Computational Visual Media
https://doi.org/10.26599/CVM.2025.9450510

Short Communication

PraNet-V2: Dual-supervised reverse attention for medical image segmentation

Bo-Cheng Hu[1,2], Ge-Peng Ji[3], Dian Shao[4], and Deng-Ping Fan[1,2] (✉)

Abstract Accurate medical image segmentation is essential for effective diagnosis and treatment. Previously we proposed PraNet-V1 as a means to enhance polyp segmentation, introducing a reverse attention (RA) module that utilizes background information. However, PraNet-V1 struggles with multi-class segmentation tasks. To address this limitation, we here propose PraNet-V2, which can effectively handle a broader range of tasks, including multi-class segmentation. At the core of PraNet-V2 is our dual-supervised reverse attention (DSRA) module, which incorporates explicit background supervision, independent background modeling, and semantically enriched attention fusion. Our PraNet-V2 framework exhibits strong performance on four polyp segmentation datasets. Moreover, the integration of DSRA into three state-of-the-art semantic segmentation models enables iterative refinement of foreground segmentation, yielding improvements of up to 1.36% in mean Dice score. Jittor code is available at `https://github.com/ai4colonoscopy/PraNet-V2/tree/main/binary_seg/jittor`.

Keywords medical image segmentation; semantic segmentation; reverse attention (RA); dual supervision

1 Introduction

Medical image segmentation plays a vital role in modern medical diagnosis and treatment by identifying regions of interest (such as lesions, organs, and tissues) in medical images. As modern medical science increasingly relies on imaging technology, segmentation tasks for medical images have evolved from binary classification to more complex multi-class segmentation. Despite advances in feature extraction and attention mechanisms, previous models still *neglect background features*. This oversight limits their ability to define boundaries accurately and diminishes performance in scenarios with low contrast between foreground and background. Our previous work, PraNet-V1 [10], introduced reverse attention (RA) to explicitly model the background, enabling effective polyp segmentation in the presence of minimal contrast [9] and significant class imbalance.

Although PraNet-V1 pioneered RA in medical image segmentation, further evaluation has revealed several limitations, as follows. (i) *Limited application scenarios*: PraNet-V1 is tailored for binary polyp segmentation, making it inadequate for multi-class segmentation. (ii) *Rule-based direct inversion*: in PraNet-V1, RA weights are generated by directly subtracting each pixel's foreground probability from one. It does not provide additional contextual information and inherits inaccuracies from the forward attention. (iii) *Semantically ambiguous attention fusion*: PraNet-V1 combines reverse and forward attention in feature space, resulting in intertwined high-dimensional foreground and background features without clear semantic boundaries.

We tackle these challenges by retuning the RA module for multi-class segmentation. See Fig. 1. We introduce a dual-supervised reverse attention (DSRA) module, which uses individual parameters to explicitly learn foreground and background attention for each class. Additionally, the DSRA module fuses

1 Nankai Institute of Advanced Research (SHENZHENFUTIAN), Shenzhen 518045, China. E-mail: thebrandonhu@gmail.com.
2 VCIP & CS, Nankai University, Tianjin 300350, China. E-mail: dengpfan@gmail.com (✉).
3 School of Computing, Australian National University, Canberra 2601, Australia. E-mail: gepengai.ji@gmail.com.
4 Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China. E-mail: shaodian@nwpu.edu.cn.

Manuscript received: 2025-01-01; accepted: 2025-08-31

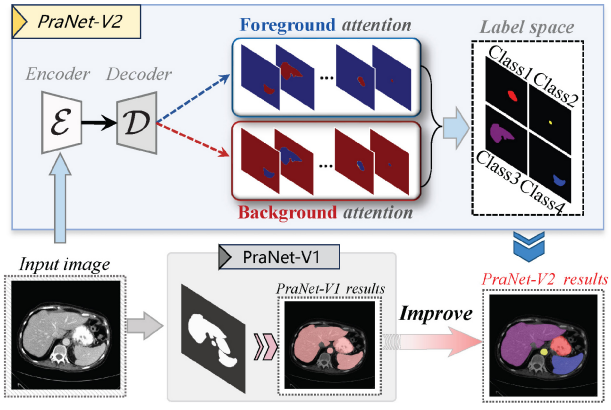

Fig. 1 Key differences between PraNet-V1 [10] and PraNet-V2in background modeling and task handling.

foreground and background information within a semantically enriched label space, thereby enhancing interpretability with respect to the traditional RA module. Overall, the main contributions of our work are as follows:

- a *modular framework:* we propose a DSRA module, which separates foreground and background processing to enhance edge and background perception, and integrate it into the PraNet-V2 framework to enable effective multi-class segmentation, and
- *performance enhancement:* we integrating DSRA into three leading segmentation models (Cascaded MERIT [30], MIST [32], and EMCAD [31]), showing that it yields consistent improvements of 0.50%–1.36% in mean Dice score.

2 Method

2.1 Dual supervised reverse attention module

In polyp segmentation, PraNet-V1 effectively employs RA for binary tasks by capturing background information. However, for multi-class segmentation, a single RA calculation for all foreground objects fails to differentiate between classes, and its rule-based RA extraction is incompatible with multi-channel pixel-level confidence outputs.

Therefore, we introduce a DSRA module, building upon the RA module from PraNet-V1. As Fig. 2 shows, we organize DSRA modules as decoder stages in a U-Net structure, leveraging the advantages of multi-scale features and skip connections. For an input image $X \in \mathbb{R}^{H\times W\times C}$, the encoder generates four multi-scale features $\{F_i \,|\, i = 1,\ldots,4\}$, where $F_i \in \mathbb{R}^{(H/2^{i+1})\times(W/2^{i+1})\times C_i}$ denotes the feature at the i-th encoder stage. The decoder then processes three high-level features $\{F_i \,|\, i = 2,3,4\}$ to generate four segmentation outputs $\{R_i \,|\, i = 1,\ldots,4\}$ across four stages, comprising a parallel partial decoder (PD) [10] followed by DSRA3 to DSRA1. Each segmentation output includes results for each semantic class (R_i^{F}) and their corresponding background regions (R_i^{B}). We modify the output layers of the PD and use it in the first decoder stage to aggregate high-level features (F_2, F_3, F_4), providing the coarse segmentation result $R_4 = \{R_4^{\mathrm{F}}, R_4^{\mathrm{B}}\}$.

