

Frequency Domain Unlocks New Perspectives for Abdominal Medical Image Segmentation

Kai Han, Siqi Ma, Chengxuan Qian, Jun Chen, Chongwen Lyu, Yuqing Song, Zhe Liu

Abstract—Accurate segmentation of tumors and adjacent normal tissues in medical images is essential for surgical planning and tumor staging. Although foundation models generally perform well in segmentation tasks, they often struggle to focus on foreground areas in complex, low-contrast backgrounds, where some malignant tumors closely resemble normal organs, complicating contextual differentiation. To address these challenges, we propose the Foreground-Aware Spectrum Segmentation (FASS) framework. First, we introduce a foreground-aware module to amplify the distinction between background and the entire volume space, allowing the model to concentrate more effectively on target areas. Next, a feature-level frequency enhancement module, based on wavelet transform, extracts discriminative high-frequency features to enhance boundary recognition and detail perception. Eventually, we introduce an edge constraint module to preserve geometric continuity in segmentation boundaries. Extensive experiments on multiple medical datasets demonstrate superior performance across all metrics, validating the effectiveness of our framework, particularly in robustness under complex conditions and fine structure recognition. Our framework significantly enhances segmentation of low-contrast images, paving the way for applications in more diverse and complex medical imaging scenarios.

Index Terms—Medical image segmentation, low-contrast images, frequency enhancement, edge constrain.

I. INTRODUCTION

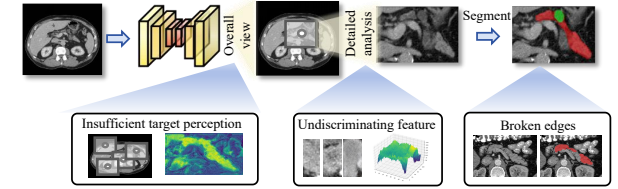
THE identification of abdominal tumors plays a pivotal role in early cancer detection, treatment planning, and improving patient survival [1]–[4]. Accurate segmentation of organs and tumors in CT images enables clinicians to assess organ conditions and precisely determine the size, location, and morphology of tumors, thereby facilitating more reliable disease evaluation and optimal therapeutic decision-making [5]–[7]. However, manual annotation is extremely time-consuming and labor-intensive, and it demands substantial clinical expertise, particularly when dealing with complex anatomical structures or ambiguous boundaries [8], [9]. Therefore, the development of efficient and accurate automated segmentation algorithms has become both essential and urgent.

This work was supported in part by the National Natural Science Foundation of China under Grant (62276116), Jiangsu Six Talent Peak Program under Grant (DZXX-122), Jiangsu Graduate Research Innovation Program under Grant (KY-CX23_3677) and National Undergraduate Training Program on Innovation and Entrepreneurship (Grant No. 202410299049Z). (Corresponding authors: Zhe Liu.)

Kai Han, Siqi Ma, Jun Chen, Chongwen Lyu, Yuqing Song and Zhe Liu are with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China (e-mail: 2112108003@stmail.ujs.edu.cn; 2212208031@stmail.ujs.edu.cn; chenjun@ujs.edu.cn; 2212308023@stmail.ujs.edu.cn; yqsong@ujs.edu.cn; 1000004088@ujs.edu.cn).

Chengxuan Qian is with the School of Mathematical Sciences, Jiangsu University, Zhenjiang 212013, China (e-mail: chengxuan.qian@stmail.ujs.edu.cn).

(a) Challenge: Automatic segmentation challenges for low contrast images



(b) Our Solution: Foreground-aware spectrum segmentation

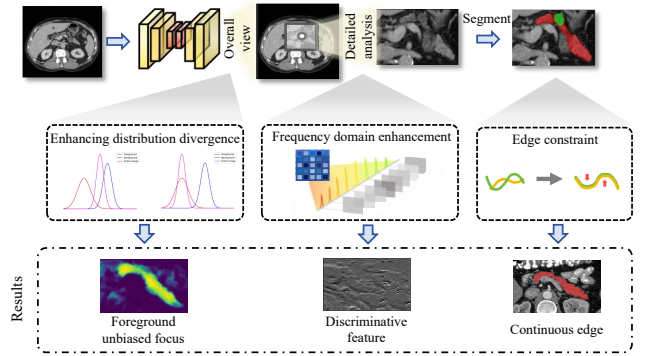


Fig. 1. Comparison between our FASS framework and previous automatic segmentation methods. (a) Segmentation process of low-contrast images by previous methods, which face challenges such as insufficient target perception, undiscriminating features, and broken edges. (b) Segmentation process of the FASS method. Our FASS framework employs adversarial training between the full image and background feature distribution to achieve focused attention on the foreground. Discriminative features are then enhanced in the frequency domain, and boundary integrity and continuity are strengthened through the edge constraint module.

These challenges are exacerbated in the segmentation of low-contrast abdominal images due to the intricate anatomical structures and limited contrast. In the abdominal region, multiple organs overlap, and tumors frequently adhere to or embed within organ surfaces, complicating foreground-background separation and leading to potential tissue misclassification. Additionally, the low contrast in these images results in blurred boundaries between tumors and surrounding tissues with similar grayscale values, making lesion contours challenging to discern. This difficulty highlights the need for advanced segmentation methods tailored to low-contrast environments.

Deep learning has shown promising potential in medical image segmentation, offering new approaches to address these complexities. Prior studies have employed two-stage strategies, progressing from manual region clipping [10] to model-driven autonomous learning [11]–[17], [17]–[19] for

initial target localization. More recently, single-stage methods have emerged, enabling end-to-end segmentation of the target region [20], [21]. In response to the challenges of low-contrast segmentation, approaches focused on boundary enhancement and multi-scale information fusion have been developed [22], [23], aiming to improve segmentation by extracting high-resolution features and refining boundary details in blurred areas. Despite these advancements, current models continue to face limitations in low-contrast and complex abdominal environments. When target regions exhibit complex internal topologies and similar boundary pixel intensities, models often struggle to capture subtle discriminative features, resulting in incomplete foreground segmentation and broken boundaries. These challenges increase the difficulty of achieving precise segmentation, as illustrated in Fig. 1 (a).

To this end, we propose the Foreground-Aware Spectrum Segmentation (FASS) framework to improve target localization and detail capture, as illustrated in Fig. 1 (b). Specifically, the Foreground-Aware (FA) module enhances foreground feature extraction and interpretation by employing an adversarial training strategy to maximize the distributional contrast between background and input image features. After identifying the target region, we designed a Feature-Level Frequency Enhancement (FLFE) module to extract discriminative features. This module performs spectral decomposition of the encoded output using wavelet transform and enhances the complementarity of high-frequency details through a cross-attention mechanism. Subsequently, it selects high-discriminative high-frequency features using channel and spatial attention mechanisms, improving the perception of the target boundaries and internal structures. Additionally, we introduce an Edge Constraint (EC) module to ensure edge integrity and geometric continuity in the segmentation results. To sum up, our main contributions are as follows:

- We propose an end-to-end Foreground-Aware Spectrum Segmentation (FASS) framework tailored for low-contrast medical image segmentation tasks.
- We design a Foreground-Aware (FA) module to deepen the model's understanding of foreground features by learning the heterogeneity between background and complete features, enabling focused attention on foreground regions.
- We introduce a Feature-Level Frequency Enhancement (FLFE) module based on wavelet transform, which selects discriminative high-frequency features to enhance detail capture.
- The Edge Constraint (EC) strategy is introduced to ensure boundary integrity and continuity, effectively preventing segmentation breaks in low-contrast settings.
- Extensive experiments demonstrate the independent performance benefits of each module within the FASS framework across multiple medical datasets, with overall performance significantly surpassing current state-of-the-art methods.

II. RELATED WORKS

A. Low-Contrast Medical Image Segmentation

Low-contrast medical image segmentation is a challenging task, especially when dealing with abdominal images from modalities like CT, MRI, or ultrasound. Automatic segmentation of such images is very difficult due to the small grayscale difference between the target and the background tissue. Traditional threshold segmentation methods or edge detection algorithms do not perform well in this context because they are highly dependent on sharp contrast differences. In recent years, deep learning has become a powerful tool in the field of medical image segmentation by learning complex features from images [24]–[27]. However, due to the subtle differences between the foreground and background, deep learning models can still encounter challenges in accurately distinguishing the boundaries of target regions in certain cases. To this end, researchers utilize generative adversarial networks (GANs) or enhancement techniques to generate clearer low-contrast images to assist in segmentation tasks [28], [29]. However, such methods often rely on the diversity of the original data, which may introduce artifacts or unrealistic features, affecting the reliability of the segmentation results. Besides, integrating information from other modalities/centers can compensate for the limitations of a single modality/center in low-contrast scenarios [30], [31], but acquiring such data is costly and requires substantial computational resources. There are also methods that attempt to enhance features or information within the network to improve the segmentation performance of low-contrast images [22], [23]. Despite some progress, low-contrast images typically exhibit minimal differences between the foreground and background, making it difficult for spatial domain enhancement to significantly improve these subtle distinctions. Our method utilizes frequency domain enhancement to amplify high-frequency components, enabling the model to better capture fine structures and accurately segment subtle features in complex scenes.

