.18	95.31 ± 0.11	91.92 ± 0.13	86.87 ± 0.18	94.84 ± 0.09
	ESFPNet-S	85.74 ± 0.16	21.03 ± 26.86	77.80 ± 0.19	95.85 ± 0.09	90.18 ± 0.15	86.61 ± 0.19	94.43 ± 0.09
	ESFPNet-S*	89.64 ± 0.12	12.03 ± 16.25	82.95 ± 0.15	96.28 ± 0.10	91.52 ± 0.13	**91.01 ± 0.14**	96.06 ± 0.07
	HiFormer-B	88.62 ± 0.13	13.23 ± 17.00	81.48 ± 0.16	96.24 ± 0.10	90.49 ± 0.14	90.42 ± 0.15	95.58 ± 0.08
	HiFormer-B*	89.58 ± 0.12	12.36 ± 18.58	82.81 ± 0.15	96.39 ± 0.09	91.70 ± 0.13	90.54 ± 0.13	96.09 ± 0.07
	PyramidalConv	89.68 ± 0.12	12.07 ± 15.94	82.83 ± 0.15	**96.69 ± 0.08**	92.15 ± 0.12	90.21 ± 0.13	96.10 ± 0.06
	PFF-Net	**90.29 ± 0.11**	**11.02 ± 14.20**	**83.75 ± 0.14**	96.42 ± 0.09	**92.39 ± 0.12**	90.92 ± 0.13	**96.40 ± 0.06**

* indicates that the model utilized pre-trained weights on the ImageNet21K dataset or pre-trained weights published in relevant literature. The bold values represent the best or second-best performance, facilitating the comparison of the methods corresponding to the top-performing results.

10.3389/fcomp.2025.1677905

TABLE 4 Performance comparison of various models on the spleen segmentation CT dataset.

Architecture	Model	Average						
		Dice%↑	HD95↓	Jaccard%↑	Specificity%↑	Sensitivity%↑	Precision%	Accuracy%↑
CNNs-based model	U-Net	92.83 ± 0.05	15.55 ± 31.91	87.00 ± 0.08	**99.98 ± 0.00**	92.79 ± 0.06	93.33 ± 0.06	99.94 ± 0.00
	U-Net++	94.65 ± 0.03	8.67 ± 19.44	90.01 ± 0.05	**99.98 ± 0.00**	94.16 ± 0.03	95.31 ± 0.05	**99.96 ± 0.00**
	ResUNet	94.00 ± 0.02	9.81 ± 16.20	88.76 ± 0.04	**99.98 ± 0.00**	93.41 ± 0.03	94.74 ± 0.04	99.95 ± 0.00
	Attention U-Net	94.03 ± 0.04	10.54 ± 19.13	88.92 ± 0.06	**99.98 ± 0.00**	94.20 ± 0.03	94.10 ± 0.06	99.95 ± 0.00
	DAUNet	94.75 ± 0.03	8.38 ± 18.31	90.16 ± 0.05	**99.98 ± 0.00**	**95.09 ± 0.03**	94.84 ± 0.04	**99.96 ± 0.00**
	R2B	**95.45 ± 0.02**	**3.62 ± 6.13**	**91.36 ± 0.03**	**99.98 ± 0.00**	**95.68 ± 0.02**	95.31 ± 0.03	**99.96 ± 0.00**
Transformer/Hybrid	TransUNet	94.06 ± 0.02	**4.12 ± 5.27**	88.88 ± 0.04	**99.98 ± 0.00**	92.15 ± 0.04	**96.23 ± 0.03**	99.95 ± 0.00
	TransUNet*	94.34 ± 0.02	7.55 ± 24.12	89.36 ± 0.04	**99.98 ± 0.00**	92.82 ± 0.04	**96.06 ± 0.03**	99.95 ± 0.00
	ViTAE V2	92.82 ± 0.03	7.76 ± 13.17	86.75 ± 0.05	99.97 ± 0.00	93.18 ± 0.04	92.65 ± 0.04	99.94 ± 0.00
	UTNet V2	94.61 ± 0.02	12.74 ± 35.53	89.87 ± 0.04	**99.98 ± 0.00**	94.20 ± 0.04	95.22 ± 0.04	99.95 ± 0.00
	MISSFormer	93.76 ± 0.04	8.46 ± 14.60	88.48 ± 0.06	**99.98 ± 0.00**	93.84 ± 0.03	93.84 ± 0.05	99.95 ± 0.00
	PVT-CASCADE*	91.29 ± 0.04	7.73 ± 12.10	84.19 ± 0.06	99.97 ± 0.00	89.81 ± 0.06	93.16 ± 0.04	99.93 ± 0.00
	ESFPNet-S	88.17 ± 0.08	16.20 ± 25.73	79.70 ± 0.12	99.96 ± 0.00	88.23 ± 0.10	88.78 ± 0.08	99.91 ± 0.00
	ESFPNet-S*	94.92 ± 0.02	6.99 ± 26.46	90.36 ± 0.03	**99.98 ± 0.00**	94.32 ± 0.02	95.59 ± 0.03	**99.96 ± 0.00**
	HiFormer-B	92.52 ± 0.04	9.82 ± 20.26	86.28 ± 0.06	**99.98 ± 0.00**	91.72 ± 0.04	93.52 ± 0.05	99.94 ± 0.00
	HiFormer-B*	94.25 ± 0.03	4.65 ± 8.80	89.24 ± 0.05	**99.98 ± 0.00**	93.62 ± 0.03	95.00 ± 0.04	99.95 ± 0.00
	PyramidalConv	**95.47 ± 0.02**	4.26 ± 6.41	**91.44 ± 0.04**	**99.98 ± 0.00**	**95.38 ± 0.03**	95.66 ± 0.03	**99.96 ± 0.00**
	PFF-Net	**95.33 ± 0.02**	**3.35 ± 5.19**	**91.13 ± 0.03**	**99.98 ± 0.00**	94.69 ± 0.02	**96.06 ± 0.03**	**99.96 ± 0.00**