Note that the subsequent three stages adopt DSRA modules to transform coarse predictions into fine results. As shown on the right side of Fig. 2, the output of the i-th DSRA module ($i = 1,2,3$), $R_i = \{R_i^{\mathrm{F}}, R_i^{\mathrm{B}}\}$, is computed as Eqs. (1) and (2):

$$R_i^{\mathrm{F}} = P_i^{\mathrm{f}} + P_i^{\mathrm{f}} \circ \gamma \tag{1}$$

$$R_i^{\mathrm{B}} = P_i^{\mathrm{b}} \tag{2}$$

Here, P_i^{f} and P_i^{b} denote the outputs of the segmentation heads under foreground and background supervision, respectively. The symbol $\circ$ represents element-wise multiplication. The term γ, referred to as the *reverse gain*, incorporates refinement information from a deeper DSRA (or PD) module to better leverage the cascade structure and background information. Unlike the RA module in PraNet-V1, the DSRA module employs dedicated supervision and structures to *separately* produce foreground and background segmentation results, avoiding the entanglement seen when using shared structures and parameters. The calculations for P_i^{f}, P_i^{b}, and γ are as Eqs. (3) and (4):

$$\{P_i^{\mathrm{f}}, P_i^{\mathrm{b}}\} = \phi(F_{i+1}) \tag{3}$$

$$\gamma = \mathrm{Softmax}(\mathcal{I}(R_{i+1}^{\mathrm{F}}; F_{i+1}) - \mathcal{I}(R_{i+1}^{\mathrm{B}}; F_{i+1})) \tag{4}$$

where $\phi(\cdot)$ denotes convolutional layers forming the decoder layers and segmentation heads. $\mathcal{I}(x; y)$ resizes the width and height of x to match y using bilinear interpolation. R_{i+1}^{F} and R_{i+1}^{B} respectively represent the foreground and background segmentation maps from the DSRA module in layer $i+1$. In the hierarchical cascade described by the above formulae, outputs from deeper DSRA modules are progressively integrated to refine coarse segmentation results. These modifications enable DSRA modules to use parameter-based learning to extract reverse attention and fuse it with forward attention in the segmentation

label space. This method resolves the spatial and semantic misalignment issues inherent in PraNet-V1's direct operation on compressed decoder features.

In summary, DSRA excels over RA in three key ways:

- *Independent structure*: DSRA utilizes *separate* segmentation heads for foreground and background, enhancing feature detail capture.
- *Background modeling*: DSRA adopts extra supervision to learn object backgrounds for each class by parameter fitting, enabling effective multi-class segmentation.
- *Semantic information refinement*: DSRA fuses foreground and background information in label space, fully leveraging pixel-level confidence to enhance boundary and background accuracy.

2.2 Background supervision and loss function

2.2.1 Background mask

To supervise each semantic class's background, we introduce a multi-channel background mask, where each channel corresponds to an object class. A pixel value of 1 indicates background regions, and 0 represents the object. See the *background mask* at the top of Fig. 2: the background for each semantic class is extracted from the ground truth, where white indicates 1 and black indicates 0.

2.2.2 Loss function

Leveraging the background mask, the total loss is defined as $\mathcal{L}_{\text{total}} = w_1\mathcal{L}_{\text{Dice}} + w_2\mathcal{L}_{\text{CE}} + w_3\mathcal{L}_{\text{BCE}}$, where w_1, w_2, and w_3 are weights. For foreground supervision, the Dice loss ($\mathcal{L}_{\text{Dice}}$) mitigates class imbalance and promotes accurate region-level segmentation, while the cross-entropy loss ($\mathcal{L}_{\text{CE}}$) refines pixel-wise classification at a finer granularity. For background supervision, the binary cross-entropy loss ($\mathcal{L}_{\text{BCE}}$) aligns background predictions with the background mask, enabling independent background learning for each class.

2.2.3 Implementation

Building on DSRA, we propose the PraNet-V2 framework, whose polyp segmentation performance is detailed in Section 3.1. In addition, the DSRA is highly versatile. Most mainstream segmentation networks can utilize two segmentation heads, like DSRA, to separately generate foreground and background segmentation results, and then iteratively enhance foreground segmentation as described in Eqs. (1)–(4). Leveraging this flexibility, we further evaluate DSRA's integration into three state-of-the-art models for multi-class medical image segmentation in Section 3.2.

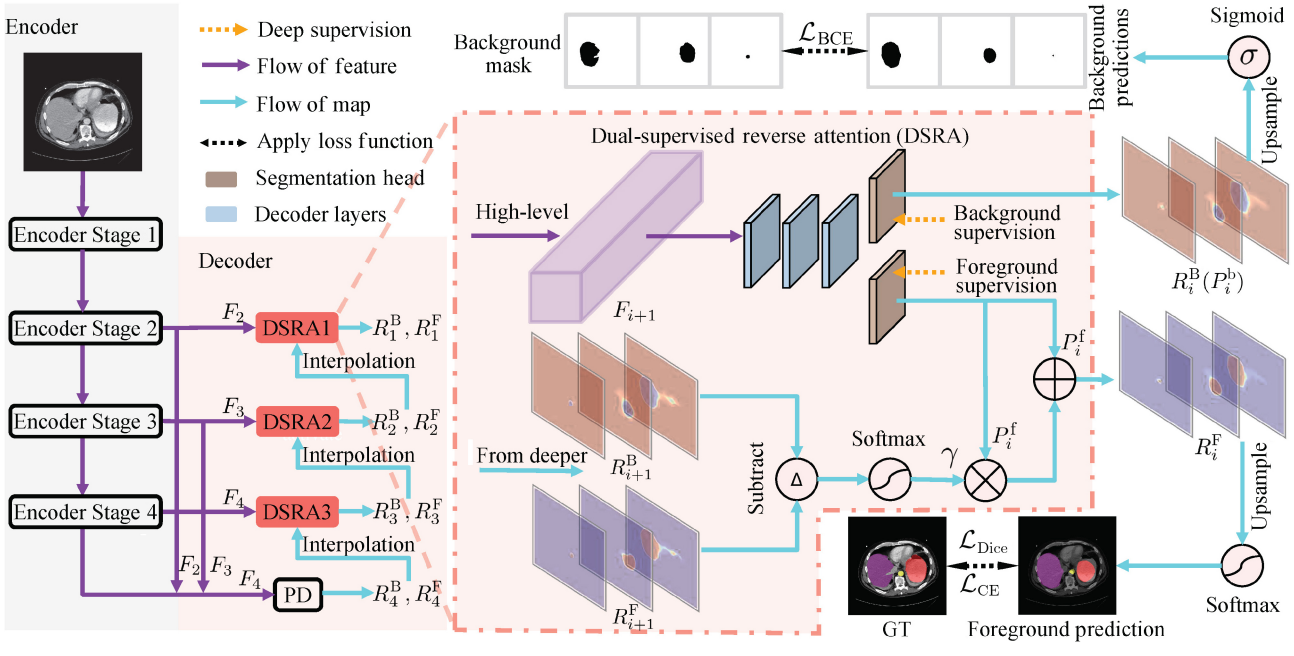

Fig. 2 Overview of the PraNet-V2 framework and DSRA module. The pipeline processes high-level features (F_2, F_3, F_4) through a parallel partial decoder (PD) and three DSRA modules. The DSRA module decodes high-level feature F_{i+1} (shown in purple) to generate foreground and background segmentation maps, while integrating outputs from deeper DSRA modules (shown in pale blue) or PD (R_{i+1}^{F},R_{i+1}^{B}) to refine the foreground segmentation maps.