B. Region of Interest Location

The inherent complexity of medical images, particularly the similarity in texture, brightness, and morphology between background and foreground, poses a significant challenge to the localization of the region of interest. Traditional methods rely on manually segmenting the foreground region, which is effective but limited by the dependence on expert knowledge and the lack of automation [10], [32]–[37], making them unsuitable for large-scale image analysis. In recent years, two-stage segmentation methods have somewhat alleviated the problem of background interference, but they have introduced increased algorithmic complexity and the risk of error accumulation [11]–[13], [19]. Against this backdrop, single-stage region-of-interest segmentation models have emerged. For example, Jiang et al. [20] proposed the axial projection attention unit, which effectively filters out redundant feature information. Li et al. [21] introduced a balanced temperature loss function, significantly enhancing the model's focus on target regions. Besides, multi-task learning frameworks [16], [38]–[43] have been used to simultaneously predict multiple

related outputs, enhancing the understanding of foreground details. Attention mechanisms [44]–[46] have also been introduced, allowing the network to adjust its focus on different regions of the image, thus concentrating on key foreground structures. Despite these advances, these methods learn an unbiased mixture of foreground and background features during training, limiting the model’s ability to deeply explore their differences. By comparison, our method ensures that the model can pay biased attention to the region of interest during the inference stage, demonstrating excellent adaptability even in low-contrast environments by effectively resisting complex background interference.

C. Frequency-Based Image Analysis Techniques

Despite significant advances in image segmentation achieved through deep learning techniques, existing methods often rely on simulating human visual perception processes. These methods tend to integrate and process high-frequency (e.g., edges and textures) and low-frequency (e.g., shapes) information at the visual level. However, this practice may face limitations in low-contrast conditions due to difficulties in precisely distinguishing subtle discriminative features. Wavelet transform, as a tool with excellent spatial representation capabilities and directional sensitivity, can decompose image features into different frequency components, offering a method to fully utilize frequency information [47], [48].

In the field of medical image segmentation, high-frequency features extracted using wavelet transform have been proven to significantly enhance neural networks’ ability to learn high-frequency details [49]–[51]. By capturing details that are easily overlooked by human vision, networks can effectively address the challenges of segmentation under low-contrast conditions. Based on this, Jin et al. [52] explored frequency feature fusion techniques to improve models’ grasp of detailed textures and overall structures. Azad et al. [53], on the other hand, advocated for moderately suppressing high-frequency information to reduce excessive reliance on texture details, though this strategy may fall short in a low-contrast environment. Although the above strategies have achieved important breakthroughs in utilizing frequency information, indiscriminate use of high-frequency features may introduce noise, potentially threatening segmentation accuracy. In light of this, this paper proposes an effective strategy to enhance and selectively utilize discriminative high-frequency information to improve model performance under low-contrast conditions.

III. METHOD

In this section, we present our Foreground-Aware Spectrum Segmentation (FASS) framework, which consists of three key modules: the Foreground-Aware (FA) module, the Feature-Level Frequency Enhancement (FLFE) module, and the Edge Constraint (EC) module. The FA module employs adversarial training to maximize the feature distribution differences between the background and the input volume, guiding the model to focus more effectively on the foreground. The FLFE module selects more discriminative high-frequency features to enhance the ability to capture details. Furthermore, the EC module

refines the morphological contours of edge predictions to further enhance segmentation accuracy. The overall framework is illustrated in Fig. 2.

A. Foreground-Aware Module

To address the issue of the minimal difference between foreground and background in low-contrast images, which makes it difficult for the model to accurately identify the foreground, we introduce a Foreground-Aware (FA) module. Suppose an input image patch space $\{R_i \in I_i, i \in N\}$ with labels $\{Y_{R_i}, i \in N\}$ random cropped from corresponding image volume $I_i \in \mathbb{R}^{w' \times h' \times d'}$, where $w' \times h' \times d'$ represent the dimension of the volume, N is the sample number, and Y_i is the corresponding label. In this section, our goal is to randomly sample a background region $\mathcal{B}_i$ of size $w \times h \times d$ from R_i , as described in Eq. 1:

$$\begin{aligned} \mathcal{B}_i &= R_i[x : x + w, y : y + h, z : z + d], \\ x &\sim \text{Uniform}(0, w' - w), \\ y &\sim \text{Uniform}(0, h' - h), \\ z &\sim \text{Uniform}(0, d' - d). \end{aligned} \quad (1)$$

where x, y and z represent the random coordinate position of R_i . $\text{Uniform}(\cdot, \cdot)$ represents the randomness of the position point value.

In order to ensure that the sampled background region contains discrimination features from the foreground region, a parameter α is introduced as a control parameter for background region sampling to quantify the overlap degree between the background and the foreground region. Only if the overlap volume between is less than α , the background region $\mathcal{B}_i$ will be selected for training. The above process can be expressed as:

$$\bar{\mathcal{B}}_i = \mathcal{B}_i \in R_i : \text{Inter}(\mathcal{B}_i, \mathcal{F}_i) < \alpha \quad (2)$$

where $\mathcal{F}_i \in R_i$ is the foreground region and $\text{Inter}(\cdot, \cdot)$ represents the intersection calculation of two sets and is defined as follow:

$$\text{Inter}(\mathcal{B}_i, \mathcal{F}_i) = |\mathcal{B}_i \cap \mathcal{F}_i| \quad (3)$$

The selection of background region sampling size and hyper-parameter α will be discussed in detail in Sec. IV-F.

Then, the dual-path encoder architecture is adopted to capture the background region feature f_i^b and the global feature f_i , respectively. The encoder architecture consists of four layers of multi-scale convolutions. In order to further optimize the feature representation, a distribution divergence loss ℓ_{KL} based on KL divergence is introduced to minimize the distribution divergence ($\min |P(f_i|R_i, \theta) - P(f_i^b|\bar{\mathcal{B}}_i, \theta)|$). The distribution difference loss ℓ_{KL} can be expressed as:

$$\begin{aligned} \mathbf{D}(P(f_i|R_i, \theta) \| P(f_i^b|\bar{\mathcal{B}}_i, \theta)) \\ = P(f_i|R_i, \theta) \log \frac{P(f_i|R_i, \theta)}{P(f_i^b|\bar{\mathcal{B}}_i, \theta)} \end{aligned} \quad (4)$$

$$\ell_{KL} = e^{-\mathbf{D}(P(f_i|R_i, \theta) \| P(f_i^b|\bar{\mathcal{B}}_i, \theta))} \quad (5)$$

where θ represents the parameters of the encoder.

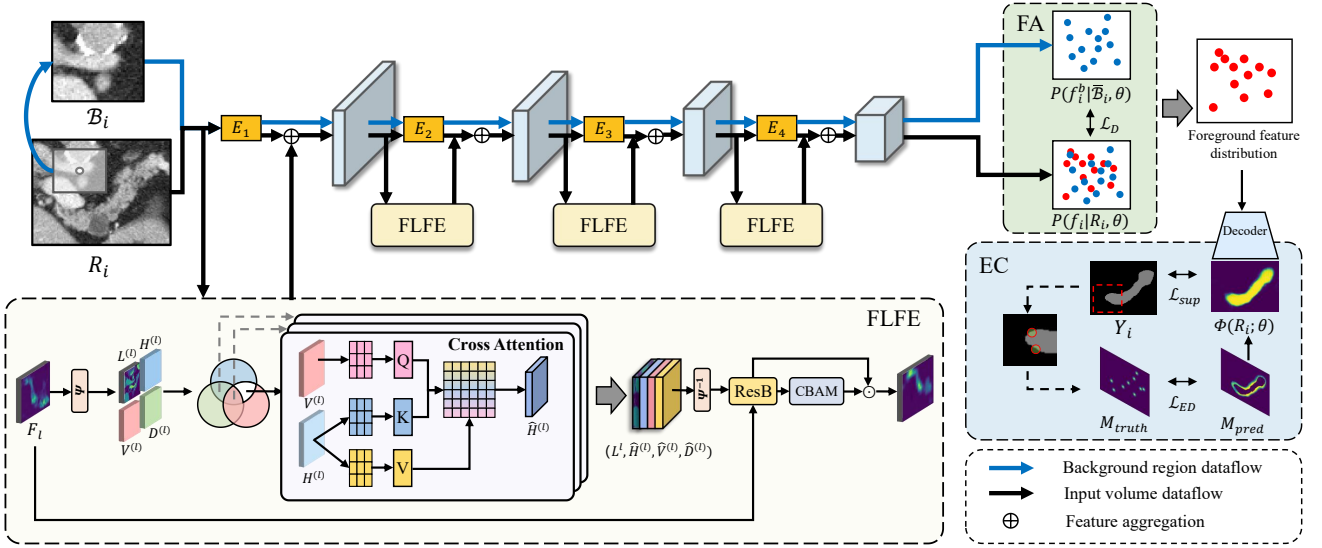


Fig. 2. Overview of the proposed FASS framework. The framework consists of a foreground-aware (FA) module (Sec. III-A), a feature-level frequency enhancement (FLFE) module (Sec. III-B), and an edge constraint (EC) module (Sec. III-C). Initially, sampled patches and background patches are fed into the encoder for feature extraction. The feature differences between encoder outputs are computed and maximized, with the FLFE module enhancing features during the encoding phase. Finally, the EC module refines the edges of the decoder output for optimized segmentation results.