* indicates that the model utilized pre-trained weights on the ImageNet21K dataset or pre-trained weights published in relevant literature. The bold values represent the best or second-best performance, facilitating the comparison of the methods corresponding to the top-performing results.

10.3389/fcomp.2025.1677905

images. Nevertheless, pretrained initialization remains beneficial, as seen in ESFPNet-S* and TransUNet*, which clearly outperform their non-pretrained counterparts, suggesting that pretraining still accelerates convergence and improves stability even under modality mismatch.

Third, our PFF-Net achieves highly competitive results, with Dice (95.33%), Jaccard (91.13%), and HD95 (3.35) all ranking among the top. Notably, its HD95 is the lowest across all models, even surpassing R2B and PyramidalConv, highlighting the robustness of our boundary-aware fusion mechanism in handling challenging CT slices where spleen boundaries are faint or ambiguous. The balance between Precision (96.06%) and Sensitivity (94.69%) further indicates that PFF-Net reduces both false positives and false negatives, offering reliable and consistent segmentation across diverse cases.

Finally, while specialized CNN designs such as R2B and PyramidalConv achieve marginally higher Dice or Specificity, they do not exhibit the same all-around superiority as PFF-Net. In contrast, our method consistently delivers near state-of-the-art Dice and Jaccard scores while leading in HD95, establishing itself as a robust and generalizable framework for CT-based organ segmentation. This suggests that the proposed pyramid feature fusion and boundary-aware branch provide strong adaptability across modalities, even when domain gaps with natural image pretraining exist.

4.3.4 Visualization results

4.3.4.1 Visual comparison of predicted results

Figure 6 presents a comparative visualization of the predicted segmentation results across three datasets: spleen segmentation, polyp segmentation, and skin lesion segmentation. Each dataset includes two representative cases to illustrate the performance of different models.

The first two columns correspond to spleen segmentation, the middle two columns display images from the ISIC 2018 skin lesion dataset, and the last two columns showcase polyp segmentation results. Each row represents the predictions of a specific model, with the bottom row highlighting the results obtained using our proposed method, while the remaining rows display the outputs of alternative approaches.

To facilitate comparison, expert-annotated contours are outlined in red, whereas the segmentation results predicted by different methods are delineated in green. These visualizations provide clear insights into the boundary adherence and segmentation accuracy of each approach.

As shown in Table 4, PyramidalConv and R2B achieve superior performance compared to other methods, including our proposed approach. However, it is worth noting that our method consistently ranks just below PyramidalConv and R2B across most evaluation metrics, highlighting its strong competitiveness. This observation is further corroborated by the visualization results, which provide intuitive insights into model performance by illustrating how each approach segments targets in medical images and enabling direct comparison of segmentation quality.

Notably, in the cases of skin lesion and polyp segmentation, our model produces results that closely resemble the ground truth masks in terms of target localization, region size, and contour accuracy. This strong alignment with the ground truth underscores the generalization capability and competitive performance of our model in medical image segmentation tasks.

4.3.4.2 Visualization of boundary-aware features

The visualization results of boundary-aware features are presented in Figure 7. To illustrate the impact of boundary perception across different data modalities, we selected two representative cases from each dataset.

The first two columns showcase examples from skin lesion segmentation tasks, where the relatively uniform background enables rough localization of boundaries. Our method effectively enhances boundary contrast, aiding in precise delineation.

The middle two columns correspond to spleen segmentation on CT images, where the spleen is relatively small and absent in most slices. In this scenario, the boundary-aware decoder efficiently localizes segmentation areas, demonstrating its adaptability to challenging anatomical structures.