3 Experiments

3.1 Binary segmentation

3.1.1 Datasets

We conducted experiments on four polyp segmentation datasets to compare the performance of PraNet-V1 and PraNet-V2, following the same split as PraNet-V1 [10]. Specifically, the datasets include CVC-ClinicDB [1], CVC-300 [37], Kvasir [18], and ETIS [35]. The CVC-ClinicDB dataset contains 612 images, of which 62 are used for testing and the rest for training. The ETIS dataset includes 196 challenging images, containing polyps that are typically small and difficult to detect. The Kvasir dataset contains 1000 images with 700 large, 48 small, and 323 medium polyps, randomly divided into 80% training, 10% validation, and 10% testing images. The ETIS and CVC-300 datasets were used exclusively for generalization evaluation, and their images omitted from training.

3.1.2 Training details

The experiments were conducted on a workstation equipped with a single NVIDIA GeForce RTX 3090 GPU, using PyTorch 2.0.1 and CUDA 12.2. The training protocols were consistent with those of PraNet-V1, including resizing inputs to 352×352 and employing a multi-scale training strategy $\{0.75\times, 1\times, 1.25\times\}$. The Adam optimizer was applied with a fixed learning rate of 10^{-4}.

3.1.3 Evaluation metrics

Following PraNet-V1, we evaluated segmentation performance using mean Dice (mDice, %), mean IoU (mIoU, %), weighted F-measure (wFm, %), structure measure (S-m, %) [7], mean enhanced measure (mEm, %) [7], and mean absolute error (MAE). Of these, mDice, mIoU, and MAE are widely used classic segmentation metrics, while wFm, S-m, and mEm provide more precise assessments beyond pixel-level accuracy.

3.1.4 Quantitative results

To ensure a fair comparison, we implemented PraNet-V1 and PraNet-V2 using the same backbone. As Table 1 shows, when using Res2Net50 [11] as the backbone, PraNet-V2 consistently outperformed PraNet-V1 on almost all datasets and metrics. Specifically, on the CVC-300 and CVC-ClinicDB datasets, PraNet-V2 achieved mDice improvements of 2.77% and 2.44%, and mIoU gains of 3.05% and 2.39%, respectively, showing enhanced segmentation accuracy. When employing PVTv2-B2 [44] as the backbone, PraNet-V2 showed even more pronounced performance gains. It consistently surpassed PraNet-V1 on all benchmark datasets and metrics, underscoring DSRA's robustness across diverse encoder architectures. On the *unseen* ETIS dataset, PraNet-V2 exhibited a remarkable 8.03% increase in mDice and an impressive 8.70% boost in mIoU, showing its strong generalization ability. Furthermore, PraNet-V2 showed a 5.12% improvement in S-m,

Table 1 Performance comparison of PraNet-V1 and PraNet-V2 on four polyp segmentation datasets, with best-performing values given in bold

Dataset	Backbone	mDice (%)		mIoU (%)		wFm (%)		S-m (%)		mEm (%)		MAE (10^{-2})	
		V1	V2	V1	V2	V1	V2	V1	V2	V1	V2	V1	V2
CVC-300	Res2Net50 [11]	87.06	**89.83**	79.61	**82.66**	84.32	**87.79**	92.55	**93.70**	94.97	**97.47**	0.99	**0.59**
ClinicDB		89.84	**92.28**	84.83	**87.22**	89.63	**91.97**	93.67	**94.87**	96.22	**97.38**	0.94	**0.91**
Kvasir		89.39	**90.70**	83.55	**85.29**	88.00	**89.59**	91.25	**91.70**	94.00	**95.07**	3.04	**2.35**
ETIS		62.75	**64.05**	**56.57**	56.54	60.07	**60.43**	79.33	**79.41**	**80.77**	79.74	3.07	**2.08**
CVC-300	PVTv2-B2 [44]	86.59	**89.89**	78.92	**83.11**	83.15	**88.48**	91.84	**93.96**	94.45	**97.04**	1.03	**0.73**
ClinicDB		90.96	**93.09**	85.42	**88.06**	89.90	**92.80**	94.34	**94.45**	96.49	**98.23**	1.02	**0.84**
Kvasir		87.09	**91.52**	81.31	**86.12**	84.52	**90.39**	89.33	**92.50**	92.58	**95.64**	4.19	**2.33**
ETIS		68.32	**76.35**	60.02	**68.72**	61.65	**72.96**	81.38	**86.50**	80.92	**88.26**	4.14	**1.45**

Table 2 Performance comparison on the Synapse dataset. Values improved by our DSRA are in bold. † indicates reproduced performance. Dice scores (%) for each organ are also included

Architecture	mDice (%)	HD95 (mm)	mIoU (%)	Aorta	GB	KL	KR	Liver	PC	SP	SM
MIST† [32]	81.91	14.93	73.19	86.15	71.43	83.09	76.43	96.02	68.20	89.39	84.59
MIST (*w/ DSRA*)	**83.27**	**14.11**	**73.89**	**87.54**	**75.36**	82.23	**76.53**	95.93	**71.51**	**91.66**	**85.41**
EMCAD-B2† [31]	82.71	21.74	74.65	87.24	69.56	85.23	80.88	95.59	65.88	92.62	84.64
EMCAD-B2 (*w/ DSRA*)	**83.75**	**17.77**	**74.81**	**88.69**	**72.79**	**85.41**	**82.91**	**95.82**	**68.47**	**93.09**	**85.85**