In addition, the random cropping strategy of the input image I_i may lead to variable proportion of the foreground region $\mathcal{F}_i$ in the image patch R_i , which has a significant impact on the learning effect of this module. A higher proportion of foreground region is conducive to the model learning how to accurately focus on the foreground region, also meaning that the patch R_i is more valuable for training. In view of this, we design an adaptive weight factor ω to ensure that the model can learn from images with different foreground proportions in a balanced way by adjusting the loss contribution of each sampled patch R_i for better understanding the essential differences between foreground and background features, as shown in Eq. 6:

$$\omega = \text{Inter}(I_i^{\text{fore}}, \mathcal{F}_i) \quad (6)$$

where I_i^{fore} is the volume of the foreground region in global image volume I_i . The total loss function of this module can be expressed as follows:

$$\min_{\theta} \mathcal{L}_D = \omega \cdot \min_{\theta} \ell_{KL} \quad (7)$$

B. Feature-Level Frequency Enhancement Module

Suppose F_l denotes the feature map of encoder layer l . To enhance the ability to capture details using frequency information, the wavelet transform $\Psi(\cdot, \varepsilon_w)$ is used to map the spatial domain Y_i to the frequency domain $Y^{(i)}$ (e.g., $F_l \rightarrow F^{(l)}$), where ε_w represents different wavelet bases. Specifically, it applies the low-pass filter f_g and the high-pass filter f_h , and the subsequent down-sampling operation $\downarrow_2$ on $F^{(l)}$ to obtain a set of approximation coefficients ($L^{(l)}$) and detail coefficients ($H^{(l)}, V^{(l)}, D^{(l)}$), which correspond to the low frequency (overall structure) and high frequency (texture and

edge) information, respectively. The above process can be expressed as follows:

$$\begin{aligned} L^{(l)}, H^{(l)}, V^{(l)}, D^{(l)} = & \downarrow_2 (f_g * F_l), \downarrow_2 (f_h * F_l), \\ & \downarrow_2 (f_h * F_l^T), \downarrow_2 (f_h * f_h^T * F_l) \end{aligned} \quad (8)$$

where $*$, $\downarrow_2$, T denote convolution operation, down-sampling operation, and transpose operation, respectively.

To further enhance the detail richness in the high-frequency components, a cross-attention mechanism is introduced. This mechanism promotes the model to learn from each other and construct a more comprehensive feature representation. Here, taking the high-frequency components $H^{(l)}$ and $V^{(l)}$ as an example, the process can be expressed as:

$$\hat{H}^{(l)} = \sigma(W_H(H^{(l)} \cdot A_V V^{(l)})) \quad (9)$$

$$\hat{V}^{(l)} = \sigma(W_V(V^{(l)} \cdot A_H H^{(l)})) \quad (10)$$

where $\hat{H}$ and $\hat{V}$ represent the horizontal and vertical high-frequency components after cross-attention fusion, respectively. σ is the activation function, and W_H and W_V are the weight matrices used to adjust the feature fusion.

The attention weight matrix A_V and A_H are calculated based on the correlation between high-frequency components, which can be expressed as:

$$A_H = \text{softmax}\left(\frac{Q_{\hat{H}^{(l)}} K_{\hat{H}^{(l)}}^T}{\sqrt{d}}\right), A_V = \text{softmax}\left(\frac{Q_{\hat{V}^{(l)}} K_{\hat{V}^{(l)}}^T}{\sqrt{d}}\right) \quad (11)$$

where $Q_{H^{(l)}}$ and $K_{H^{(l)}}$ denote the query and key matrices of $H^{(l)}$, respectively. The softmax function is used to normalize the attention weights; d is the dimension of the key vector used to scale the inner product.

After the above steps, the enhanced high-frequency components ($\hat{H}^{(l)}, \hat{V}^{(l)}, \hat{D}^{(l)}$) are obtained. In order to reconstruct back to the spatial domain, the inverse wavelet transform is

applied to obtain the enhanced feature map. The process is shown in Eq. 12:

$$F'_l = \Psi^{-1}(L^l, \bar{H}^{(l)}, \bar{V}^{(l)}, \bar{D}^{(l)}) \quad (12)$$

where $\Psi^{-1}(\cdot, \varepsilon_w)$ denotes the inverse wavelet transform operation. Subsequently, in order to preserve long-range dependencies, residual block $Res(\cdot)$ with batch normalization is used. We further introduce a $CBAM(\cdot)$ module [54] to ensure the model focuses on high-frequency information that is critical to the task. The generated attention map P_l can be expressed as:

$$P_l = CBAM(Res(F'_l)) \quad (13)$$

In view of the structural characteristics of the U-shaped network, its shallow layers tend to capture high-frequency detailed information, while the deep layers focus more on low-frequency global semantic features. Therefore, we gradually aggregate F'_l into deeper network layers in the encoders so as to ensure feature representation retains detail richness and semantic understanding ability. The aggregated feature map F_{l+1}^{agg} can be expressed as:

$$F_{l+1}^{agg} = FA_E(F_{l+1}, (F'_l \odot P_l)) \quad (14)$$

where $FA_E(\cdot)$ represents the feature aggregation of the encoder layers, and $\odot$ represents Hadamard product.

C. Edge Constraint Module

In order to further improve the integrity and the continuity of the geometric shape under the condition of low contrast, the Edge Constraint (EC) module is introduced. The module integrates the prior knowledge of the physical model into the deep learning framework to generate segmentation results that are more in line with the ground truth geometry.

Specifically, an initial set of boundary points B is first extracted from the input image with the help of traditional edge detection algorithms. Then, define a circular window centered $O(r, b_i)$ at the boundary point $b_i \in B$ with a radius r of (10 pixels by default), and calculate the proportion $p(b_i)$ of the foreground area within this window. As shown in Eq. 15:

$$p(b_i) = \frac{Inter(O(r, b_i), I^{fore})}{O(r, b_i)} \quad (15)$$

When the ratio $p(b_i)$ is closer to 0 or 1, it indicates irregular boundaries within that window. To quantify this irregularity, a scoring function $s(b_i)$ is introduced as shown in Eq. 16:

$$s(b_i) = |p(b_i) - 0.5| \quad (16)$$

The higher the score function $s(b_i)$, the greater the irregular near the boundary point b_i , and the more helpful it is for edge continuity learning. In order to highlight the key boundary features, the non-maximum suppression (NMS) technique is used to filter the local maximum of the scoring function $s(b_i)$. Specifically, for each boundary point b_i , we compare its scores with those of its k nearest neighbors (k is set to 10 by default) and keep only those boundary points whose scores are greater than nearest neighbors. Based on the filtered set of boundary points, the label of the size region around the

retained boundary point b_i is set to 1, while other points are set to 0. Thus, a ground truth map of the boundary key point set M_{truth} is obtained. Different from M_{truth} , the predicted boundary key point set M_{pred} is retained as the predicted probability value.

By imposing constraints on the key point set M_{pred} predicted by the network and the ground truth key point set M_{truth} , the network is guided to generate a more complete and continuous boundary representation. The loss $\mathcal{L}_{match}$ between M_{pred} and M_{truth} is measured in the form of a cross-entropy loss function $\mathcal{L}_{CE}$, which is defined as:

$$\mathcal{L}_{match} = \mathcal{L}_{CE}(M_T, M_{pred}) \quad (17)$$

The boundary coherence loss $\mathcal{L}_{cont}$ focuses on the spatial relationship between key points in M_{pred} , ensuring that they form a continuous path on the image. This is achieved by calculating the predicted boundary key point difference between adjacent pixels, which is defined as follows:

$$\mathcal{L}_{cont} = \sum_{i=1}^{N-1} \rho_{i,i+1} \cdot |M_{pred}^i - M_{pred}^{i+1}| \quad (18)$$

where $\rho_{i,i+1}$ is a weight calculated by the distance of $\|M_{pred}^i - M_{truth}^j\|_2 - \|M_{pred}^{i+1} - M_{truth}^{j+1}\|_2$ and $\|M_{pred}^i - M_{pred}^{i+1}\|_2$, to reflect the coherence requirements between pixels, j denotes the point nearest i . The total loss of the EC module is defined as:

$$\mathcal{L}_{EC} = \frac{1}{2} \cdot (\mathcal{L}_{match} + \mathcal{L}_{cont}) \quad (19)$$

D. Loss function

Our FASS framework overall loss consists of three parts: the supervised loss $\mathcal{L}_{sup}$, the distribution difference loss $\mathcal{L}_D$, and the edge constraint loss $\mathcal{L}_{EC}$, as shown in Eq. 20.

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda(t)(\mathcal{L}_D + \mathcal{L}_{EC}) \quad (20)$$

where $\lambda(t)$ is the time-varying Gaussian heating coefficient [55] used to balance the above loss, which can be expressed as $\lambda(t) = 0.1 * \exp(-5(1 - \frac{t}{t_{max}})^2)$, t and t_{max} denote the current iteration number and the total iteration number, respectively. It is worth noting that we adopt hybrid loss as the supervision loss $\mathcal{L}_{sup}$, which can be expressed as:

$$\mathcal{L}_{sup} = \frac{1}{2} \cdot (\mathcal{L}_{Dice}(R_i, Y_{R_i}) + \mathcal{L}_{CE}(R_i, Y_{R_i})) \quad (21)$$

IV. EXPERIMENT

A. Datasets

To comprehensively evaluate our algorithm, we selected three representative low-contrast abdominal image datasets.

