CALDGSEG: A Medical Image Segmentation Framework via Interactive Feature Fusion and Dynamic Guidance

Zihan Xiong, Siyan Xiao
School of Computer Science and Engineering, Northeastern University, Shenyang, China
2401938@stu.neu.edu.cn, 2472114@stu.neu.edu.cn

***Abstract*—Accurate segmentation remains challenging due to ambiguous boundaries and background interference, where foreground and background features are highly entangled. In medical image segmentation, this issue is further aggravated by soft boundaries and noise, while existing models typically address these problems in isolation and rely on fixed feature fusion strategies, resulting in limited robustness. We propose CALDGSeg, a neural network framework that explicitly restructures feature interaction and fusion through feature disentanglement, contrast-aware local attention, and dynamic guidance. A Foreground–Background Attention Disentangler first decouples foreground and background representations, reducing early-stage semantic interference. Subsequently, a Contrast-Aware Local Attention module performs foreground–background interactive weighting within windowed self-attention, enhancing discriminative boundary representation while suppressing background noise. Finally, a dynamically guided decoder learns per-pixel adaptive fusion weights with cascaded channel–spatial optimization to alleviate semantic conflicts during multi-scale decoding. Experiments on three public benchmarks demonstrate that CALDGSeg consistently outperforms state-of-the-art methods and exhibits strong generalization.This framework offers generalizable design insights for robust feature interaction and fusion in dense prediction networks. The code is available at https://github.com/zihan1215/CALDG**

***Index Terms*—Medical image segmentation, soft boundary, interactive feature fusion, dynamic guidance**

I. Introduction

Medical image segmentation is a cornerstone of computer-aided diagnosis and is widely applied in lesion detection and treatment planning, providing critical quantitative evidence for clinical evaluation [1]–[3].However, clinical images often suffer from soft boundaries, complex backgrounds, and noise [4], [5], which make accurate segmentation particularly challenging. From a representation learning perspective, these difficulties are not merely data-specific artifacts, but rather manifestations of foreground–background feature entanglement and semantic conflicts during multi-scale feature fusion. In dense prediction networks, features extracted at different resolutions exhibit heterogeneous semantic reliability, yet they are fused using fixed or heuristic strategies. This mismatch leads to unstable feature interaction, where background responses contaminate foreground representations, especially in regions with weak contrast or complex structures, resulting in blurred boundaries and degraded robustness even when powerful backbones or attention mechanisms are employed.

In recent years, the introduction of U-Net [6] and its variants has significantly advanced medical image segmentation performance. Nevertheless, fuzzy boundaries and background noise remain persistent challenges. Most existing approaches attempt to mitigate these issues by enhancing boundary cues or suppressing noise at isolated stages of the network. However, such strategies typically operate on already entangled feature representations and fail to explicitly model how foreground and background features should interact and be fused across scales. Consequently, semantic conflicts accumulate across layers, limiting both robustness and generalization. These observations suggest that effectively addressing ambiguous boundaries and noise requires a fundamental redesign of feature interaction and fusion mechanisms, rather than incremental architectural modifications.

These limitations show that robust segmentation requires a unified framework for foreground–background disentanglement, interaction, and adaptive fusion. Effective solutions should: (1) reduce early semantic interference via disentanglement, (2) enhance local foreground–background contrast, and (3) dynamically regulate multi-scale fusion. We thus propose CALDGSeg for robust segmentation under ambiguous boundaries and background interference. Our main contributions are:

- We propose the FBAD module, which achieves preliminary separation of foreground and background features by suppressing uncertain regions, thereby providing subsequent processing stages with feature maps rich in information.
- We design the CALA module, which conducts foreground–background attention interaction and fusion to strengthen local feature associations and regional contrast, enabling reliable discrimination of adjacent regions under severe soft boundary and improving the accuracy of segmentation.
- We optimize the DG-Decoder, where a dynamic weight-learning mechanism alleviates the semantic conflicts inherent to equal-weight fusion, and a cascaded "channel-selection-spatial-localization" strategy further enhances robustness to noise.
- We propose CALDGSeg that serially integrates FBAD, CALA, and DG-Decoder, yielding a targeted solution to the dual challenges of fuzzy boundary and noise.