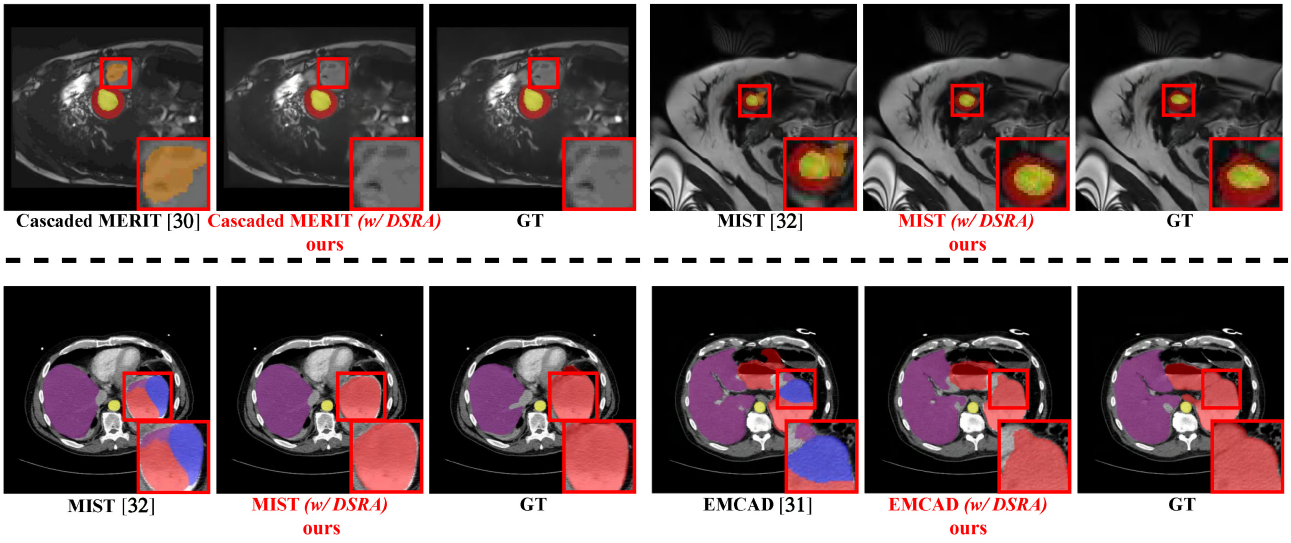

Fig. 3 Segmentation results on ACDC (above) and Synapse (below) datasets, with segmentation errors highlighted using red boxes.

indicating that our model preserves the overall shape, boundary consistency, and geometric integrity of the segmented regions.

3.2 Multi-class segmentation

3.2.1 Models and datasets

We conducted experiments on two multi-class medical image segmentation datasets to evaluate the performance of state-of-the-art models after integrating DSRA. The models used include Cascaded MERIT [30], MIST [32], and EMCAD [31]. For datasets, we used the automated cardiac diagnosis challenge (ACDC) dataset and the Synapse multi-organ segmentation (Synapse) dataset. The ACDC dataset contains cardiac MRI scans from 100 patients, annotated for three structures: right ventricle (RV), left ventricle (LV), and myocardium (Myo). The Synapse dataset comprises 30 contrast-enhanced CT scans (3779 slices) with annotations for 8 abdominal organs, including the aorta (Aorta), gallbladder (GB), left kidney (KL), right kidney (KR), liver (Liver), pancreas (PC), spleen (SP), and stomach (SM). We followed the dataset splits used by the aforementioned three models to ensure evaluation consistency.

3.2.2 Training details

The original training protocols of the three models were adopted, with minor batch size adjustments made to better accommodate the DSRA module.

3.2.3 Evaluation metrics

Following Refs. [30–32], segmentation performance was evaluated using mean Dice (mDice, %) on the ACDC and Synapse datasets. Additionally, the Synapse dataset evaluation incorporated mean IoU (mIoU, %) and 95th percentile Hausdorff distance (HD95, mm) following Ref. [31]. HD95 measures the maximum deviation between predicted and ground-truth boundaries within the 95th percentile. Lower HD95 values indicate better boundary alignment.

3.2.4 Quantitative results

On the Synapse dataset with eight categories of organs, the benefits of DSRA integration are evident (see Table 2). MIST (*w/ DSRA*) achieves an average Dice score of 83.27% and reduces HD95 to 14.11 mm, surpassing its original version by 1.36% and 0.82 mm, respectively. This improvement is further reflected in organ-level performance, where the Dice score for the GB (gallbladder) increases significantly from 71.43% to 75.36% (+3.93%). Similarly, EMCAD-B2 (*w/ DSRA*) shows robust performance gains, with the average Dice score increasing from 82.71% to 83.75% (+1.04%) and HD95 dropping from 21.74 to 17.77 mm. These results emphasize the DSRA module's ability to boost segmentation accuracy and reduce errors.

On the ACDC dataset (see Table 3), integration of the DSRA module also enhanced the performance of both MIST and Cascaded MERIT. MIST (*w/ DSRA*) and Cascaded MERIT (*w/ DSRA*) achieved average Dice scores of 92.31% and 92.28%, respectively, reflecting consistent advances over their original versions. Notably, both models show nearly a 1% improvement in Dice score for the RV (right ventricle),

a challenging class due to its irregular shape.

3.2.5 Qualitative result

Figure 3 shows segmentation results for the ACDC and Synapse datasets. It compares the performance of the three models before and after integrating the DSRA. We can see that with DSRA integration, all three models present more accurate results with fewer errors and redundancies.

3.3 Ablation study

We conducted an experiment on the MIST (*w/ DSRA*) model to assess the impact of different combinations of loss functions on segmentation performance. As Table 4 shows, the network achieves optimal performance with the combined use of $\mathcal{L}_{\text{BCE}}$, $\mathcal{L}_{\text{CE}}$, and $\mathcal{L}_{\text{Dice}}$, achieving an mDice of 92.31% and an mIoU of 86.02%. Removing any of these loss functions results in a performance drop. Specifically, excluding $\mathcal{L}_{\text{Dice}}$ impairs the network's ability to handle class-imbalanced input images, while removing $\mathcal{L}_{\text{BCE}}$ or $\mathcal{L}_{\text{CE}}$ weakens the ability of the two segmentation heads in the DSRA module to map features to the label space.

4 Discussion

Although integrating DSRA enhances the performance of three state-of-the-art segmentation models, they remain constrained by the fixed classes of the training dataset, restricting their ability to handle unknown categories in open-world scenarios. Inspired by how medical practitioners rely on expert consensus for unclassified diseases, future work could explore anomaly detection and incorporate approaches such as dual-decoder architectures with feature space manipulation to identify and adapt to unknown categories [36].