MSD Pancreas Dataset provided by the MICCAI 2018 Medical Segmentation Decathlon (MSD) challenge [56]. It comprises 281 CT scans along with their corresponding pancreas and tumor labels, with a median resolution of $0.8 \times 0.8 \times 2.5$ mm³. Out of these, 225 scans were used for training, while the rest were reserved for testing.

NIH Dataset consists of 82 abdominal CT scans annotated with pancreas labels [57]. Volume sizes range from 512 ×

512×181 to $512 \times 512 \times 466$. Following [19], we selected 62 samples for training and reserved 20 samples for testing.

LiMT Dataset were collected from the Affiliated Hospital of Jiangsu University. It covers four types of liver diseases: hepatocellular carcinoma (HCC), metastatic liver cancer, hemangioma, and liver cyst. The dataset includes 100 volumes of arterial phase CT scans, each annotated and verified by experienced clinical experts. For the experiments, 80 scans were used for training, while the remaining 20 were used for testing.

The MSD pancreas dataset is used to highlight the relationship between organs and small tumors; the NIH dataset evaluates segmentation performance for a single organ within complex backgrounds; and the LiMT dataset further assesses the model's ability to segment organs and distinguish various tumor subtypes.

B. Evaluation Metrics

To comprehensively evaluate the proposed segmentation method, we employ four metrics: Dice Similarity Coefficient (Dice), Jaccard Index (Jaccard), 95th Hausdorff Distance (95HD), and Average Surface Distance (ASD). The Dice and Jaccard metrics evaluate overlap with ground truth, where higher values indicate better overlap. 95HD measures boundary discrepancy, and ASD reflects average surface distance, with lower values indicating closer alignment.

C. Implementation Details

All experiments were conducted on a computing platform with an NVIDIA RTX A6000 GPU and 48 GB of RAM. Given U-Net's [24] strong performance and adaptability in medical image segmentation, we adopted it as our baseline network. Each experiment ran for 30,000 iterations to thoroughly optimize the model parameters. Model optimization was performed using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001 to enhance generalization. Random rotation is introduced to do data augmentation. A five-fold cross-validation was employed for systematic evaluation. The selection of parameter α and its impact on model performance will be detailed discussed in the subsequent parameter sensitivity analysis section IV-F.

D. Comparison with State-of-the-Art Methods

To demonstrate the superiority of our framework, we conducted comprehensive evaluations across three medical datasets to systematically compare our approach against methods from three different research directions. These directions cover: i) *Mainstream medical image segmentation methods*: 3D U-Net [24], V-Net [25], Swin UNETR [58], RC-3DUNet [19], nnU-Net (3D) [27], and U-Mamba [59]; ii) *Frequency domain-enhanced segmentation methods*: WU-Net [60], XNet [61], FET [62], and SASAN [63]; iii) *Low-contrast medical image segmentation methods*: HMedN [22] and TBNNet [23].

1) *Quantitative Analysis*: We report the comparison of our FASS with state-of-the-art segmentation methods on the MSD pancreas dataset, as shown in Table I. From this table, our FASS achieves a Dice score of 87.85% and a Jaccard index of 78.02% in pancreas segmentation, marking an improvement of at least 2% over other general and frequency-domain methods. In tumor segmentation, our model reaches a Dice score of 60.49%, which is at least 3% higher than the best frequency-domain methods (FET and WU-Net), and shows improved overlap in the Jaccard index. In addition, our model achieves low boundary error values, with ASD scores of 1.06 mm for pancreas and 4.55 mm for tumor segmentation, both notably lower than those of other methods.

Table II presents the segmentation results on the NIH dataset, where our method consistently outperforms other state-of-the-art approaches across all metrics. In pancreas segmentation, FASS achieves a Dice score of 87.76% and a Jaccard index of 78.34%, indicating improved accuracy. For boundary accuracy, our method records 2.76 mm (HD95) and 0.88 mm (ASD), demonstrating a significant advantage over competing approaches.

Table III shows segmentation results on the LiMT dataset, with our method consistently achieving the highest scores across all metrics. In Dice and Jaccard scores, our method demonstrates substantial improvements in both liver and liver tumor segmentation. Specifically, it achieves a Dice score of 96.77% and a Jaccard index of 93.79% in liver segmentation, outperforming others by at least 1%, underscoring its strength in liver region segmentation. For liver tumor segmentation, our model reaches a Dice score of 60.31% and a Jaccard index of 43.47%, exceeding frequency-domain methods like SASAN and TBNNet by at least 2%, reflecting improved precision in capturing tumor boundaries. Besides, our method achieves lower boundary error metrics, with a 95HD of 3.64 mm and an ASD of 1.01 mm for liver segmentation and 23.51 mm (95HD) and 5.64 mm (ASD) for tumor segmentation, confirming our model's robustness in managing complex liver and tumor boundaries.

2) *Qualitative Analysis*: Fig. 3 shows a qualitative comparison of our framework with other methods. In the pancreas and small tumor co-segmentation example in Fig. 3 (a), U-Mamba fails to completely segment the tumor due to insufficient handling of the tumor-pancreas relationship; TBNNet, with limited attention to global structure, results in false positive outputs (indicated by the yellow arrow); and RC-3DUNet and SASAN are affected by background tissues, resulting in missegmentation regions. In contrast, our framework focuses more effectively on the foreground region and considers the intrinsic connection between the organ and tumor. In Fig. 3(b), the method by RC-3DUNet produces discontinuous pancreatic boundary segmentation with a Dice score of only 74.63% (indicated by the yellow arrow), while our framework, with the introduction of the EC module, achieves finer and smoother segmentation, especially in the pancreatic head, closely matching the ground truth. Fig. 3(c) presents a large primary liver cancer segmentation example from the LiMT dataset. Due to the tumor's large size and distortion of liver morphology, RC-3DUNet, U-Mamba, and SASAN struggle

TABLE I
SEGMENTATION RESULTS ON THE MSD DATASET COMPARED WITH OTHER STATE-OF-THE-ART APPROACHES.

Methods	Metrics							
	Dice [%]↑		Jaccard [%]↑		95HD [mm]↓		ASD [mm]↓	
	Pancreas	Tumor	Pancreas	Tumor	Pancreas	Tumor	Pancreas	Tumor
3D U-Net [24]	79.52±7.54	43.18±28.50	65.96±5.32	31.53±20.51	7.69±0.83	35.49±8.39	3.03±0.56	12.20±2.39
V-Net [25]	78.19±8.45	41.75±30.49	64.75±6.51	29.38±24.55	8.00±1.23	37.74±8.67	2.95±0.64	13.30±3.64
Swin UNETR [58]	79.71±8.61	40.83±26.78	66.28±7.87	26.15±22.57	5.86±0.76	38.70±9.73	2.71±0.27	10.88±3.83
RC-3DUNet [19]	84.83±6.55	48.36±26.50	73.25±6.94	42.99±17.83	4.44±0.67	26.98±5.14	1.70±0.38	5.45±2.29
nnU-Net(3D) [27]	84.24±7.53	49.83±25.14	72.78±5.17	42.83±16.75	4.22±0.49	27.07±4.67	1.94±0.37	8.47±1.30
U-Mamba [59]	84.85±6.01	58.55±27.28	73.69±6.54	32.06±19.30	3.73±0.39	26.74±5.28	1.35±0.42	5.83±1.20
WU-Net [60]	81.95±8.08	51.73±27.43	70.07±6.28	34.85±21.82	4.54±0.45	26.51±6.25	1.85±0.26	5.34±1.42
XNet [61]	85.73±7.41	52.15±25.17	75.40±6.75	35.50±18.55	3.82±0.32	27.44±7.31	1.46±0.55	4.90±1.82
FET [62]	84.98±5.71	57.38±23.34	73.95±5.81	42.73±16.38	4.38±0.51	24.46±5.23	1.28±0.35	5.74±1.29
SASAN [63]	85.29±5.96	54.55±22.12	74.71±6.93	37.42±20.72	5.49±0.78	25.47±4.63	1.99±0.48	5.58±1.55
HMEDN [22]	83.95±8.64	50.37±24.88	72.59±6.76	33.54±26.43	5.76±0.81	26.09±5.52	1.53±0.27	6.32±1.47
TBNet [23]	84.73±9.92	57.08±21.78	73.68±7.53	39.84±19.30	5.22±0.62	25.70±5.42	1.15±0.29	5.13±1.42
Ours	87.85±5.30	60.49±18.26	78.02±4.98	43.58±15.30	3.55±0.29	23.99±4.02	1.06±0.21	4.55±0.87