The last two columns display polyp segmentation cases, where polyps share similar textures with surrounding tissues. In this context, our method not only highlights polyp boundaries but also captures relevant contextual information, improving overall segmentation accuracy.

Compared to existing methods, our approach demonstrates superior boundary adherence in lesion segmentation, which is particularly beneficial for clinical applications requiring precise delineation. This improvement is driven by our novel feature fusion strategy, which enhances spatial consistency and effectively suppresses segmentation artifacts. By integrating multi-scale contextual cues, our model ensures sharper and more reliable boundary predictions, reinforcing its robustness across diverse medical imaging tasks.

These visualization results provide an intuitive understanding of boundary-aware features, revealing how boundary perception varies across different tasks and data modalities. This underscores the adaptability and diversity of model performance in various scenarios.

4.3.5 Comparison of model parameters and computational complexity

Table 5 presents the number of parameters and computational complexity of each model. All models have an input dimension of $1 \times 3 \times 224 \times 224$ and an output dimension of $1 \times 2 \times 224 \times 224$. Additionally, the performance results shown in Tables 2–4 are taken into account.

Compared to other methods, the proposed model achieves a balance between model parameters, computational complexity, and performance. It maintains a relatively low number of parameters and computational cost while delivering excellent segmentation results. This highlights the effectiveness and efficiency of the proposed method in the field of medical image segmentation.

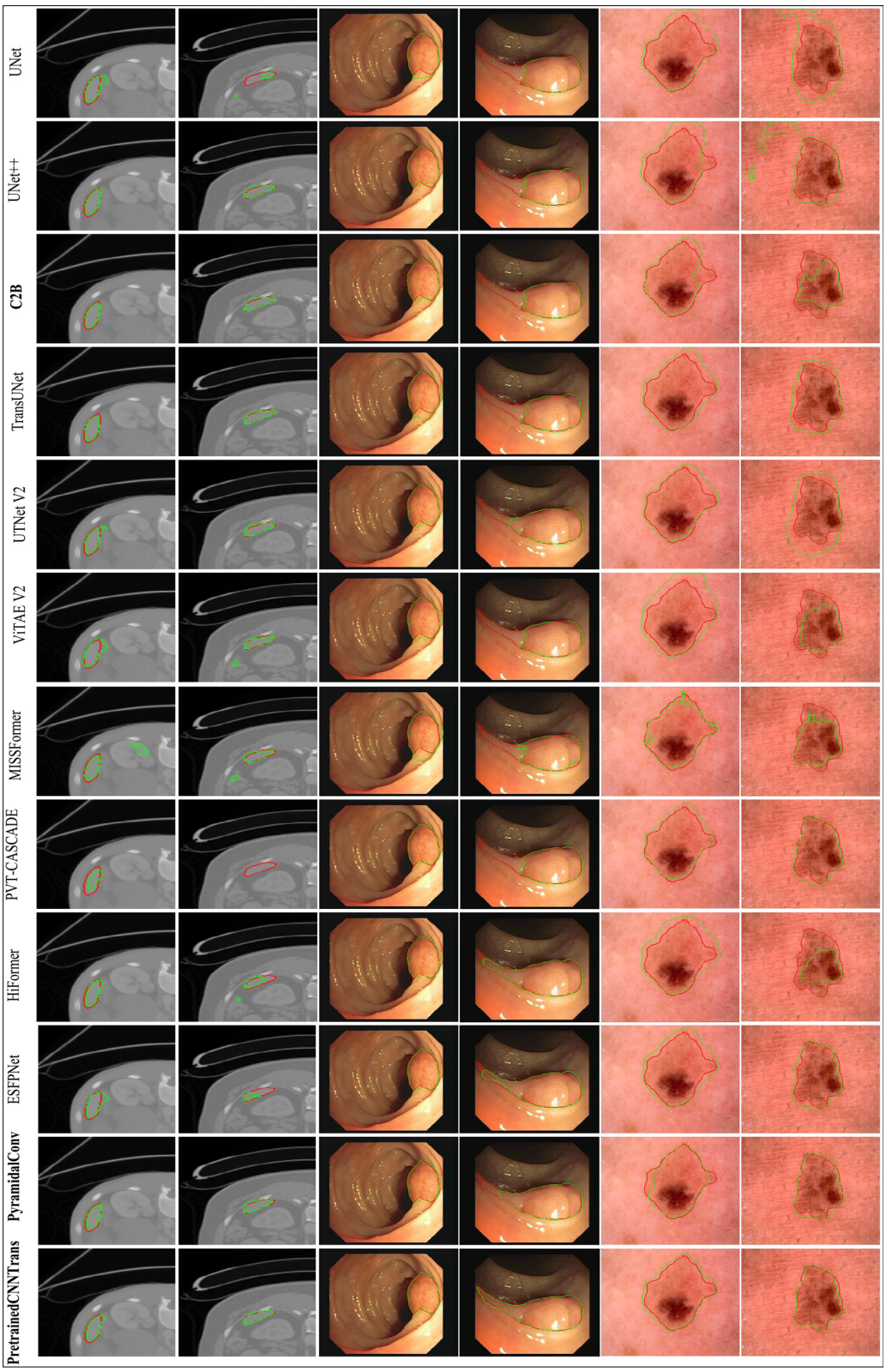

FIGURE 6
Visualization examples.