5 Conclusions

Table 3 Comparison on the ACDC dataset, with Dice scores (%) for each class. Values improved by our DSRA are in bold. † denotes reproduced performance

Architecture	mDice (%)	RV	Myo	LV
MIST† [32]	91.73	89.98	89.39	95.84
MIST (*w/ DSRA*)	**92.31**	**90.82**	**90.07**	**96.04**
Cascaded MERIT† [30]	91.78	90.36	89.21	95.79
Cascaded MERIT (*w/ DSRA*)	**92.28**	**91.27**	**89.38**	**96.19**

Table 4 Ablation on loss function

Loss function			mDice (%)	mIoU (%)
$\mathcal{L}_{\text{BCE}}$	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{Dice}}$		
✓	✓		91.91	85.43
✓		✓	92.14	85.72
	✓	✓	92.16	85.84
✓	✓	✓	**92.31**	**86.02**

In this paper, we have addressed the limitations in PraNet-V1 by introducing a DSRA module, resulting in the improved PraNet-V2 framework. By decoupling foreground and background feature computations, DSRA enhances feature separation and significantly refines segmentation accuracy in the polyp segmentation task. Furthermore, the versatility of DSRA is demonstrated through its integration into three state-of-the-art segmentation models. This integration achieves mean Dice score improvements ranging from 0.50% to 1.36% across two benchmark datasets. Such results underscore the module's broad applicability and effectiveness.

Appendix A Related work

As modern medical science increasingly relies on imaging technology, segmentation tasks in medical images have progressed from binary classification [3, 10, 12, 19, 20, 22, 34, 43] to more complex multi-class segmentation [5, 41, 46, 53]. For example, the U-Net series [8, 15, 33, 52] utilizes an encoder–decoder architecture with various skip connections to capture multi-scale information, providing refined semantic and spatial features for segmentation. Building upon various U-Net models, the nnU-Net [17] emphasizes versatility over architectural innovation. In contrast, the DeepLab series [6] expands receptive fields through dilated convolutions but remains constrained by the limitations of CNNs in holistic modeling. Furthermore, Transformer-based models [4, 5, 16, 40] improve segmentation by combining local features with global context, effectively managing long-range dependencies. Recently, many studies have explored multi-organ and multi-lesion segmentation due to its clinical importance. Some adopt SAM variants [23, 25, 38, 39, 45, 50], while others enhance model representations using geometric priors inherent to medical images [47, 49]. Inspired by other domains, recent works also introduce diffusion models [14, 21,

28], MoE frameworks [38, 48], and novel activation functions [2] for organ and lesion segmentation.

Despite advances in feature extraction and attention mechanisms, these models still *neglect background features*. This oversight limits their ability to define boundaries accurately and diminishes performance in scenarios with low contrast between foreground and background. Our previous work, PraNet-V1 [10], introduced RA to explicitly model the background, enabling effective polyp segmentation under low contrast [9] and significant class imbalance. Since then, multiple subsequent studies have adopted the core design principle of *background suppression and foreground-edge enhancement* introduced by PraNet-V1. Some approaches focus on data augmentation to enrich boundary and background supervision [12, 24], while others employ parallel-branch architectures to separately model the foreground, background and transitional regions [24, 26, 34].

Appendix B SAM-based comparison

We fine-tuned two promptable models, MedSAM [25] and BiomedParse [50], based on their publicly available pretrained weights, for comparison to our model. Unlike our method, MedSAM depends on bounding box input to define the segmentation region. During testing, following the published protocol, we simulated this requirement by expanding the ground truth bounding box outward by a random offset. Specifically, we evaluated two variants: MedSAM-2%, which corresponds to the default setting in the published implementation, where the margin varies from 0 to 20 pixels (approximately 2% of a 1024×1024 image), and MedSAM-8%, with a margin up to 80 pixels. As Table 5 shows, PraNet-V2 consistently outperforms BiomedParse across both datasets and all metrics. While MedSAM-2% achieves higher scores, it benefits from a bounding box input that roughly localizes the lesion. In contrast, our model operates entirely without ground truth-derived prompts and is capable of handling multi-class semantic segmentation tasks that the two promptable models cannot address. When the bounding box is less precise (MedSAM-8%), PraNet-V2 surpasses MedSAM on all metrics, highlighting our model's robustness to localization uncertainty and boundary ambiguity.

Table 5 Comparison of PraNet-V2 and two SAM-based promptable segmentation models on four polyp segmentation datasets. The best results are highlighted in bold. The percentage after MedSAM denotes the bounding box offset from ground truth as a fraction of image size during testing

Dataset	Model	mDice (%)	mIoU (%)	S-m (%)
CVC-300	PraNet-V2	**89.89**	**83.11**	**93.96**
ClinicDB		**93.09**	**88.06**	**94.45**
CVC-300	BiomedParse [50]	88.57	82.47	86.60
ClinicDB		91.13	86.34	90.25
CVC-300	MedSAM-8% [25]	87.65	79.46	91.56
ClinicDB		90.80	84.29	92.48
CVC-300	MedSAM-2% [25]	93.76	88.73	95.59
ClinicDB		94.97	90.82	95.58

Appendix C Multi-organ segmentation

For completeness, we report the full quantitative results of multi-organ segmentation on the Synapse and ACDC datasets in Tables 6 and 7. While the main paper highlights representative comparisons, here we provide the complete results covering all baseline methods, along with the improvements brought by our DSRA module.

Appendix D Lesion segmentation study

We conducted experiments on the kidney tumor segmentation challenge dataset (KiTS19) [13], consisting of 210 CT volumes with expert annotations for kidney and tumor regions. Following standard practices in prior works such as MT-UNet [40], we preprocessed the dataset by sampling axial slices that simultaneously contain both kidney and tumor structures, thereby ensuring clinically relevant supervision signals. To establish a reliable evaluation protocol, we randomly split the 210 cases into 168 for training and 42 for testing.

As Table 8 shows, the integration of the DSRA module significantly improved lesion-wise segmentation performance. EMCAD-B2 (*w/ DSRA*) achieved a substantial gain of 10.61% in mDice and a 21.22 mm reduction in HD95, indicating more precise and stable boundary predictions. MIST (*w/ DSRA*) also benefited from DSRA, with moderate improvements.

Appendix E Ablation study analysis

Table 4 shows the performance gain from $\mathcal{L}_{\text{BCE}}$ is the smallest among the three loss functions, as

Table 6 Complete segmentation results on the Synapse dataset, supplementing representative methods presented in Table 2 of the main paper