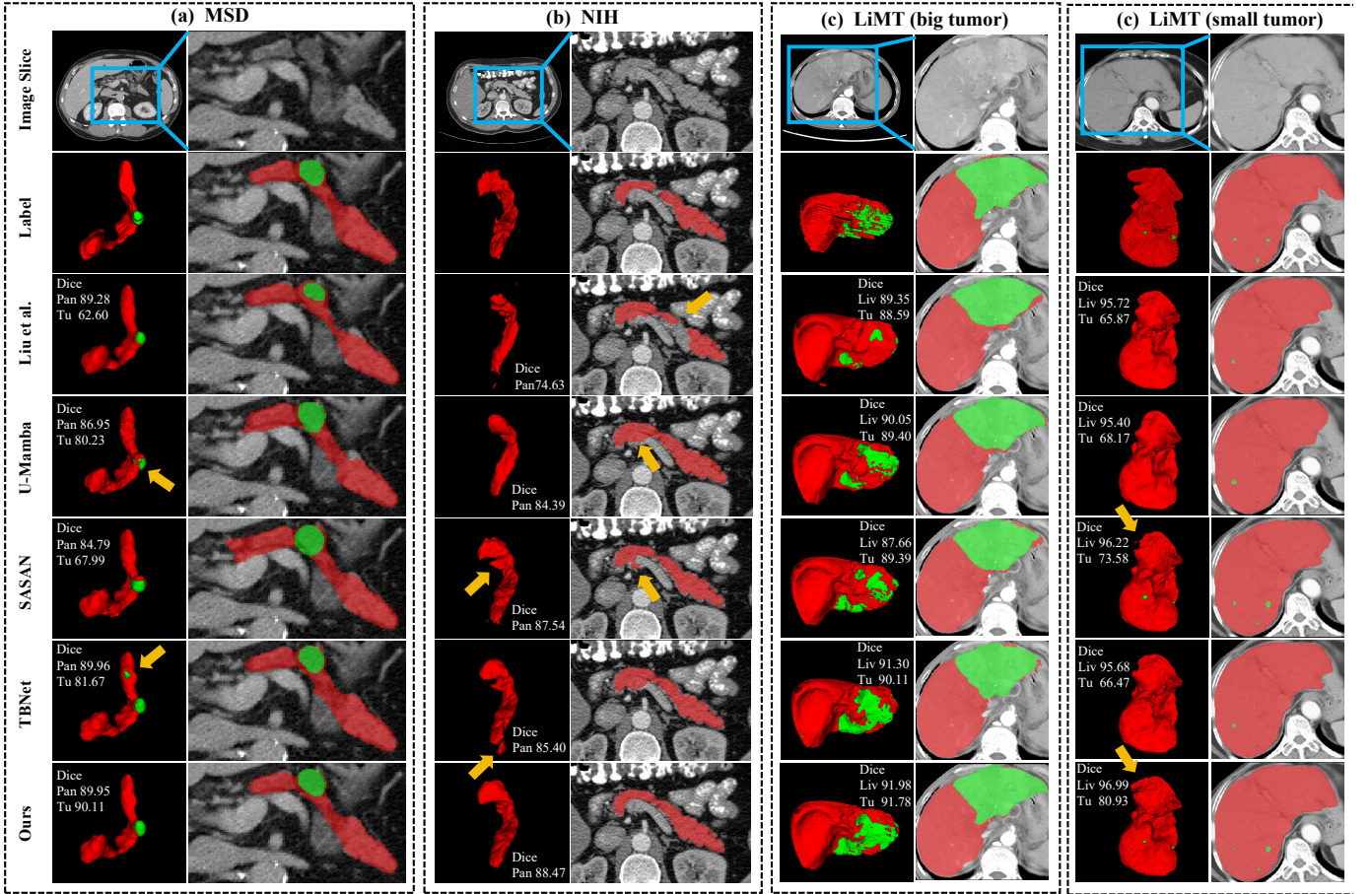


Fig. 3. Qualitative segmentation examples of our framework compared to competing approaches across three datasets demonstrate that our framework significantly enhances segmentation accuracy and integrity, especially in capturing complex tumor shapes and detecting small tumors.

to fully segment the tumor. Our framework achieves higher segmentation completeness, with Dice scores of 91.98% and 91.78% for liver and tumor, respectively. In the hemangioma segmentation example in Fig. 3(d), despite the small tumor size, only SASAN and our framework successfully detect all tumors (indicated by the yellow arrow). In contrast, our

framework closely aligns with the ground truth, achieving a Dice score of 96.99% for the liver and 80.93% for the tumor.

E. Ablation Study

To evaluate the effectiveness of each module in FASS, we conducted detailed ablation experiments and selected 3D U-

TABLE II
SEGMENTATION RESULTS ON NIH DATASET COMPARED WITH OTHER
STATE-OF-THE-ART METHODS.

Methods	Metrics			
	Dice [%]↑	Jaccard [%]↑	95HD [mm]↓	ASD [mm]↓
3D U-Net [24]	79.48±5.06	66.07±4.64	5.53±1.40	2.57±1.14
V-Net [25]	78.53±7.81	64.89±5.32	5.83±1.23	2.66±0.93
Swin UNETR [58]	80.74±6.23	67.78±5.41	4.66±1.56	1.74±1.02
RC-3DUNet [19]	84.62±4.85	74.51±1.92	3.23±0.89	1.13±0.51
nnU-Net(3D) [27]	85.67±4.25	75.02±3.34	3.36±0.57	1.12±0.67
U-Mamba [59]	86.08±3.95	75.57±2.14	3.29±0.53	0.93±0.37
XU-Net [60]	82.45±7.56	70.36±5.73	4.25±1.26	1.07±0.52
XNet [61]	85.67±8.01	74.97±4.19	5.24±0.82	1.53±0.91
FET [62]	83.29±5.15	71.49±4.51	3.73±0.97	1.20±0.73
SASAN [63]	85.30±3.99	74.39±5.73	3.17±1.05	1.16±0.67
HMEDN [22]	84.89±6.17	73.68±4.61	3.20±0.72	1.28±0.88
TBNet [23]	85.03±6.83	73.84±5.08	2.97±0.83	1.03±0.53
Ours	87.76±3.52	78.34±1.62	2.76±0.50	0.88±0.40

Net [24] as the baseline network. The quantitative results summarized in Table II highlight the significant performance improvements contributed by each module. Among them, the FA module demonstrated the core value in the pancreas segmentation task on the MSD and NIH datasets, effectively reducing the false positive prediction of the background region. This improvement is attributed to the model’s ability to accurately focus on pancreas and tumors while neighboring tissues have similar grayscale values with them. However, on the LiMT dataset, where the surrounding tissues of the liver are relatively less similar and the liver shape is more regular, the FA module showed minimal improvement in liver segmentation accuracy. Its advantage became more evident in liver tumor segmentation. The FLFE module performed exceptionally well in liver tumor segmentation on the LiMT dataset. Thanks to the enhancement of features in the frequency domain, different types of liver tumors are more discriminative, thereby optimizing the discriminant ability of the model. The EC module showed limited boundary constraint effects on the NIH dataset, possibly due to the annotation quality, as the unsmooth boundary of ground truth in the NIH dataset may limit the effective application of edge constraints. Furthermore, the combination of different modules can enhance the model’s performance to varying degrees. In summary, the experimental results show that the collaborative use of the FA, FLFE, and EC modules leads to superior segmentation performance.

1) *FA Module*: To evaluate the effectiveness of the FA module, we performed a visual analysis of the features extracted by the encoder layer during the inference stage. Fig. 4 shows the performance of the module’s ability to focus on foreground regions on the LiMT dataset and the MSD pancreas dataset. Pancreas and liver are usually surrounded by complex backgrounds, resulting in low contrast at the boundaries with surrounding tissue. Compared with the baseline method, the introduction of the FA module led to greater unbiased attention to the foreground features. Specifically, Fig. 4 (a) and Fig. 4 (b) demonstrate the feature extraction of liver images with tumors within a complex scenario. In this scenario, the features extracted by the baseline model contain irrelevant background information, resulting in cluttered and unfocused representations. In contrast, when the FA module is introduced,

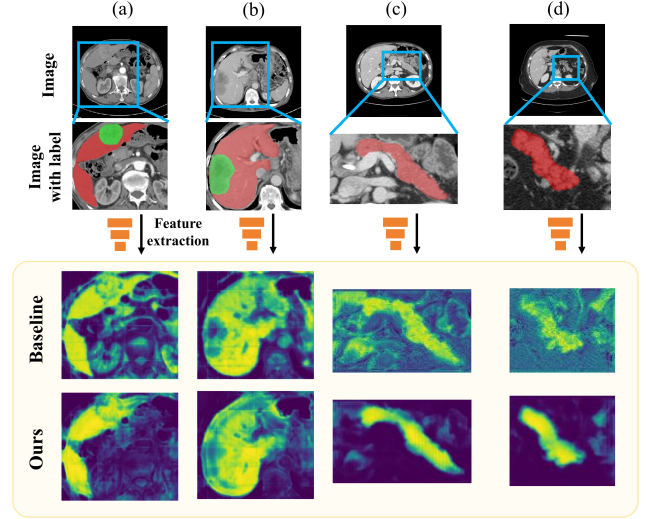


Fig. 4. Visual analysis of encoder feature extraction. With the introduction of the FA module, our method effectively focuses on foreground areas, filtering out the complex background information.