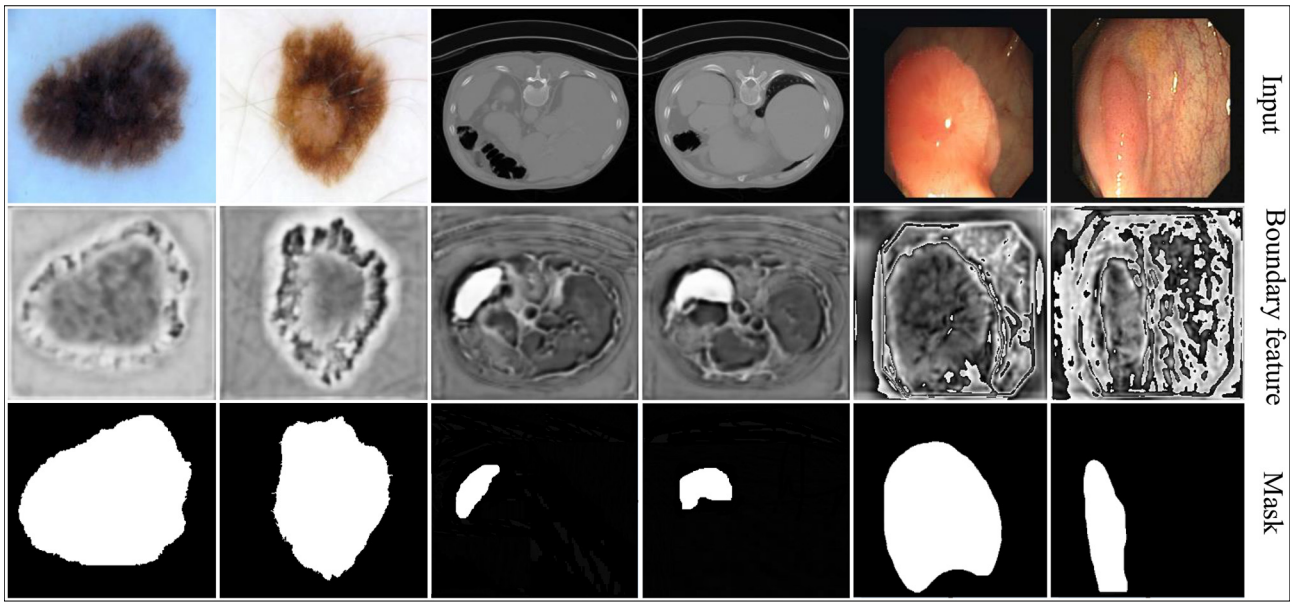

FIGURE 7
Visualization examples of boundary features.

4.4 Ablation experiments

The ablation experiments on the polyp dataset (Table 6) provide deeper insights into the role of different components in our architecture. Several key observations can be drawn.

4.4.1 Impact of each component

First, comparing the first two rows, the Mix Transformer encoder (90.32% Dice) clearly outperforms the ResNet34 encoder (89.96% Dice), confirming the advantage of Transformer-based architectures in modeling long-range dependencies. However, when both encoders are fused (third row), performance improves further (91.50% Dice), demonstrating the complementarity between CNNs (local spatial features) and Transformers (global contextual modeling). This validates our motivation for designing a dual-backbone architecture.

Second, the introduction of the Boundary-Aware Decoder brings additional benefits. In the dual-task setup (fourth row, variant A), explicit boundary supervision enables the network to refine edge details, leading to a slight improvement in Dice and HD95. This highlights the importance of boundary modeling in medical image segmentation, where accurate delineation of lesion contours is clinically critical.

Third, the comparison between variant A and variant B reveals that merely injecting boundary information into the Region-Aware Decoder, without direct supervision of the Boundary-Aware Decoder, results in a slight drop in performance. This suggests that boundary features need explicit supervision to be fully effective, otherwise their contribution becomes diluted in the feature fusion process.

Finally, the optimized design (variant C), where dual supervision is applied to both decoders achieves the best balance, with Dice maintained at 91.61% and HD95 reduced to 12.22, the lowest among all settings. This demonstrates that our Pyramid Fusion + dual-supervision strategy not only maximizes average segmentation accuracy but also significantly improves robustness against boundary ambiguity.