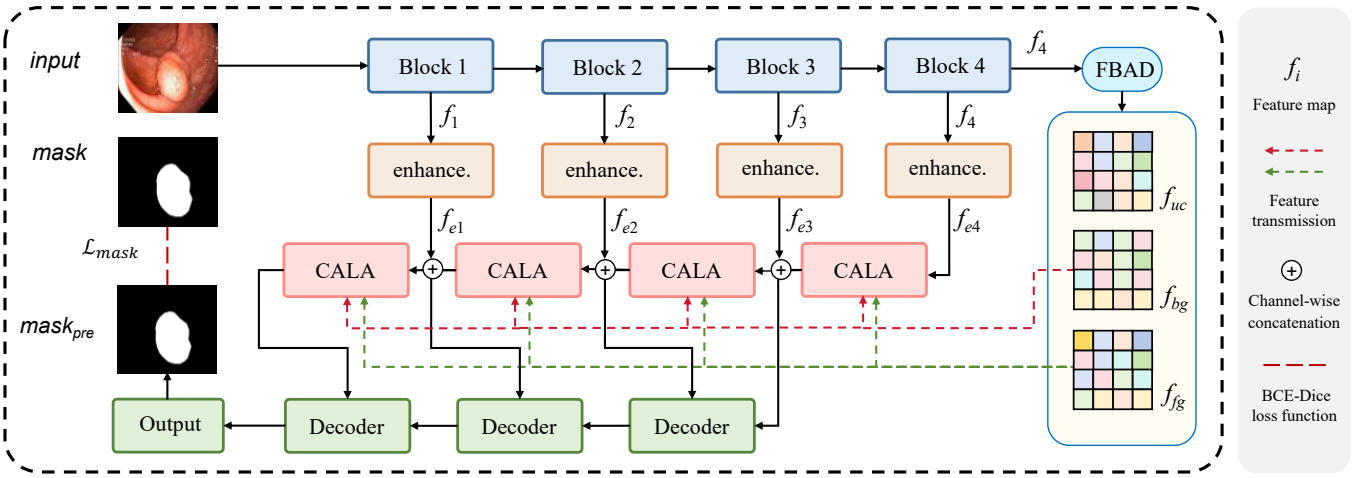

Fig. 1. CALDGSeg Model Structure Diagram.

II. RELATED WORK

A. Encoder–Decoder Architectures

Encoder–decoder architectures are the dominant paradigm for medical image segmentation, integrating low-level spatial details with high-level semantics. Over the past decade, advances in deep learning have driven great progress in medical imaging, producing numerous segmentation models.The U-Net [6] architecture, with its encoder-decoder framework, revolutionised the landscape of medical image segmentation. By skip connections to fuse high- and low-dimensional features, it preserves image detail and semantic information, as the benchmark framework for medical image segmentation. Subsequently, numerous U-Net variants emerged. UNet++ [7] optimised feature fusion through a nested decoder structure, while ResUNet [8] introduced residual connections to mitigate gradient vanishing during deep network training, significantly enhancing segmentation models' representational capabilities. These improvements achieved notable breakthroughs in medical image segmentation such as polyps, glands, and tumours. In recent years, attention mechanisms and Transformer models have opened new avenues for medical image segmentation. U-Net Transformer [9] combines the self-attention mechanism of Transformers with the encoder-decoder architecture of U-Net, preserving the fine details required for medical image segmentation while capturing long-range feature dependencies. Swin Transformer [10] proposes a layered self-attention mechanism based on sliding windows, reducing computational complexity to achieve a balance between high accuracy and efficiency. Despite their success, most encoder–decoder methods use fixed feature fusion, ignoring varied semantic reliability across scales and causing foreground–background conflicts.