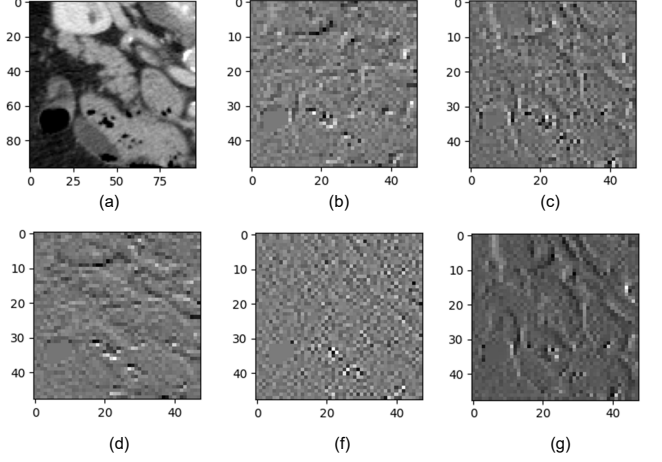


Fig. 5. Visualization comparison of high-frequency features: (a) shows the original image, (b) displays the feature map extracted after the separable convolution operation, and (c)-(f) present the high-frequency components in the vertical, horizontal, and diagonal directions, respectively. (g) illustrates the feature map after integrating each component through the cross-attention mechanism.

the model successfully separates the foreground features from the entire image features, effectively filtering out the complex background information.

Table V reports the performance and efficiency of the algorithm under different background sampling sizes. The results show that when the sampling size is set to $18 \times 18 \times 18$, the model tends to focus excessively on smaller local background areas, which may limit its learning ability and, in turn, affect segmentation performance. In contrast, when the sampling size is set to $48 \times 48 \times 48$, it becomes more difficult to select background areas with low overlap with the foreground regions, hindering the design goals of the foreground perception module and increasing sampling time. In comparison, a sampling size of $32 \times 32 \times 32$ allows the model to efficiently sample appropriate background areas in a shorter time, demonstrating

TABLE III
SEGMENTATION RESULTS ON LiMT DATASET COMPARED WITH OTHER STATE-OF-THE-ART METHODS.

Methods	Metrics							
	Dice [%]↑		Jaccard [%]↑		95HD [mm]↓		ASD [mm]↓	
	Liver	Tumor	Liver	Tumor	Liver	Tumor	Liver	Tumor
3D U-Net [24]	92.99±2.85	53.84±5.25	86.94±2.27	36.87±5.03	8.13±5.17	37.92±10.83	3.41±1.73	9.34±4.76
V-Net [25]	91.56±3.17	54.84±7.38	84.38±2.35	37.89±7.28	7.75±6.38	39.49±12.58	3.48±1.48	9.97±4.24
Swin UNETR [58]	90.23±5.74	50.52±7.65	82.49±4.63	33.97±7.03	6.75±5.02	31.98±9.33	4.23±1.80	8.76±4.45
RC-3DUNet [19]	90.17±3.58	51.55±9.48	82.17±3.34	34.75±8.98	4.69±2.52	24.06±5.60	1.27±0.75	7.07±2.80
nnU-Net (3D) [27]	94.16±2.13	57.19±4.36	88.97±2.26	40.15±4.01	4.54±2.59	27.77±5.00	1.29±0.61	6.43±2.42
U-Mamba [59]	95.83±2.19	56.65±3.86	92.19±1.98	39.56±3.50	4.31±1.43	25.78±4.96	1.46±0.79	6.62±2.49
XU-Net [60]	94.35±2.18	53.14±7.21	89.43±2.03	35.56±6.36	5.43±2.70	32.01±7.16	2.09±1.32	5.94±3.58
XNet [61]	92.73±3.02	58.55±4.83	86.48±2.56	42.48±4.27	6.23±3.87	29.39±10.28	1.66±0.78	6.67±3.30
FET [62]	93.39±2.35	56.64±4.02	87.98±1.75	39.58±3.81	4.83±1.67	26.93±7.88	1.83±1.05	6.31±2.43
SASAN [63]	95.88±2.06	58.81±3.88	92.06±1.75	41.69±3.26	3.98±2.01	24.46±5.36	1.08±0.67	5.73±3.13
HMEDN [22]	92.76±2.41	53.37±8.46	86.75±1.87	36.57±5.89	4.27±1.50	25.58±6.03	1.20±0.73	5.82±2.82
TBNet [23]	95.83±1.76	57.92±3.97	92.03±1.44	40.65±3.51	4.78±1.83	24.07±5.26	1.13±0.89	6.08±2.45
Ours	96.77±1.46	60.31±3.44	93.79±1.32	43.47±2.28	3.64±1.15	23.51±4.79	1.01±0.59	5.64±2.21

TABLE IV
ABLATION STUDY OF EACH MODULE ON THREE DATASETS.

Methods				MSD dataset		NIH dataset	LiMT dataset	
Baseline	FA	FLFE	EC	Pancreas	Tumor	Pancreas	Liver	Tumor
✓				79.52±7.54	43.18±28.50	79.48±5.06	92.99±2.85	53.84±5.25
✓	✓			83.02±5.44	56.47±20.23	83.57±4.38	93.07±2.61	57.18±4.07
✓		✓		82.77±5.97	49.37±22.79	82.23±4.82	94.18±1.93	57.21±3.89
✓			✓	81.03±6.67	46.29±27.91	80.04±5.36	94.60±2.75	55.61±4.47
✓	✓		✓	85.36±5.72	56.95±22.57	85.15±2.37	95.42±2.53	58.48±3.81
✓		✓	✓	84.91±5.98	52.73±25.41	83.44±4.93	95.82±2.44	58.56±3.56
✓	✓	✓		86.73±5.85	58.36±19.55	86.61±3.89	96.23±1.57	59.28±3.72
✓	✓	✓	✓	87.85±5.30	60.49±18.26	87.76±3.52	96.77±1.46	60.31±3.44

TABLE V
COMPARISON OF PERFORMANCE AND TIME FOR DIFFERENT BACKGROUND REGION SAMPLING SIZES ON THE MSD DATASET.

Size	Dice (%) ↑		Sampling time (ms) ↓
	Pancrea	Tumor	
18×18×18	84.78±7.35	57.15±20.06	1.67
32×32×32	87.85±5.30	60.49±18.26	2.18
48×48×48	83.02±6.64	58.30±21.23	8.54

the best segmentation performance across the three datasets.

2) *FLFE Module*: In medical images, the precise capture and utilization of high-frequency features are crucial for detail recognition and edge delineation. The proposed method enhances the representation of high-frequency details by applying cross-attention among high-frequency components in the horizontal, vertical, and diagonal directions, effectively leveraging their complementary advantages. As shown in Fig. 5(b), experiments on the MSD pancreas dataset demonstrate that although the direct convolution operations can extract basic features, the edge and texture performance are slightly blurred. In contrast, as shown in Fig. 5(g), after integrating high-frequency components through the cross-attention mechanism, the resulting feature map exhibits clearer edges and texture

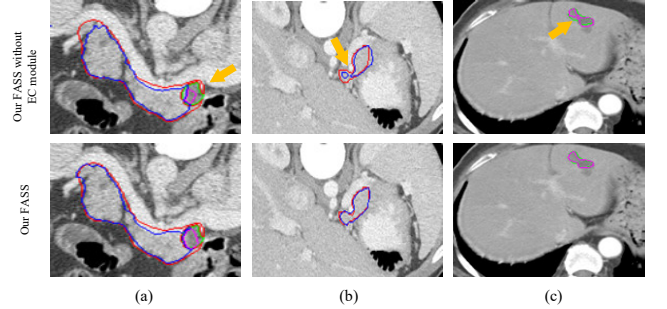


Fig. 6. Visualization of example results with and without the edge constraint module. (a), (b), and (c) show sample results from the MSD pancreas dataset, the NIH dataset, and the LiMT dataset, respectively. Red and green lines represent the ground truth for organs and tumors, while blue and purple lines indicate the predicted boundaries for organs and tumors. With the introduction of the EC module, our method achieves improved boundary continuity and smoothness.

details. This demonstrates the effective complementarity of high-frequency components from Fig. 5(c) to Fig. 5(f).

The selection of different wavelet bases affects the performance and time efficiency of the experiment. Therefore, we compare the application effects of various wavelet bases on the MSD pancreas dataset, including Haar, Biorthogonal (Bior

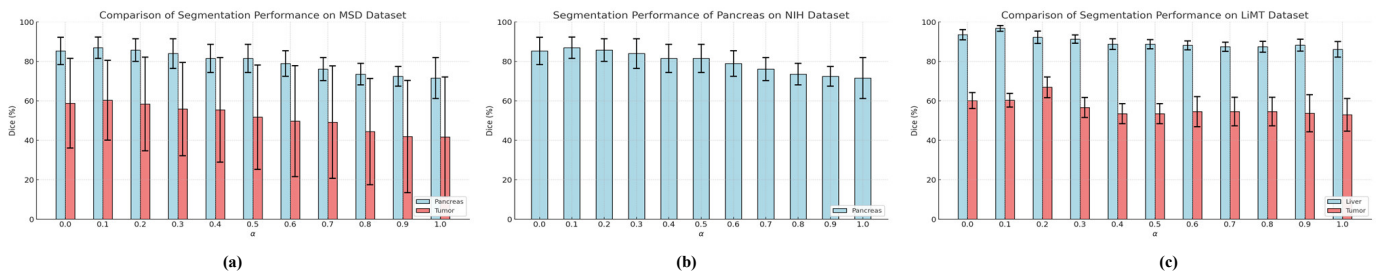


Fig. 7. Comparison of the impact of different α values on model performance. Where (a), (b), and (c) represent the Dice scores and the corresponding standard deviations on the MSD pancreas dataset, NIH dataset, and LiMT dataset, respectively.