Overall, the ablation results confirm three key principles: (1) hybrid CNN–Transformer backbones provide complementary feature representations, (2) explicit boundary modeling is essential for accurate lesion delineation, and (3) dual-task supervision ensures the effective utilization of boundary information. These insights validate the architectural choices in PFF-Net and explain its consistent superiority in the main experiments.

4.4.2 Impact of multi-scale feature fusion strategy extracted by CNN and transformer backbones

Pre-trained CNN and Transformer backbones provide multiple strategies for multi-scale feature fusion. In this study, we focus on same-level fusion methods, including two basic strategies: element-wise addition (*Addition Fusion*) and channel-wise concatenation (*Concatenation Fusion*). Beyond these simple baselines, we also compare several representative fusion modules proposed in recent literature:

- **STCF (Spatial-Transformer-CNN Fusion)** from CoTrFuse (Chen et al., 2023), which explicitly fuses spatial and contextual features from CNNs and Transformers.

TABLE 5 Parameters and FLOPs of all models.

Metric	UNet	UNet++	R2B	TransUNet	SwinUNet	UTNet V2	MISSFormer	ESFPNet-S	PVT-CASCADE	HiFormer-B	Ours
Param(M)	7.85	9.16	44.06	93.19	41.30	12.79	35.37	25.02	35.25	34.12	24.40
FLOPs(G)	10.70	26.67	29.02	24.64	8.64	15.26	7.18	3.39	6.23	35.37	17.46

- **FCM (Feature Complementary Module)** from CTC-Net (Yuan et al., 2023), designed to enhance feature complementarity via cross-branch interactions.
- **APF (Addition and Product Fusion)**, an enhanced variant of addition fusion that incorporates both element-wise addition and multiplication before channel alignment.
- **CFM (Complementary Fusion Module)**, our proposed lightweight module (Figure 7), which achieves efficient yet effective feature integration with lower computational overhead.

This setup ensures a fair and comprehensive comparison, covering both classical fusion schemes and advanced modules. The ablation results are summarized in Table 7.

First, among the two simplest baselines, channel-wise concatenation (91.80% Dice, HD95 = 12.14) consistently outperforms element-wise addition (91.61% Dice, HD95 = 12.22). This suggests that concatenation preserves richer complementary representations from CNN and Transformer backbones, whereas addition may oversimplify the integration.

Second, more sophisticated fusion modules from prior works, such as APF, STCF, and FCM, generally improve Dice scores but at the cost of increased complexity. For example, STCF incurs higher FLOPs and only marginally improves Dice (91.66%) while degrading HD95 (13.47). Similarly, APF achieves moderate gains over addition but still underperforms concatenation in terms of both Dice and HD95. Notably, FCM reaches the same Dice score as our CFM (91.87%) but requires almost twice the parameters (121M vs. 68M) and produces inferior boundary precision (HD95 = 12.40). These comparisons highlight that not all fusion strategies translate into better accuracy–efficiency trade-offs.

Finally, our proposed CFM achieves the best overall balance, obtaining both the highest Dice score (91.87%) and the lowest HD95 (11.68), while maintaining computational efficiency (68M parameters). This indicates that the lightweight design of CFM enables effective feature complementarity between CNNs and Transformers without introducing unnecessary overhead. In particular, the superior HD95 results confirm that CFM enhances boundary delineation robustness, which is critical for precise lesion segmentation.

In summary, the comparative analysis demonstrates that while traditional concatenation already provides a solid improvement over addition, our CFM further advances segmentation performance by delivering both accuracy and efficiency. This validates the necessity of a tailored lightweight fusion design for hybrid CNN–Transformer segmentation networks.

5 Discussion

We discuss the impact of pretraining and the practical significance of our hybrid architecture across different modality as observed from Tables 2–7. PFF-Net fuses local feature extraction from pretrained CNNs and the global context modeling ability of Transformers, leading to two key findings supported by our experiments.

TABLE 6 Impact of each component.

Backbone/encoder	Boundary-aware decoder	Params(M)	FLOPs(G)	AVG	
				Dice%↑	HD95↓
Resnet34*	×	43.03	12.45	89.96 ± 0.15	16.20 ± 46.44
Mix Transformer*	×	24.49	3.22	90.32 ± 0.15	15.47 ± 42.57
Resnet34* + mix transformer*	×	67.20	15.58	91.50 ± 0.12	13.00 ± 36.04
Resnet34* + mix transformer*	√ + A	67.42	17.50	**91.61 ± 0.13**	12.88 ± 38.14
Resnet34* + mix transformer*	√ + B	67.42	17.50	91.51 ± 0.13	13.53 ± 39.87
Resnet34* + mix transformer*	√ + C	67.42	17.50	**91.61 ± 0.13**	**12.22 ± 32.51**