Architecture	mDice (%)	HD95 (mm)	mIoU (%)	Aorta	GB	KL	KR	Liver	PC	SP	SM
UNet [33]	70.11	44.69	59.39	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
AttnUNet [27]	71.70	34.47	61.38	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
R50+UNet [5]	74.68	36.87	—	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet [5]	75.57	36.97	—	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
SSFormer [42]	78.01	25.72	67.23	82.78	63.74	80.72	78.11	93.53	61.53	87.07	76.61
PolypPVT [3]	78.08	25.61	67.43	82.34	66.14	81.21	73.78	94.37	59.34	88.05	79.40
TransUNet [5]	77.61	26.90	67.32	86.56	60.43	80.54	78.53	94.33	58.47	87.06	75.00
SwinUNet [4]	77.58	27.32	66.88	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
MT-UNet [40]	78.59	26.59	—	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer [16]	81.96	18.20	—	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
PVT-CASCADE [29]	81.06	20.23	70.88	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
TransCASCADE [29]	82.68	17.34	73.48	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.52
MIST† [32]	81.91	14.93	73.19	86.15	71.43	83.09	76.43	96.02	68.20	89.39	84.59
MIST (*w/ DSRA*)	**83.27**	**14.11**	73.89	**87.54**	**75.36**	82.23	**76.53**	95.93	**71.51**	**91.66**	**85.41**
EMCAD-B2† [31]	82.71	21.74	74.65	87.24	69.56	85.23	80.88	95.59	65.88	92.62	84.64
EMCAD-B2 (*w/ DSRA*)	**83.75**	**17.77**	**74.81**	**88.69**	**72.79**	**85.41**	**82.91**	**95.82**	**68.47**	**93.09**	**85.85**

Table 7 Complete results on the ACDC dataset, supplementing representative methods presented in Table 3 of the main paper

Architecture	mDice (%)	RV	Myo	LV
R50+UNet [5]	87.55	87.10	80.63	94.92
R50+AttnUNet [5]	86.75	87.58	79.20	93.47
ViT+CUP [5]	81.45	81.46	70.71	92.18
R50+ViT+CUP [5]	87.57	86.07	81.88	94.75
TransUNet [5]	89.71	88.86	84.53	95.73
SwinUNet [4]	90.00	88.55	85.62	95.83
MT-UNet [40]	90.43	86.64	89.04	95.62
MISSFormer [16]	90.86	89.55	88.04	94.99
PVT-CASCADE [29]	91.46	88.90	89.97	95.50
nnUNet [17]	91.61	90.24	89.24	95.36
nnFormer [51]	91.78	90.22	89.53	95.59
MIST† [32]	91.73	89.98	89.39	95.84
MIST *(w/ DSRA)*	**92.31**	**90.82**	**90.07**	**96.04**
Cascaded MERIT† [30]	91.78	90.36	89.21	95.79
Cascaded MERIT *(w/ DSRA)*	**92.28**	**91.27**	**89.38**	**96.19**

Table 8 PComparison on the KITS19 dataset for lesion-wise segmentation; values improved by our DSRA are in bold

Architecture	mDice (%)	mIoU (%)	HD95 (mm)
EMCAD-B2 *(w/ DSRA)*	82.95	73.96	12.19
EMCAD-B2 [31]	72.35	62.66	33.41
Improvement	**10.61**	**11.30**	**Δ21.22**
MIST *(w/ DSRA)*	78.20	69.23	13.18
MIST [32]	75.48	66.54	16.05
Improvement	**2.72**	**2.69**	**Δ2.87**

it only supervises intermediate outputs, unlike the Dice loss, which directly optimizes the evaluation metric, and the CE loss, which serves as the core objective for multi-class segmentation. Nevertheless, it still provides auxiliary pixel-level supervision that complements the decoupled dual-branch structure of the DSRA module.

Appendix F Supplementary visual results

Here, we provide additional visualizations on both the Synapse and ACDC datasets, covering more organs and cardiac structures. Figure 4 shows that the DSRA consistently refines boundary delineation and reduces segmentation errors across different architectures, further validating its effectiveness.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (62476143, 62306239).

Author contributions

Bo-Cheng Hu's primary contribution was in writing the main content of the paper and conducting all experiments, including the Jittor version. Ge-Peng Ji's was in finalizing the experimental direction and in the writing of the paper. Dian Shao's was in assisting with the writing and consultation. Deng-Ping Fan's contributed in conceptualizing and leading the project, as well as in writing the paper.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

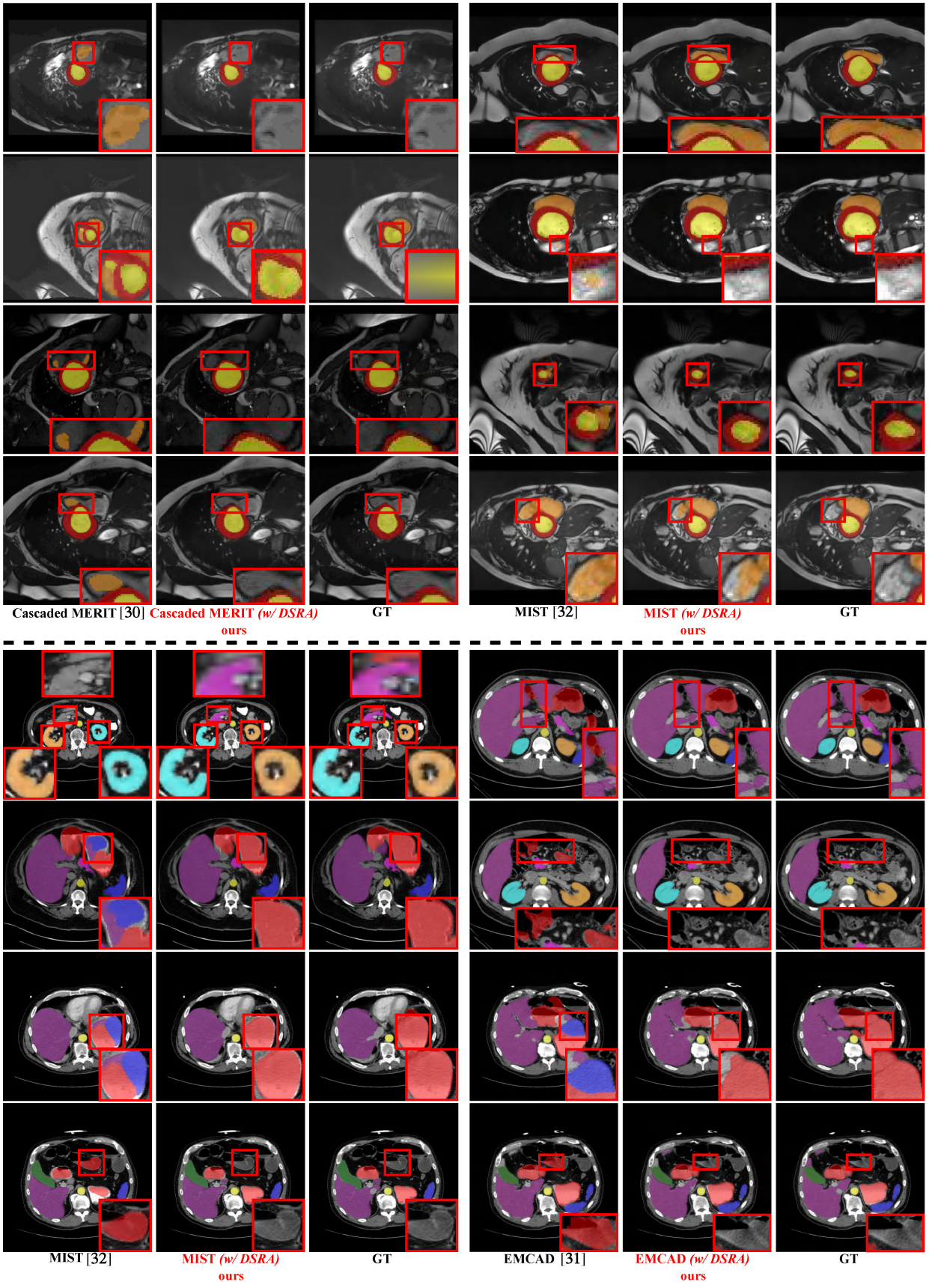

Fig. 4 Full results on ACDC (above) and Synapse (below) datasets, with segmentation errors highlighted in red boxes.