TABLE VI
COMPARISON OF DIFFERENT WAVELET BASES ON MSD DATASET.

Wavelet	Dice (%) $\uparrow$		ASD (mm) $\downarrow$		Wavelet decomposition time (ms) $\downarrow$
	Pancreas	Tumor	Pancreas	Tumor	
Haar	84.76 $\pm$ 6.33	58.96 $\pm$ 22.82	1.65 $\pm$ 0.81	7.72 $\pm$ 2.08	0.26
Bior 2.4	84.13 $\pm$ 6.94	57.44 $\pm$ 21.76	1.66 $\pm$ 0.65	11.09 $\pm$ 2.53	1.06
Coif 1	86.59 $\pm$ 5.27	60.78$\pm$20.54	1.41 $\pm$ 0.34	6.12 $\pm$ 1.11	2.31
Db 2	87.85$\pm$5.30	60.49 $\pm$ 18.26	1.06$\pm$0.21	4.55$\pm$0.87	0.73

2.4), Coiflets (Coif 1), and Daubechies (Db 2), as shown in Table VI. The experimental results show that although the Haar wavelet basis exhibits high boundary sensitivity, due to its simple structure, its segmentation accuracy is limited in dealing with texture details, particularly in tumor segmentation. The Coif 1 wavelet basis achieves the best performance in tumor segmentation due to its excellent symmetry and attenuation characteristics, but this comes at the cost of high computational complexity and training time. In contrast, the Db 2 wavelet basis maintains both high segmentation accuracy and time efficiency. Therefore, Db 2 is chosen as the benchmark wavelet basis for the wavelet transform in this paper.

3) *EC Module*: To further demonstrate the effectiveness of the EC module, visual results of selected samples are presented, as shown in Fig. 6. The baseline segmentation results (see the first row of Fig. 6) exhibit poor boundary continuity and smoothness, particularly in the areas indicated by the yellow arrows. In contrast, after incorporating the EC module, the segmentation boundaries in the second row of Fig. 6 show significant improvement. This change indicates that the EC module plays a crucial role in enhancing the segmentation results, improving the model's adaptability to low-contrast images, and consequently increasing overall segmentation accuracy.

F. Parameter Sensitivity Analysis

Impact of parameter α : As a key regulatory factor in the FA module, the parameter α dominates the sampling position of the background region and ranges from $[0, 1]$. As shown in Fig. 7, this experiment systematically explored the specific impact of different α values on model performance. Theoretically, an α value closer to 1 indicates fewer background elements in the sampling region. Ideally, $\alpha = 0$ represents an ideal adversarial training pattern that is sampled entirely

from the background. However, the experimental results show that the peak performance is not achieved when $\alpha = 0$ but rather in a small range near zero. Specifically, when α is set to a relatively low value (e.g., 0.1), the model shows excellent performance on three datasets. In this case, the selected background region is close to the foreground boundary with minimal overlap, but it still moderately expands the background features compared to $\alpha > 0.1$. This phenomenon may be due to the fact that a small number of overlaps will deepen the model's understanding of low-contrast adjacent background tissues. On the contrary, a higher α value results in excessive inclusion of foreground information during sampling, contrary to the design intent of the FA module, thus negatively affecting its efficacy. When α is set to 0.2, the model shows slight improvement in tumor segmentation performance on the LiMT dataset compared to $\alpha = 0.1$. This may be due to the model incorrectly identifying liver tissue as part of the background, unintentionally reducing phenotypic contrast between the liver and tumors, which, in turn, stimulates the model to pay more attention to the tumor region, thereby improving the segmentation accuracy to some extent. Based on the analysis of the experimental results of parameter α , when $\alpha = 0.1$, the model shows excellent segmentation performance on the three datasets, which confirms the importance of reasonable choice of α value for the overall segmentation performance.

V. CONCLUSION

In this paper, we proposed a foreground-aware spectrum segmentation (FASS) framework for low-contrast medical images. First, the foreground-aware module forces the model to focus on the target areas through adversarial training of background features and global features. Then, a feature-level frequency enhancement strategy is designed to better segment fine anatomical structures, along with an edge constraint that

aligns edge prediction with expected contours, enhancing boundary continuity. Extensive experiments demonstrate that the proposed method outperforms others in segmentation performance across multiple medical image datasets. FASS not only holds promise for advancing the clinical application of low-contrast medical image segmentation but also provides more reliable data support for clinical decision-making.

REFERENCES

- [1] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: a cancer journal for clinicians*, vol. 74, no. 1, 2024.
- [2] K. Han, C. Lyu, L. Ma, C. Qian, S. Ma, J. Chen, and Z. Liu, "Climd: A curriculum learning framework for imbalanced multimodal diagnosis," *arXiv preprint arXiv:2508.01594*, 2025.
- [3] K. Han, S. Wang, J. Chen, C. Qian, C. Lyu, S. Ma, V. S. Sheng, Q. Huang, and Z. Liu, "Region uncertainty estimation for medical image segmentation with noisy labels," *IEEE Transactions on Medical Imaging*, 2025.
- [4] C. Qian, K. Han, S. Ma, C. Lyu, Z. Yuan, J. Chen, and Z. Liu, "Adaptive label correction for robust medical image segmentation with noisy labels," *arXiv preprint arXiv:2503.12218*, 2025.
- [5] J. Liu, Y. Hu, J. Yang, Y. Chen, H. Shu, L. Luo, Q. Feng, Z. Gui, and G. Coatrieux, "3d feature constrained reconstruction for low-dose ct imaging," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1232–1247, 2016.
- [6] J. Dong, Y. Cong, G. Sun, Y. Yang, X. Xu, and Z. Ding, "Weakly-supervised cross-domain adaptation for endoscopic lesions segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, 2020.
- [7] J.-H. Shi, Q. Zhang, Y.-H. Tang, and Z.-Q. Zhang, "Polyp-mixer: An efficient context-aware mlp-based paradigm for polyp segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 30–42, 2022.
- [8] K. Han, V. S. Sheng, Y. Song, Y. Liu, C. Qiu, S. Ma, and Z. Liu, "Deep semi-supervised learning for medical image segmentation: A review," *Expert Systems with Applications*, p. 123052, 2024.
- [9] X. Wang, D. Cai, S. Yang, Y. Cui, J. Zhu, K. Wang, and J. Zhao, "Sac-net: enhancing spatiotemporal aggregation in cervical histological image classification via label-efficient weakly supervised learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [10] Y. Zhou, Y. Li, Z. Zhang, Y. Wang, A. Wang, E. K. Fishman, A. L. Yuille, and S. Park, "Hyper-pairing network for multi-phase pancreatic ductal adenocarcinoma segmentation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 155–163.
- [11] X. Chen, Z. Chen, J. Li, Y.-D. Zhang, X. Lin, and X. Qian, "Model-driven deep learning method for pancreatic cancer segmentation based on spiral-transformation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 1, pp. 75–87, 2021.
- [12] Y. Ding, C. Zhang, M. Cao, Y. Wang, D. Chen, N. Zhang, and Z. Qin, "Tostagan: An end-to-end two-stage generative adversarial network for brain tumor segmentation," *Neurocomputing*, vol. 462, pp. 141–153, 2021.
- [13] Q. Yu, D. Yang, H. Roth, Y. Bai, Y. Zhang, A. L. Yuille, and D. Xu, "C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4126–4135.
- [14] Z. Yuan, J. Cao, Z. Li, H. Jiang, and Z. Wang, "Sd-mvs: Segmentation-driven deformation multi-view stereo with spherical refinement and em optimization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 7, 2024, pp. 6871–6880.
- [15] Z. Yuan, J. Cao, Z. Wang, and Z. Li, "Tsar-mvs: Textureless-aware segmentation and correlative refinement guided multi-view stereo," *Pattern Recognition*, vol. 154, p. 110565, 2024.
- [16] Z. Yuan, X. Qu, C. Qian, R. Chen, J. Tang, L. Sun, X. Chu, D. Zhang, Y. Wang, Y. Cai, and S. Li, "Video-star: Reinforcing open-vocabulary action recognition with tools," *arXiv preprint arXiv:2510.08480*, 2025.
- [17] Z. Yuan, C. Liu, F. Shen, Z. Li, J. Luo, T. Mao, and Z. Wang, "Msp-mvs: Multi-granularity segmentation prior guided multi-view stereo," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9753–9762.
- [18] Z. Yuan, Z. Yang, Y. Cai, K. Wu, M. Liu, D. Zhang, H. Jiang, Z. Li, and Z. Wang, "Sed-mvs: Segmentation-driven and edge-aligned deformation multi-view stereo with depth restoration and occlusion constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- [19] Z. Liu, S. Ma, Y. Liu, W. Wang, Y. Song, J. Su, Y. Tang, A. Yu, and X. Liu, "Pancreas segmentation in ct based on rc-3dunet with som," *Multimedia Systems*, vol. 30, no. 2, p. 66, 2024.
- [20] Y. Jiang, Z. Zhang, S. Qin, Y. Guo, Z. Li, and S. Cui, "Apaunet: axis projection attention unet for small target in 3d medical segmentation," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 283–298.
- [21] Q. Li, X. Liu, Y. He, D. Li, and J. Xue, "Temperature guided network for 3d joint segmentation of the pancreas and tumors," *Neural Networks*, vol. 157, pp. 387–403, 2023.
- [22] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder-decoder networks for low-contrast medical image segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 461–475, 2019.
- [23] Y. Zhang, R. Xi, W. Wang, H. Li, L. Hu, H. Lin, D. Towey, R. Bai, H. Fu, R. Higashita *et al.*, "Low-contrast medical image segmentation via transformer and boundary perception," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.
- [24] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.
- [25] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.
- [26] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "3d transunet: Advancing medical image segmentation through vision transformers," *arXiv preprint arXiv:2310.07781*, 2023.
- [27] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [28] B. Subramani, M. Veluchamy, and A. K. Bhandari, "Optimal fuzzy intensification system for contrast distorted medical images," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.
- [29] C. Xu, T. Zhang, D. Zhang, D. Zhang, and J. Han, "Deep generative adversarial reinforcement learning for semi-supervised segmentation of low-contrast and small objects in medical images," *IEEE Transactions on Medical Imaging*, 2024.
- [30] D. Xiang, T. Peng, Y. Bian, L. Chen, J. Zeng, F. Shi, W. Zhu, and X. Chen, "Unpaired dual-modal image complementation learning for single-modal medical image segmentation," *IEEE Transactions on Biomedical Engineering*, 2024.
- [31] W. Fu, H. Hu, X. Li, R. Guo, T. Chen, and X. Qian, "A generalizable causal-invariance-driven segmentation model for peripancreatic vessels," *IEEE Transactions on Medical Imaging*, 2024.
- [32] K. Han, S. Ma, C. Lyu, C. Qian, X. Qiu, J. Chen, Y. Liu, Y. Song, Y. Zhu, L. Tian *et al.*, "Limt: A multi-task liver image benchmark dataset," *arXiv preprint*, 2024.
- [33] Z. Yuan, D. Zhang, Z. Li, C. Qian, J. Chen, Y. Chen, K. Chen, T. Mao, Z. Li, H. Jiang *et al.*, "Dvp-mvs++: Synergize depth-normal-edge and harmonized visibility prior for multi-view stereo," *arXiv preprint arXiv:2506.13215*, 2025.
- [34] —, "Dvp-mvs++: Synergize depth-normal-edge and harmonized visibility prior for multi-view stereo," *IEEE Transactions on Circuits and Systems for Video Technology (Under Review)*, 2025.
- [35] J. Chen, Z. Li, Y. Cai, H. Jiang, C. Qian, J. Kang, S. Gao, H. Zhao, T. Mao, and Y. Zhang, "Haif-gs: Hierarchical and induced flow-guided gaussian splatting for dynamic scene," *arXiv preprint arXiv:2506.09518*, 2025.
- [36] Z. Zhu, P. Zhou, C. Qian, R. Yang, Y. Ye, and J. Zhu, "Contrastive intra- and inter-modal clustering for multimodal semantic discovery," *arXiv preprint*, 2025.
- [37] Z. Li, Z. Zhu, J. Zhu, P. Zhou, J. Du, K. Chen, and C. Qian, "Rusc: Integrating regularization and unsupervised contrastive learning for automatic speech recognition," *arXiv preprint*, 2025.
- [38] S. Chen, G. Bortsova, A. García-Uceda Juárez, G. Van Tulder, and M. De Bruijne, "Multi-task attention-based semi-supervised learning for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019: 22nd International Con-*