1. represents performing only dual-tasking, where the Boundary-aware Decoder and the Region-aware Decoder branches are independent, and supervision is applied at both decoder ends.
2. represents having dual decoders, with the boundary-aware results applied to the Region-aware Decoder branch, but no supervision at the Boundary-aware Decoder.
3. represents implementing dual supervision in situation B, i.e., supervision is also applied at the end of the Boundary-aware Decoder.
4. * indicates that the MixTransformer backbone and ResNet backbone are initialized with pre-trained weights from the ImageNet dataset. √ denotes the existence of the Boundary-aware Decoder, and × indicates the absence of the Boundary-aware Decoder.
5. For simplicity, the fusion of ResNet34 and Mix Transformer uses the simplest addition strategy.
The bold values represent the best or second-best performance, facilitating the comparison of the methods corresponding to the top-performing results.

TABLE 7 Impact of different feature fusion strategy.

Fusion strategy	Params(M)	FLOPs(G)	AVG	
			Dice% ↑	HD95 ↓
Addition Fusion	67.42	17.50	91.61 ± 0.13	12.22 ± 32.51
Concatenation Fusion	68.19	17.61	91.80 ± 0.12	12.14 ± 32.48
APF	68.19	17.61	91.68 ± 0.13	12.79 ± 36.92
STCF	74.57	18.58	91.66 ± 0.13	13.47 ± 41.00
FCM	121.04	25.65	**91.87 ± 0.13**	12.40 ± 36.53
CFM	68.21	17.62	**91.87 ± 0.12**	**11.68 ± 27.07**

The decimal point of the Dice average score has been shifted two places to the right. The optimal metric for each column has been highlighted in bold.

5.1 Role of pretraining and modality-dependent gains

In our experiments, pretraining mostly speeded up convergence and boosted segmentation scores immensely w.r.t Transformer-only models. SwinUNet without pretraining, for example, showed poor performance on any of the three datasets (thus not given in full table). But the scale of these benefits is tightly connected to how well the pretraining data matches up with the target modality. Tables 2–7 show HiFormer and ESFPNet both provided **+8.85%**, **+4.33%**, and **9.00%**, improvements for Polyp segmentation and ISIC2018 in terms of Dice scores, separately, those two tasks are more similar to the domain distribution our original task. By contrast, HiFormer only improved the Dice value of Spleen CT segmentation by **+1.87%** (Table 4). ESFPNet still accumulated a significant improvement of **+7. 76%**, indicating that the model's architecture allows effective cross-domain feature reuse even in modalities with low similarity (Mei et al., 2022). In addition, since the effectiveness of model pretraining is modality dependent, more specific approaches such as large-scale medical imaging pretraining with unsupervised or self-supervised learning should also be explored (Xu et al., 2025).

5.2 Practical value of PFF-Net

PFF-Net integrated pretrained CNN and Transformer backbones into a unified multi-scale pyramid fusion architecture with a boundary-aware refinement branch. Notably, the boundary-aware branch helped to better capture fine structural details and consequently improved the Dice and HD95 scores across several datasets.

6 Conclusion

The focus of this study is to propose a hybrid CNN–Transformer architecture, called **PFF-Net**, by combining pyramid feature fusion with boundary-aware refinement for medical image segmentation. Experiments on polyp, skin lesion, and splenic CT datasets showed comparable improvements suggesting that the architecture preserves local information well while modeling global context. In future work, we plan to investigate (1) modality-aware pretraining methods using large-scale medical image datasets, particularly in unsupervised/self-supervised learning settings to further learn domain-specific features as

well; (2) more lightweight Transformer formulations with inherently inductive biases for better generalization on small-size datasets; and (3) adaptive fusion strategies for processing 3D volumetric scans and fusing multi-modality sources in clinical contexts.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

XL: Investigation, Writing – original draft, Writing – review & editing, Funding acquisition. JT: Funding acquisition, Writing – review & editing. SH: Writing – review & editing, Software, Investigation, Methodology, Project administration. WS: Investigation, Writing – review & editing, Visualization, Resources.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the Excellent Youth Fund of the Hunan Provincial Education Department (Grant No. 24B0680) and the Hunan Province General University Teaching Reform Research Project (Grant No: HNJG-20230957).