B. Boundary Ambiguity

Boundary ambiguity represents a challenge in dense prediction tasks, where foreground and background features are often entangled near object boundaries. In medical image segmentation, the problem is attributed to boundary co-occurrence and soft boundary effects. The issue of boundary co-occurrence or soft boundaries represents a challenge in medical imaging, a bottleneck constraining clinical image segmentation technology. Its essence lies in the gradients between lesions and surrounding tissues, in ambiguous boundary definitions that impact segmentation accuracy and clinical diagnostic reliability. PraNet [11] introduced an inverse attention module to enhance boundary feature capture; ConvSegNet [12] and ECTransNet [13] employed multi-scale feature aggregation strategies, fusing features from different receptive fields to strengthen boundary representation. However, they failed to resolve semantic conflicts arising from feature fusion in noisy environments, demonstrating insufficient robustness in complex soft boundary scenarios. ConDSeg [14] introduces a semantic information decoupling mechanism, separating feature maps into foreground, background, and uncertain regions. However, ConDSeg focuses on representation disentanglement and still relies on fixed feature fusion strategies, without modeling foreground–background interaction or resolving semantic conflicts during multi-scale decoding, and its fixed fusion strategy lacks adaptive adjustment for soft boundary areas, hindering exploitation of boundary information features and the enhancement of soft boundary region recognition.

Meanwhile, clinical demands for practical segmentation models keep growing. Besides accuracy, generalization, deployment efficiency, and interpretability have become key goals. Current models suffer from performance drops across datasets and modalities, and complex structures are too computationally costly for clinical use. Moreover, most methods only handle blurred boundaries or noise separately, without explicit foreground–background interaction and adaptive fusion, leaving semantic conflicts unresolved.

III. METHOD

To address feature entanglement, ambiguous boundaries, and semantic conflicts, we propose a unified framework named CALDGSeg. As illustrated in Fig. 1, the CALDGSeg architecture comprises four components: an encoder, a FBAD module,

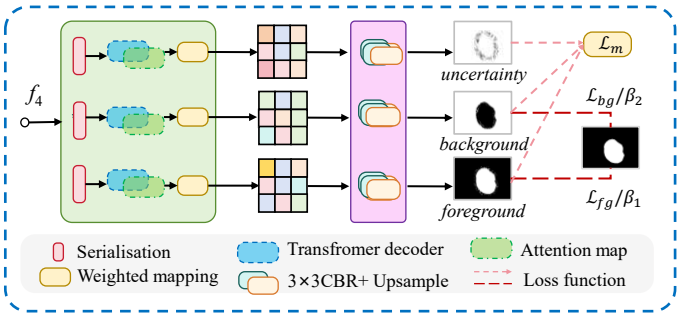

Fig. 2. The structure of FBAD.

a CALA module, and a DG-Decoder. The encoder adopts ResNet-50 [15] as the backbone to extract multi-scale features. The FBAD decouples high-level features into foreground, background, and uncertainty regions. CALA performs feature enhancement by strengthening local feature associations and reinforcing foreground–background contrastive interactions. Finally, the DG-Decoder performs layer-wise decoding to generate the final mask.

A. Encoder

We use ResNet-50 as the encoder to extract multi-scale feature maps f_1–f_4. The encoder is initialized with the official pretrained weights and robustly fine-tuned on the target task. Specifically, ResNet-50 is a mature general-purpose model. We will employ two strategies to specifically optimise the ResNet-50 model for better adaptation to the experimental data. First, we conduct image alignment and use random photometric adjustments to simulate complex real imaging conditions. The augmented and original datasets are then trained together, which strengthens the encoder's feature extraction and improves its robustness to illumination and chromatic aberration changes. Second, a smaller learning rate is used for ResNet-50 blocks during training for better performance. For fairness, all other ResNet-50-based methods are treated identically.

B. Foreground-Background Attention Disentangler (FBAD)

The Foreground-Background Attention Disentangler aims to decompose an image into foreground (M^{fg}), background (M^{bg}), and uncertainty regions (M^{uc}), whilst minimising the proportion of uncertainty regions (M^{uc}) to provide reliable, information-rich foreground (M^{fg}) and background (M^{bg}) for subsequent processing. Among these, (M^{fg}) and (M^{bg}) constitute the core content of this module, while (M^{uc}) serves merely as a process variable for continuously optimizing (M^{fg}) and (M^{bg}). To achieve this functionality, we now define the relationship among the three variables:

$$M^{fg} + M^{bg} + M^{uc} = 1 \tag{1}$$

where $M^{fg}, M^{bg}, M^{uc} \in [0, 1]$.