References

[1] Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* Vol. 43, 99–111, 2015.

[2] Biswas, K.; Jha, D.; Tomar, N. K.; Karri, M.; Reza, A.; Durak, G.; Medetalibeyoglu, A.; Antalek, M.; Velichko, Y.; Ladner, D.; et al. Adaptive smooth activation function for improved organ segmentation and disease diagnosis. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15009*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 65–74, 2024.

[3] Dong, B.; Wang, W.; Fan, D. P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research* Article No. 9150015, 2023.

[4] Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure Transformer for medical image segmentation. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13803*. Karlinsky, L.; Michaeli, T.; Nishino, K. Eds. Springer Cham, 205–218, 2023.

[5] Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint* arXiv:2102.04306, 2021.

[6] Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 4, 834–848, 2018.

[7] Cheng, M. M.; Fan, D. P. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision* Vol. 129, No. 9, 2622–2638, 2021.

[8] Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2016. Lecture Notes in Computer, Vol. 9901*. Ourselin, S.; Joskowicz, L.; Sabuncu, M. R.; Unal, G.; Wells, W. Eds. Springer Cham, 424–432, 2016.

[9] Fan, D. P.; Ji, G. P.; Xu, P.; Cheng, M. M.; Sakaridis, C.; Van Gool, L. Advances in deep concealed scene understanding. *Visual Intelligence* Vol. 1, No. 1, Article No. 16, 2023.

[10] Fan, D. P.; Ji, G. P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel reverse attention network for polyp segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer, Vol. 12266*. Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Eds. Springer Cham, 263–273, 2020.

[11] Gao, S. H.; Cheng, M. M.; Zhao, K.; Zhang, X. Y.; Yang, M. H.; Torr, P. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 2, 652–662, 2021.

[12] Guo, X.; Yang, C.; Liu, Y.; Yuan, Y. Learn to threshold: ThresholdNet with confidence-guided manifold mixup for polyp segmentation. *IEEE Transactions on Medical Imaging* Vol. 40, No. 4, 1134–1146, 2021.

[13] Heller, N.; Sathianathen, N.; Kalapara, A.; Walczak, E.; Moore, K.; Kaluzniak, H.; Rosenberg, J.; Blake, P.; Rengel, Z.; Oestreich, M.; et al. The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. *arXiv preprint* arXiv:1904.00445, 2019.

[14] Hu, X.; Chen, Y. J.; Ho, T. Y.; Shi, Y. Conditional diffusion models for weakly supervised medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer, Vol. 14223*. Greenspan, H.; Madabhushi, A.; Mousavi, P.; Salcudean, S.; Duncan, J.; Syeda-Mahmood, T.; Taylor, R. Eds. Springer Cham, 756–765, 2023.

[15] Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y. W.; Wu, J. UNet 3+: A full-scale connected UNet for medical image segmentation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1055–1059, 2020.

[16] Huang, X.; Deng, Z.; Li, D.; Yuan, X.; Fu, Y. MISSFormer: An effective transformer for 2D medical image segmentation. *IEEE Transactions on Medical Imaging* Vol. 42, No. 5, 1484–1494, 2023.

[17] Isensee, F.; Jaeger, P. F.; Kohl, S. A. A.; Petersen, J.; Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* Vol. 18, No. 2, 203–211, 2021.

[18] Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H. D. Kvasir-SEG: A segmented polyp dataset. In: *MultiMedia Modeling. Lecture Notes in Computer, Vol. 11962*. Ro, Y. M.; Cheng, W.-H.; Kim, J.; Chu, W.-T.; Cui, P.; Choi, J.-W.; Hu, M.-C.; De Neve, W. Eds. Springer Cham, 451–462, 2020.

[19] Ji, G. P.; Fan, D. P.; Chou, Y. C.; Dai, D.; Liniger, A.; Van Gool, L. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research* Vol. 20, No. 1, 92–108, 2023.

[20] Ji, G. P.; Xiao, G.; Chou, Y. C.; Fan, D. P.; Zhao, K.; Chen, G.; Van Gool, L. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research* Vol. 19, No. 6, 531–549, 2022.

[21] Konz, N.; Chen, Y.; Dong, H.; Mazurowski, M. A. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15007*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 88–98, 2024.

[22] Li, H.; Zhang, D.; Yao, J.; Han, L.; Li, Z.; Han, J. ASPS: Augmented segment anything model for polyp segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15009*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 118–128, 2024.

[23] Liang, J.; Cao, P.; Yang, W.; Yang, J.; Zaiane, O. R. 3D-SAutoMed: Automatic segment anything model for 3D medical image segmentation from local–global perspective. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15009*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 3–12, 2024.

[24] Liu, Z.; Zheng, S.; Sun, X.; Zhu, Z.; Zhao, Y.; Yang, X.; Zhao, Y. The devil is in the boundary: Boundary-enhanced polyp segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 34, No. 7, 5414–5423, 2024.

[25] Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nature Communications* Vol. 15, Article No. 654, 2024.

[26] Nguyen, T. C.; Nguyen, T. P.; Diep, G. H.; Tran-Dinh, A. H.; Nguyen, T. V.; Tran, M. T. CCBANet: Cascading context and balancing attention for polyp segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Lecture Notes in Computer, Vol. 12901*. de Bruijne, M.; Cattin, P. C.; Cotin, S.; Padoy, N.; Speidel, S.; Zheng, Y.; Essert, C. Eds. Springer Cham, 633–643, 2021.

[27] Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa,K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; Glocker, B.; Rueckert, D. Attention u-net: Learning where to look for the pancreas. In: Proceedings of the MIDL, 2018.

[28] Rahman, A.; Valanarasu, J. M. J.; Hacihaliloglu, I.; Patel, V. M. Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11536–11546, 2023.

[29] Rahman, M. M.; Marculescu, R. Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 6211–6220, 2023.