- ference, Shenzhen, China, October 13–17, 2019, *Proceedings, Part III* 22. Springer, 2019, pp. 457–465.
- [39] C. Qian, S. Xing, S. Li, Y. Zhao, and Z. Tu, “Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning,” *arXiv preprint arXiv:2503.11892*, 2025.
- [40] C. Qian, K. Han, J. Wang, Z. Yuan, C. Lyu, J. Chen, and Z. Liu, “Dyncim: Dynamic curriculum for imbalanced multimodal learning,” *arXiv preprint arXiv:2503.06456*, 2025.
- [41] Q. Hao, Q. Gan, Z. Liu, J. Chen, Q. Shen, C. Qian, and Y. Liu, “Ssdnet: An effective classification method of steel surface defects based on salient local features,” in *International Conference on Intelligent Computing*. Springer Nature Singapore, 2024, pp. 490–503.
- [42] Z. Yuan, J. Tang, J. Luo, R. Chen, C. Qian, L. Sun, X. Chu, Y. Cai, D. Zhang, and S. Li, “Autodrive-r2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving,” *arXiv preprint arXiv:2509.01944*, 2025.
- [43] Z. Xu, T. Li, Y. Liu, Y. Zhan, J. Chen, and T. Lukasiewicz, “Pacnet: Multi-pathway fpn with position attention guided connections and vertex distance iou for 3d medical image detection,” *Frontiers in Bioengineering and Biotechnology*, vol. 11, p. 1049555, 2023.
- [44] X. Wang, S. Gou, J. Li, Y. Zhao, Z. Liu, C. Jiao, and S. Mao, “Self-paced feature attention fusion network for concealed object detection in millimeter-wave image,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 224–239, 2021.
- [45] J. M. J. Valanarasu, P. Oza, I. Hacıhaliloglu, and V. M. Patel, “Medical transformer: Gated axial-attention 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, 2021, pp. 36–46.
- [46] Y. Yu, K. Zhang, X. Wang, N. Wang, and X. Gao, “An adaptive region proposal network with progressive attention propagation for tiny person detection from uav images,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [47] C. Tian, M. Zheng, W. Zuo, B. Zhang, Y. Zhang, and D. Zhang, “Multi-stage image denoising with the wavelet transform,” *Pattern Recognition*, vol. 134, p. 109050, 2023.
- [48] X. Yin and X. Xu, “A method for improving accuracy of deeplabv3+ semantic segmentation model based on wavelet transform,” in *International Conference in Communications, Signal Processing, and Systems*. Springer, 2021, pp. 315–320.
- [49] Y. Zhao, S. Wang, Y. Zhang, S. Qiao, and M. Zhang, “Wranet: wavelet integrated residual attention u-net network for medical image segmentation,” *Complex & intelligent systems*, vol. 9, no. 6, pp. 6971–6983, 2023.
- [50] W. Gao, X. Li, Y. Wang, and Y. Cai, “Medical image segmentation algorithm for three-dimensional multimodal using deep reinforcement learning and big data analytics,” *Frontiers in Public Health*, vol. 10, p. 879639, 2022.
- [51] T. T. Showrav and M. K. Hasan, “Hi-gmisnet: generalized medical image segmentation using dwt based multilayer fusion and dual mode attention into high resolution pgan,” *Physics in Medicine & Biology*, vol. 69, no. 11, p. 115019, 2024.
- [52] S. Jin, S. Yu, J. Peng, H. Wang, and Y. Zhao, “A novel medical image segmentation approach by using multi-branch segmentation network based on local and global information synchronous learning,” *Scientific Reports*, vol. 13, no. 1, p. 6762, 2023.
- [53] R. Azad, A. Bozorgpour, M. Asadi-Aghbolaghi, D. Merhof, and S. Escalera, “Deep frequency re-calibration u-net for medical image segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3274–3283.
- [54] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [55] S. Adiga, J. Dolz, and H. Lombaert, “Anatomically-aware uncertainty for semi-supervised image segmentation,” *Medical Image Analysis*, vol. 91, p. 103011, 2024.
- [56] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” *arXiv preprint arXiv:1902.09063*, 2019.
- [57] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, “Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part I* 18. Springer, 2015, pp. 556–564.
- [58] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, “Self-supervised pre-training of swin transformers for 3d medical image analysis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 730–20 740.
- [59] J. Ma, F. Li, and B. Wang, “U-mamba: Enhancing long-range dependency for biomedical image segmentation,” *arXiv preprint arXiv:2401.04722*, 2024.
- [60] Y. Li, Y. Wang, T. Leng, and W. Zhijie, “Wavelet u-net for medical image segmentation,” in *Artificial Neural Networks and Machine Learning—ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I* 29. Springer, 2020, pp. 800–810.
- [61] Y. Zhou, J. Huang, C. Wang, L. Song, and G. Yang, “Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 085–21 096.
- [62] R. Azad, A. Kazerouni, A. Sulaiman, A. Bozorgpour, E. K. Aghdam, A. Jose, and D. Merhof, “Unlocking fine-grained details with wavelet-based high-frequency enhancement in transformers,” in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2023, pp. 207–216.
- [63] X. Huang, J. Huang, K. Zhao, T. Zhang, Z. Li, C. Yue, W. Chen, R. Wang, X. Chen, Q. Zhang *et al.*, “Sasan: Spectrum-axial spatial approach networks for medical image segmentation,” *IEEE Transactions on Medical Imaging*, 2024.