The specific operation of FBAD is as follows, and its schematic diagram is shown in Fig.2. First, construct a 1×1 convolutional projection layer $\mathcal{P}(\cdot)$ to project the input feature $X \in \mathbb{R}^{B\times C_{\text{in}}\times H\times W}$ into a hidden dimension, namely:

$$E = \mathcal{P}(X) \in \mathbb{R}^{B\times D\times H\times W} \tag{2}$$

Simultaneously, initialise three learnable query vectors $Q = \{q_f, q_b, q_u\} \in \mathbb{R}^{3\times D}$, corresponding respectively to foreground (M^{fg}), background (M^{bg}), and uncertainty regions (M^{uc}). Next, flatten the projected features into a sequence format:

$$E_{\text{flat}} = \text{reshape}(E, [B, N, D]) \tag{3}$$

where $N = H \times W$ denotes the total spatial pixels, and extend the query vectors to batch dimension $Q_B = Q \otimes \mathbf{1}_B \in \mathbb{R}^{B\times 3\times D}$ for input to a two-layer Transformer decoder. The decoded query vector $\hat{Q} \in \mathbb{R}^{B\times 3\times D}$ is then output.
For each decoded query vector $\hat{Q}[:, i, :]$ ($i \in \{f, b, u\}$), compute its attention score with the feature sequence, normalise it, and reshape it into a spatial attention map:

$$\alpha_i = \text{reshape}\left(\text{softmax}\left(\frac{\hat{Q}[:, i, :] \cdot E_{\text{flat}}^T}{\sqrt{D}}\right), [B, 1, H, W]\right) \tag{4}$$

Subsequently, the attention map is element-wise multiplied with the projection feature E to obtain the decoupled regional features:

$$E_f = E \odot \alpha_f, E_b = E \odot \alpha_b, E_u = E \odot \alpha_u \tag{5}$$

(where $\odot$ denotes element-wise multiplication). Finally, each regional feature is mapped to the target dimension $C_{\text{out}} = 128$ via the output projection layer $\mathcal{O}(\cdot)$, comprising a 3×3 convolution, ReLU activation, and 1×1 convolution, yielding

$$O_i = \mathcal{O}(E_i) \in \mathbb{R}^{B\times C_{\text{out}}\times H\times W}, \quad i \in \{f, b, u\} \tag{6}$$

yielding the disentangled features for foreground, background, and uncertain regions: $\{O_f, O_b, O_u\}$. This module explicitly separates feature representations across distinct semantic regions via an attention mechanism, providing refined feature inputs for subsequent segmentation tasks.

To achieve separation of the foreground (M^{fg}) and background (M^{bg}) and suppress overlap between the three image layers, an exclusive penalty term is introduced:

$$\mathcal{L}_m = \frac{1}{N}\sum_{i=1}^{N}\left(M_i^{fg}\cdot M_i^{bg} + M_i^{fg}\cdot M_i^{uc} + M_i^{bg}\cdot M_i^{uc}\right) \tag{7}$$

where $N = W \times H$ is the number of pixels in an image, M_i^{fg}, M_i^{bg} and M_i^{uc} denote the response at the i-th pixel.

The sample ratio between foreground and background classes is typically imbalanced. To balance the weight of each loss component, we introduce penalty factors:

$$\beta_1 = \tanh\left(\frac{\Sigma_{i=1}^{N} M_i^{fg}}{N}\right), \beta_2 = \tanh\left(\frac{\Sigma_{i=1}^{N} M_i^{bg}}{N}\right) \tag{8}$$

We use convolution and upsampling for prediction, and the overall loss function of FBAD is:

$$\mathcal{L}_{FBAD} = \frac{1}{\beta_1}\mathcal{L}_{fg} + \frac{1}{\beta_2}\mathcal{L}_{bg} + \mathcal{L}_m \tag{9}$$

where $\mathcal{L}_{fg}$ and $\mathcal{L}_{bg}$ are supervision losses for M^{fg} and M^{bg}, respectively, computed using a BCE–Dice Loss.

C. Contrast-Aware Local Attention (CALA)

The CALA module leverages the contrastive features f_{fg} and f_{bg} decoupled by the FBAD module to further enhance the contrast between foreground and background features and perform fine-grained feature fusion, thus facilitating reliable separation of target structures from complex backgrounds. The CALA module is designed to explicitly enhance local foreground–background contrast under soft boundary conditions, rather than performing generic attention-based feature aggregation.