[30] Rahman, M. M.; Marculescu, R. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. *arXiv preprint* arXiv:2303.16892, 2023.

[31] Rahman, M. M.; Munir, M.; Marculescu, R. EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11769–11779, 2024.

[32] Rahman, M. M.; Shokouhmand, S.; Bhatt, S.; Faezipour, M. MIST: Medical image segmentation transformer with convolutional attention mixing (CAM) decoder. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 403–412, 2024.

[33] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2015. Lecture Notes in Computer, Vol. 9351*. Navab, N.; Hornegger, J.; Wells, W. M.; Frangi, A. F. Eds. Springer Cham, 234–241, 2015.

[34] Shao, H.; Zhang, Y.; Hou, Q. Polyper: Boundary sensitive polyp segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 38, No. 5, 4731–4739, 2024.

[35] Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery* Vol. 9, No. 2, 283–293, 2014.

[36] Sodano, M.; Magistri, F.; Nunes, L.; Behley, J.; Stachniss, C. Open-world semantic segmentation including class similarity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3184–3194, 2024.

[37] Vázquez, D.; Bernal, J.; Sánchez, J.; Fernández-Esparrach, G.; López, A. M.; Romero, A.; Drozdzal, M.; Courville, A. A benchmark for endoluminal scene segmentation of colonoscopy images. *arXiv preprint* arXiv:1612.00799, 2016.

[38] Wang, G.; Ye, J.; Cheng, J.; Li, T.; Chen, Z.; Cai, J.; He, J.; Zhuang, B. SAM-Med3D-MoE: Towards a

non-forgetting segment anything model via mixture of experts for 3D medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15009*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 552–561, 2024.

[39] Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; et al. SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. In: *Computer Vision – ECCV 2024. Lecture Notes in Computer Science, Vol. 15638*. Del Bue, A.; Canton, C.; Pont-Tuset, J.; Tommasi, T. Eds. Springer Cham, 51–67, 2025.

[40] Wang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.-H.; Chen, Y.-W.; Tong, R. Mixed transformer U-Net for medical image segmentation. In: Proceedings of the ICASSP, 2022.

[41] Wang, J.; Chen, J.; Chen, D.; Wu, J. LKM-UNet: Large kernel vision mamba UNet for medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15008*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 360–370, 2024.

[42] Wang, J.; Huang, Q.; Tang, F.; Meng, J.; Su, J.; Song, S. Stepwise feature fusion: Local guides global. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science, Vol. 13433*. Wang, L.; Dou, Q.; Fletcher, P. T.; Speidel, S.; Li, S. Eds. Springer Cham, 2022: 110–120.

[43] Wang, W.; Sun, H.; Wang, X. LSSNet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15007*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 446–456, 2024.

[44] Wang, W.; Xie, E.; Li, X.; Fan, D. P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVT v2: Improved baselines with Pyramid Vision Transformer. *Computational Visual Media* Vol. 8, No. 3, 415–424, 2022.

[45] Wei, X.; Cao, J.; Jin, Y.; Lu, M.; Wang, G.; Zhang, S. I-MedSAM: Implicit medical image segmentation with segment anything. In: *Computer Vision – ECCV 2024. Lecture Notes in Computer Science, Vol. 15068*. Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; Varol, G. Eds. Springer Cham, 90–107, 2025.

[46] Wu, J.; Xu, M. One-prompt to segment all medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11302–11312, 2024.

[47] Yang, X.; Chen, L.; Zheng, Y.; Ma, L.; Chen, F.; Ning, G.; Liao, H. Airway segmentation based on topological structure enhancement using multi-task learning. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15009*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 86–95, 2024.

[48] Zhang, X.; Ou, N.; Basaran, B. D.; Visentin, M.; Qiao, M.; Gu, R.; Ouyang, C.; Liu, Y.; Matthews, P. M.; Ye, C.; et al. A foundation model for brain lesion segmentation with mixture of modality experts. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15012*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 379–389, 2024.

[49] Zhao, F.; Tang, Y.; Lu, L.; Zhang, L. A curvature-guided coarse-to-fine framework for enhanced whole brain segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15012*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 13–22, 2024.

[50] Zhao, T.; Gu, Y.; Yang, J.; Usuyama, N.; Lee, H. H.; Kiblawi, S.; Naumann, T.; Gao, J.; Crabtree, A.; Abel, J.; et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods* Vol. 22, No. 1, 166–176, 2025.

[51] Zhou, H. Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; Yu, Y. nnFormer: Volumetric medical image segmentation via a 3D transformer. *IEEE Transactions on Image Processing* Vol. 32, 4036–4045, 2023.

[52] Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging* Vol. 39, No. 6, 1856–1867, 2020.

[53] Zhu, W.; Chen, X.; Qiu, P.; Farazi, M.; Sotiras, A.; Razi, A.; Wang, Y. SelfReg-UNet: Self-regularized UNet for medical image segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer, Vol. 15008*. Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Springer Cham, 601–611, 2024.

Bo-Cheng Hu is currently an undergraduate student in the School of Cybersecurity and Cyberspace Security, Nankai University, supervised by Prof. Deng-Ping Fan. He expects to graduate in 2025 and his primary research interests include medical image segmentation, multimodal large models, and AIGC.

Ge-Peng Ji commenced his Ph.D. studies at the Australian National University's School of Computing in 2022, under the supervision of Prof. Nick Barnes. He completed his master degree at Wuhan University in 2021. His research focuses on addressing complex perception challenges using computer vision and machine learning, with an emphasis on developing scalable algorithms for practical applications.

Dian Shao is an associate professor at Northwestern Polytechnical University's Unmanned System Research Institute. She is also a member of the National Key Laboratory of UAV Technology. She received her Ph.D. degree from The Multimedia Lab, Chinese University of Hong Kong, and earned her bachelor degree in intelligent science from Peking University in 2017. Her research interests include deep video analysis, AIGC, AI for science, and AI for aviation.

Deng-Ping Fan is a full professor and deputy director of the Media Computing Lab (MCLab) in the College of Computer Science, Nankai University. Before that, he was a postdoctor, working with Prof. Luc Van Gool in the Computer Vision Lab at ETH Zurich. From 2019 to 2021, he was a research scientist (PI) and team lead of IIAI-CV&Med in IIAI, working with Prof. Ling Shao and Prof. Jianbing Shen. He received his Ph.D. degree from Nankai University in 2019 under the supervision of Prof. Ming-Ming Cheng. His research interests are computer vision, machine learning, and medical image analysis. Specifically, he focuses on dichotomous image segmentation, and multi-modal AI. He is a senior member of IEEE.

To submit a manuscript, please go to `https://jcvm.org`.