Acknowledgments

We sincerely appreciate the reviewers for their diligent efforts and invaluable feedback, which have significantly contributed to improving the quality of this manuscript. Their constructive suggestions have been instrumental in refining our work. We also extend our gratitude to the Hunan Malanshan Computing Media Research Institute for providing essential GPU resources, which were pivotal in enabling the successful execution of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Azad, R., Jia, Y., Aghdam, E. K., Cohen-Adad, J., and Merhof, D. (2023). Enhancing medical image segmentation with transception: A multi-scale feature fusion approach. *arXiv preprint arXiv:2301.10847*. doi: 10.48550/arXiv.2301.10847

Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., and Vilariño, F. (2015). Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* 43, 99–111. doi: 10.1016/j.compmedimag.2015.02.007

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*. doi: 10.48550/arXiv.2105.05537

Chang, Q., Ahmad, D., Toth, J., Bascom, R., and Higgins, W. E. (2023). "Esfpnet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video," in *Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging, volume 12468* (San Diego, CA: SPIE), 1246803. doi: 10.1117/12.2647897

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., et al. (2024). Transunet: rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* 97:103280. doi: 10.1016/j.media.2024.103280

Chen, Y., Wang, T., Tang, H., Zhao, L., Zhang, X., Tan, T., et al. (2023). Cotrfuse: a novel framework by fusing cnn and transformer for medical image segmentation. *Phys. Med. Biol.* 68:175027. doi: 10.1088/1361-6560/acede8

Ding, X., Zhang, X., Han, J., and Ding, G. (2022). "Scaling up your kernels to 31 × 31: Revisiting large kernel design in CNNS," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway: IEEE), 11963–11975. doi: 10.1109/CVPR52688.2022.01166

Dong, B., Wang, W., Fan, D., Li, J., Fu, H., and Shao, L. (2021). Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*. doi: 10.48550/arXiv.2108.06932

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). "Dual attention network for scene segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway: IEEE), 3146–3154. doi: 10.1109/CVPR.2019.00326

Gao, Y., Zhang, S., Shi, L., Zhao, G., and Shi, Y. (2025). Collaborative transformer u-shaped network for medical image segmentation. *Appl. Soft Comput.* 173:112841. doi: 10.1016/j.asoc.2025.112841

Gao, Y., Zhou, M., Liu, D., and Metaxas, D. N. (2022). A multi-scale transformer for medical image segmentation: architectures, model efficiency, and benchmarks. *CoRR,abs*/2203.00131. doi: 10.48550/arXiv.2203.00131

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., et al. (2022). "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Piscataway: IEEE), 574–584. doi: 10.1109/WACV51458.2022.00181

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E. K., Cohen-Adad, J., et al. (2023). "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Piscataway: IEEE), 6202–6212. doi: 10.1109/WACV56688.2023.00614

Huang, X., Deng, Z., Li, D., Yuan, X., and Fu, Y. (2023). Missformer: an effective transformer for 2D medical image segmentation. *IEEE Trans. Med. Imaging* 42, 1484–1494. doi: 10.1109/TMI.2022.3230943

Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., et al. (2020). "KVASIR-SEG: a segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II 26* (Springer: New York), 451–462. doi: 10.1007/978-3-030-37734-2_37

Kanopoulos, N., Vasanthavada, N., and Baker, R. L. (1988). Design of an image edge detection filter using the sobel operator. *IEEE J. Solid-State Circuits* 23, 358–367. doi: 10.1109/4.996

Kumar, P. A., and Gunasundari, R. (2025). A lightweight adaptive spatial channel attention efficient net B3 based generative adversarial network approach for mr image reconstruction from under sampled data. *Magn. Reson. Imaging* 117:110281. doi: 10.1016/j.mri.2024.110281

Kumar, S. (2025). Advancements in medical image segmentation: a review of transformer models. *Comput. Electr. Eng.* 123:110099. doi: 10.1016/j.compeleceng.2025.110099

Lei, W., Xu, W., Li, K., Zhang, X., and Zhang, S. (2025). Medlsam: localize and segment anything model for 3D CT images. *Med. Image Anal.* 99:103370. doi: 10.1016/j.media.2024.103370

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., and Zhang, D. (2022). Ds-transunet: dual swin transformer u-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* 71, 1–15. doi: 10.1109/TIM.2022.3178991

Liu, X., Hu, Y., and Chen, J. (2023). Hybrid cnn-transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron. *Biomed. Signal Process. Control* 86:105331. doi: 10.1016/j.bspc.2023.105331

Liu, X., Yang, L., Chen, J., Yu, S., and Li, K. (2022). Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation. *Biomed. Signal Process. Control* 71:103165. doi: 10.1016/j.bspc.2021.103165

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision* (Piscataway: IEEE), 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Lou, M., and Yu, Y. (2025). "Overlock: an overview-first-look-closely-next convnet with context-mixing dynamic kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Piscataway: IEEE), 128–138. doi: 10.1109/CVPR52734.2025.00021

Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., et al. (2022). Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiol. Artif. Intell.* 4:e210315. doi: 10.1148/ryai.210315

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M. C. H., Heinrich, M. P., Misawa, K., et al. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. doi: 10.48550/arXiv.1804.03999

Park, K.-B., and Lee, J. Y. (2022). Swine-net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and swin transformer. *J. Comput. Des. Eng.* 9, 616–632. doi: 10.1093/jcde/qwac018