Specifically, each f_i is enhanced by its own convolution layer to yield f_{ei}. CALA takes as input feature maps $\hat{f}_i$ together with contrastive features f_{fg} and f_{bg}. Except for the top stage $\hat{f}_4$=f_{e4}, each $\hat{f}_i$ is formed by concatenating the enhanced encoder feature f_{ei} with the previous CALA output, i.e., $\hat{f}_i$=$f_{ei} \oplus F_{i-1}$, and the module outputs a fused feature map F_i (here, $\oplus$ denotes channel-wise concatenation).

The CALA schematic is depicted in Fig.3. First, dual-channel residual feature fusion is applied to the feature map $\hat{f}$: the primary channel concentrates on semantic features, while the residual channel concentrates on detailed features to supplement the original information, thereby providing richer and more refined initial features for subsequent processing. To enhance the local connectivity of feature maps and preserves more critical image details (i.e., edge textures at lesions, branch junctions, etc.), the feature map $x \in \mathbb{R}^{B\times H\times W\times C}$ is partitioned into multiple $x_w \in \mathbb{R}^{B\times G\times N\times C}$, where G denotes the total number of local windows and N denotes the number of pixels within each local window. We achieve this through linear projection:

$$Q = x_w W_Q, K = x_w W_K, V = x_w W_V \tag{10}$$

Where W_Q, W_K, $W_V \in \mathbb{R}^{c\times c}$ are learnable linear projections matrices, and $Q, K, W \in \mathbb{R}^{B\times G\times h\times N\times d}$. For each window we compute a spatial attention:

$$A\left(Q,K,V\right) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{11}$$

Subsequently, a linear projection is applied to obtain the processed features, which are then reshaped into the vectors $P \in \mathbb{R}^{B\times H\times N\times k^2\times d}$, where k denotes the receptive field size. On the one hand, the contrastive features f_{fg} and f_{bg} are enhanced by fusing their respective average pooling and max pooling to improve feature receptivity, and then passed through linear layers to generate multi-head attention weights over the local receptive fields, A_{fg} and A_{bg}. To markedly reduce background interference, A_{fg} and A_{bg} are concatenated and the result is fused via a linear projection:

$$A_{fusion} = \text{Linear}_{2H\rightarrow H}\left(A_{fg} \oplus A_{bg}\right) \tag{12}$$

The linear layer learns the interaction between foreground and background attentions, more effectively representing regional importance differences and contrast to form a discriminative hybrid attention map, which facilitates more precise and clearer identification of target features. An activation function

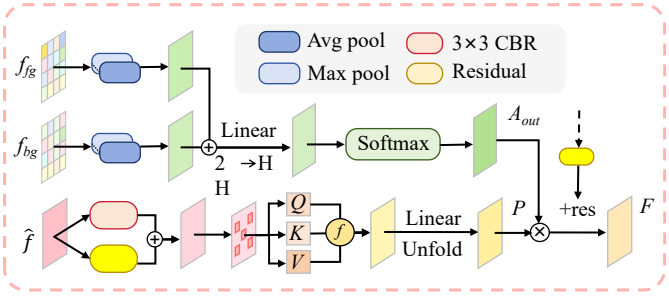

Fig. 3. The structure of CALA.

is then applied to produce A_{out}. Finally, A_{out} and P are fused via weighted multiplication:

$$x_{weight} = A_{out}P \in \mathbb{R}^{B\times H\times W\times C} \tag{13}$$

After residual correction, the output F_i is obtained for subsequent processing.

D. Dynamically Guided Decoder (DG-Decoder)

In conventional decoders, the upsampled features and the skip features are typically concatenated and treated with equal weight, which often leads to semantic conflicts and the loss of fine details. To address this, we design a decoder with a dynamic weight-learning mechanism: a 1×1 convolution compresses channels, followed by a sigmoid activation to produce a per-pixel adaptive weight $w \in [0,1]$ for the upsampled feature and a complementary weight 1 - w for the skip features. This improves the fusion of high-resolution and low-resolution features while preserving critical details, thereby enhancing the effectiveness of feature fusion.

To further strengthen discriminability, we jointly exploit the channel features and spatial features. A channel-aware mechanism first filters irrelevant channels, and the filtered channel feature is then mapped via a 1×1 convolution to a spatial guidance signal that guides spatial perception. This cascaded channel-selection-spatial-localization strategy endows the decoder with strong fusion and discrimination capabilities.

E. Loss Function

We use the BCE-Dice Loss to calculate the supervised loss of the predictions, and integrate the loss functions of all components to obtain the following loss function:

$$\mathcal{L}_{Total} = \mathcal{L}_{mask} + \mathcal{L}_{FBAD}. \tag{14}$$

IV. EXPERIMENTS

A. Experimental Settings

a) Datasets: We conduct experiments on three public benchmarks for medical image segmentation. The details of each dataset are as follows: CVC-ClinicDB [16] and Kvasir-SEG [17] are endoscopic colon polyp segmentation datasets, with 612 and 1,000 images, and training/test splits of 490/122 and 880/120, respectively.GlaS [18] is a whole-slide imaging (WSI) dataset for colorectal glands, consisting of 165 images with a training/test splits of 85/80.

TABLE I
COMPARISON WITH OTHER METHODS ON THE KVASIR-SEG, CVC-CLINICDB, AND GLAS DATASETS.

Network	CVC-ClinicDB				Kvasir-SEG				Glas			
	IoU	DSC	Pre.	Rec.	IoU	DSC	Pre.	Rec.	IoU	DSC	Pre.	Rec.
U-Net	74.99	83.64	89.52	82.44	73.62	82.86	90.25	80.55	75.80	85.50	82.80	90.30
UNet++	77.79	86.13	90.41	85.21	76.11	84.42	87.48	85.67	77.60	86.90	85.50	89.60
ResNet	79.57	87.58	89.87	87.67	80.65	88.24	92.95	85.84	74.70	84.95	81.75	90.25
PraNet	85.30	89.50	89.90	91.20	83.80	89.60	92.10	89.40	71.80	83.00	78.00	90.90
ConvSegNet	86.50	91.77	90.48	**95.18**	79.36	86.18	86.92	91.24	80.60	88.75	84.99	93.93
ECTransNet	87.80	92.30	93.30	93.10	83.90	90.10	93.20	89.00	84.21	91.02	90.41	92.59
ConDSeg	87.98	92.63	**93.70**	93.75	84.60	90.50	92.30	**91.70**	85.10	91.60	93.50	90.50
Ours	**88.51**	**93.55**	93.47	94.20	**85.74**	**91.05**	**93.81**	91.37	**85.67**	**91.95**	**92.07**	**92.50**

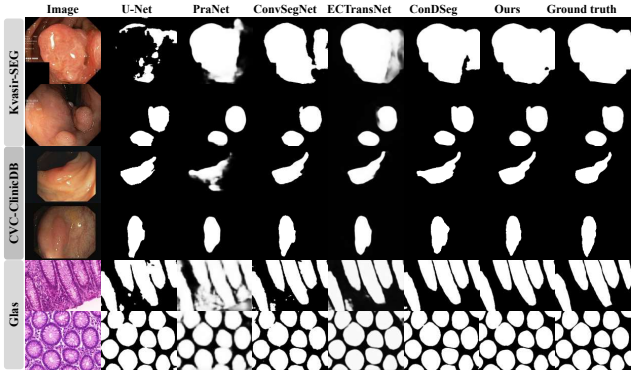

Fig. 4. Visualisation of segmentation results of different methods on datasets with different modalities.

b) Implementation Details: We implement the proposed model in PyTorch and train it on an NVIDIA Tesla T4 GPU. During training, all input images are uniformly resized to 256×256, consistent with the resolution adopted by most existing methods in the field to ensure a fair comparison and align with common practice in polyp segmentation tasks. We employ the Adam optimizer with an initial learning rate of $1e{-}4$, while the ResNet-50 backbone is trained with a lower learning rate of $1e{-}5$ to stabilize feature extraction. The batch size is set to 8, and training proceeds for 200 epochs under these configurations before termination.

c) Metrics: To objectively evaluate the performance, we report four widely-used metrics: IoU, DSC, Recall, and Precision [14]. Among these, IoU is adopted as the primary metric due to its sensitivity to boundary errors, which is crucial for precisely segmenting polyps with ambiguous margins.

B. Comparison with Other Methods

We compare our method with representative medical image segmentation approaches in terms of segmentation accuracy, generalization ability, and computational efficiency.