Qayoom, A., Xie, J., and Ali, H. (2025). Polyp segmentation in medical imaging: challenges, approaches and future directions. *Artif. Intell. Rev.* 58:169. doi: 10.1007/s10462-025-11173-2

Rahman, M. M., and Marculescu, R. (2023). "Medical image segmentation via cascaded attention decoding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Piscataway: IEEE), 6222–6231. doi: 10.1109/WACV56688.2023.00616

Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., and Kumar, K. (2024). Self-supervised learning for medical image analysis: a comprehensive review. *Evol. Syst.* 15, 1607–1633. doi: 10.1007/s12530-024-09581-w

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer: New York), 234–241. doi: 10.1007/978-3-319-24574-4_28

Shah, O. I., Rizvi, D. R., and Mir, A. N. (2025). Transformer-based innovations in medical image segmentation: a mini review. *SN Comput. Sci.* 6:375. doi: 10.1007/s42979-025-03929-y

Silva, J., Histace, A., Romain, O., Dray, X., and Granado, B. (2014). Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* 9, 283–293. doi: 10.1007/s11548-013-0926-3

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). "Segmenter: transformer for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision* (Piscataway: IEEE), 7262–7272. doi: 10.1109/ICCV48922.2021.00717

Tajbakhsh, N., Gurudu, S. R., and Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* 35, 630–644. doi: 10.1109/TMI.2015.2487997

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 6000–6010. doi: 10.5555/3295222.3295349

Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., et al. (2017). A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* 2017:4037190. doi: 10.1155/2017/4037190

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. (2021a). "Transbts: multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer: New York), 109–119. doi: 10.1007/978-3-030-87193-2_11

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., et al. (2021b). "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *IEEE/CVF International Conference on Computer Vision* (Piscataway: IEEE), 568–578. doi: 10.1109/ICCV48922.2021.00061

Xia, Q., Zheng, H., Zou, H., Luo, D., Tang, H., Li, L., et al. (2025). A comprehensive review of deep learning for medical image segmentation. *Neurocomputing* 613:128740. doi: 10.1016/j.neucom.2024.128740

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021a). Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* 34, 12077–12090. doi: 10.5555/3540261.3541185

Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021b). "COTR: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part III 24* (Springer: New York), 171–180. doi: 10.1007/978-3-030-87199-4_16

Xu, J., Yang, W., Kong, L., Liu, Y., Zhou, Q., Zhang, R., et al. (2025). Visual foundation models boost cross-modal unsupervised domain adaptation for 3D semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* 1–15. doi: 10.1109/TITS.2025.3587430

Yan, S., Yang, B., Chen, A., Zhao, X., and Zhang, S. (2025). Multi-scale convolutional attention frequency-enhanced transformer network for medical image segmentation. *Inf. Fusion* 119:103019. doi: 10.1016/j.inffus.2025.103019

Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., and Xie, Y. (2024). From CNN to transformer: a review of medical image segmentation models. *J. Imaging Inform. Med.* 37, 1–19. doi: 10.1007/s10278-024-00981-7

Yuan, F., Zhang, Z., and Fang, Z. (2023). An effective CNN and transformer complementary network for medical image segmentation. *Pattern Recognit.* 136:109228. doi: 10.1016/j.patcog.2022.109228

Zhang, J., Chen, X., Yang, B., Guan, Q., Chen, Q., Chen, J., et al. (2025). Advances in attention mechanisms for medical image segmentation. *Comput. Sci. Rev.* 56:100721. doi: 10.1016/j.cosrev.2024.100721

Zhang, Q., Xu, Y., Zhang, J., and Tao, D. (2023). Vitae V2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *Int. J. Comput. Vis.* 131, 1–22. doi: 10.1007/s11263-022-01739-w

Zhang, Y., Liu, H., and Hu, Q. (2021). "Transfuse: fusing transformers and CNNS for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24* (Springer: New York), 14–24. doi: 10.1007/978-3-030-87193-2_2

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway: IEEE), 6881–6890. doi: 10.1109/CVPR46437.2021.00681

Zhou, Z., and Abawajy, J. (2025). Reinforcement learning-based edge server placement in the intelligent internet of vehicles environment. *IEEE Trans. Intell. Transp. Syst.* 1–11. doi: 10.1109/TITS.2025.3557259

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "U-net++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer: New York), 3–11. doi: 10.1007/978-3-030-00889-5_1

Zhou, Z., Shojafar, M., Abawajy, J., Yin, H., and Lu, H. (2022a). ECMS: an edge intelligent energy efficient model in mobile edge computing. *IEEE Trans. Green Commun. Netw.* 6, 238–247. doi: 10.1109/TGCN.2021.3121961

Zhou, Z., Shojafar, M., Alazab, M., and Li, F. (2022b). IECL: an intelligent energy consumption model for cloud manufacturing. *IEEE Trans. Ind. Informat.* 18, 8967–8976. doi: 10.1109/TII.2022.3165085