Firstly, we compare our method quantitatively and qualitatively with several established medical image segmentation models, including baseline models such as U-Net [6], U-Net++ [7], and ResNet [8], as well as state-of-the-art approaches PraNet [11], ConvSegNet [12], ECTransNet [13], and ConDSeg [14]. TABLE I summarizes results on CVC-ClinicDB, Kvasir-SEG and GlaS, our method outperforms existing models on most metrics. On CVC-ClinicDB, it achieves 88.51% IoU and 93.55% DSC, exceeding the previous best by +0.53 IoU and +0.92 DSC. On Kvasir-SEG, it achieves 85.74% IoU and 91.05% DSC, exceeding the previous best by +1.14 IoU and +0.55 DSC. On GlaS, it achieves 85.67% IoU and 91.95% DSC, exceeding the previous best by +0.57 IoU and +0.35 DSC. Consistent improvements across all datasets demonstrate that CALDGSeg effectively handles soft boundaries and suppresses background interference.

Fig. 4 shows qualitative comparisons. CALDGSeg yields sharper boundaries and fewer false positives than other methods. For ambiguous regions, it retains fine details and clear foreground–background separation, avoiding blurring and leakage. This validates the effectiveness of our foreground–background interaction modeling and dynamic feature fusion, consistent with the quantitative results in TABLE I.

TABLE II
COMPARISON OF MODEL COMPLEXITY WITH EXISTING METHODS.

Model	#Params	GFLOPs	FPS
U-Net	31.04	54.75	26.13
PraNet	32.55	6.96	53.72
ConvSegNet	15.58	136.08	9.14
ECTransNet	52.02	81.34	23.95
ConDSeg	45.55	101.57	26.63
Ours	34.83	112.01	25.70

C. Model Complexity Analysis

As shown in TABLE II, we compare the parameters, complexity and throughput of each model. Our method achieves 25.70 FPS with 34.83 M parameters, approaching state-of-the-art speed. Despite a moderate rise in complexity, CALDGSeg effectively balances accuracy and efficiency, meeting the accuracy and real-time requirements of medical image analysis.

D. Generalization Experiments

To demonstrate strong generalization of our method, we conduct two types of evaluations: cross-dataset generalization [19] (within the same task domain) and cross-modality [20] generalization. For cross-dataset generalization, we conduct evaluations in both directions: CVC-ClinicDB served as the

TABLE III
CROSS-DATASET COMPARISON WITH EXISTING METHODS.

Train / Test	Kvasir-SEG / CVC-ClinicDB		CVC-ClinicDB / Kvasir-SEG	
	IoU	DSC	IoU	DSC
U-Net	61.33	71.72	55.88	62.22
PraNet	53.01	62.24	71.94	79.35
ConvSegNet	65.03	72.31	68.43	77.84
ECTransNet	70.50	76.86	71.30	80.27
ConDSeg	71.67	78.82	73.88	81.99
Ours	**73.97**	**81.16**	**76.22**	**83.72**

TABLE IV
ABLATION EXPERIMENTS FOR THE MODEL.
* DENOTES THE STRATEGY MENTIONED IN THE ENCODER

Modules					Matrics	
Baseline	Fine-tuning*	FBAD	CALA	DG-Decoder	IoU	DSC
✓					77.30	85.10
✓	✓				82.70	88.90
✓	✓	✓			83.59	90.52
✓	✓		✓		83.30	90.10
✓	✓	✓	✓		84.60	90.58
✓	✓	✓	✓	✓	**85.74**	**91.05**

training set and Kvasir-SEG as the test set, and vice versa. As shown in TABLE III, our approach achieves the best results with substantial improvements over prior state-of-the-art methods. This suggests that the model learns genuinely task-relevant features and has significant practical value.

To verify broader applicability, we further evaluate on the GlaS dataset. As reported in TABLE I, our method also performs strongly on GlaS, with IoU exceeding the previous state-of-the-art by +0.57. Fig.4 shows effective boundary segmentation on GlaS, meaning CALDGSeg learns robust representations rather than dataset-specific cues.

E. Ablation Studies

We conduct comprehensive ablations studies to verify the effectiveness of the proposed modules. On Kvasir-SEG, we design three groups of experiments to assess the necessity of each module. Modules are added progressively while the remaining parts use simple convolutional blocks. Experimental results show that IoU and DSC are significantly improved when modules are added gradually, and reach the optimum when all modules are used. We also validate the data augmentation strategy, observing enhanced robustness under adverse conditions and additional performance improvements.

V. CONCLUSION

In this paper, we present CALDGSeg,a strongly interactive and guided framework for medical image segmentation. It introduces an FBAD module for foreground–background disentanglement, a CALA module for contrast-aware local attention weighting, and a DG-Decoder for pixel-wise adaptive fusion with cascaded channel–spatial optimization. Experiments on public datasets show significant gains over recent SOTA with strong cross-dataset and cross-modal generalization.

REFERENCES

[1] X. Hu, L. Liu, M. Xiong, and J. Lu, "Application of artificial intelligence-based magnetic resonance imaging in diagnosis of cerebral small vessel disease," *CNS Neuroscience & Therapeutics*, vol. 30, no. 7, pp. e14841, 2024.

[2] J. Ma, G. Yuan, C. Guo, and et al., "SW-UNet: a U-Net fusing sliding window transformer block with CNN for segmentation of lung nodules," *Frontiers in medicine*, vol. 10, pp. 1273441, 2023.

[3] X. Yang, X. Ye, D. Zhao, and et al., "Multi-threshold image segmentation for melanoma based on Kapur's entropy using enhanced ant colony optimization," *Frontiers in Neuroinformatics*, vol. 16, pp. 1041799, 2022.

[4] M. Xu, Q. Ma, H. Zhang, and et al., "MEF-UNet: An end-to-end ultrasound image segmentation algorithm based on multi-scale feature extraction and fusion," *Computerized Medical Imaging and Graphics*, vol. 114, pp. 102370, 2024.

[5] X. Zhan, J. Liu, H. Long, and et al., "An intelligent auxiliary framework for bone malignant tumor lesion segmentation in medical image analysis," *Diagnostics*, vol. 13, no. 2, pp. 223, 2023.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[7] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Proc. Int. Workshop Deep Learning in Medical Image Analysis*, 2018, pp. 3–11.

[8] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[9] O. Petit, N. Thome, C. Rambour, and et al., "U-net transformer: Self and cross attention for medical image segmentation," in *International Workshop on Machine Learning in Medical Imaging*, 2021, pp. 267–276.

[10] Z. Liu, Y. Lin, Y. Cao, and et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[11] D.-P. Fan, G.-P. Ji, T. Zhou, and et al., "PraNet: Parallel Reverse Attention Network for Polyp Segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020, pp. 263–273.

[12] A. O. Ige, N. K. Tomar, F. O. Aranuwa, and et al., "ConvSegNet: Automated Polyp Segmentation from Colonoscopy Using Context Feature Refinement with Multiple Convolutional Kernel Sizes," *IEEE Access*, vol. 11, pp. 16142–16155, 2023.

[13] W. Liu, Z. Li, C. Li, and H. Gao, "ECTransNet: An Automatic Polyp Segmentation Network Based on Multi-Scale Edge Complementary," *Journal of Digital Imaging*, vol. 36, no. 6, pp. 2427–2440, 2023.

[14] M. Lei, H. Wu, X. Lv, and X. Wang, "ConDSeg: A General Medical Image Segmentation Framework via Contrast-Driven Feature Enhancement," in *Proc. AAAI Conf. Artificial Intelligence*, 2025, vol. 39, pp. 4571–4579.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Proc. European Conf. Computer Vision (ECCV)*, 2016, pp. 630–645.

[16] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, and et al., "WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps from Physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[17] D. Jha, P. H. Smedsrud, M. A. Riegler, and et al., "Kvasir-SEG: A Segmented Polyp Dataset," in *Proc. Int. Conf. MultiMedia Modeling (MMM)*, 2020, pp. 451–462.

[18] K. Sirinukunwattana et al., "Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest," *Medical Image Analysis*, vol. 35, pp. 489–502, 2017.

[19] S. Graham, Q. D. Vu, S. E. A. Raza, and et al., "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Medical image analysis*, vol. 58, pp. 101563, 2019.

[20] F. Isensee, J. Petersen, A. Klein, and et al., "nnu-net: Self-adapting framework for u-net-based medical image segmentation," in *Bildverarbeitung für die Medizin 2019 (BVM 2019): Proceedings des Workshops*, 2019, p. 22.