

Dual-encoder cross-scale dependency fusion network for placenta segmentation using MR and ultrasound images

Sisi Song

Deyang People's Hospital

Yanfen Xin

dyxinyanfen@126.com

Deyang People's Hospital

Hong Fan

Sichuan Polytechnic University

Article

Keywords:

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11026374/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Dual-encoder cross-scale dependency fusion network for placenta segmentation using MR and ultrasound images

Sisi Song¹, Yanfen Xin^{2,*}, and Hong Fan³

¹Department of Radiology, Deyang People's Hospital, Deyang, Sichuan, 618000, China

²Department of Ultrasound, Deyang People's Hospital, Deyang, Sichuan, 618000, China

³School of Mechanical Engineering, Sichuan Polytechnic University, Deyang, Sichuan, 618000, China

*Correspondence: dyxinyanfen@126.com

ABSTRACT

Accurate placenta segmentation is essential for prenatal assessment, pregnancy management, and the diagnosis of placenta-related abnormalities. However, automatic placenta segmentation remains challenging due to complex anatomical structures, ambiguous boundaries, intensity variations, and substantial differences between imaging modalities. In this study, we propose a dual-encoder cross-scale dependency fusion network (DECSDF-Net) for placenta segmentation using MR and ultrasound images. The proposed DECSDF-Net adopts a dual-encoder architecture to extract complementary hierarchical features from two parallel encoding pathways. To enhance spatial representation, a cross-directional coordinate attention module (CDCAM) is introduced to capture directional dependencies and strengthen boundary-related features. Furthermore, a guided cross-scale dependency module (GCSDM) is designed to establish adaptive interactions among multi-level features, enabling effective information exchange between low-level spatial details and high-level semantic representations. An attention-guided enhanced receptive field block (AGERFB) is developed to improve contextual feature extraction by integrating multi-scale receptive fields with attention refinement. In addition, a progressive multi-scale feature fusion module (PMSFFM) is incorporated into the decoder to progressively aggregate features from different resolutions and enhance segmentation accuracy. To validate the effectiveness of DECSDF-Net, two placenta segmentation databases, namely Placenta-MR and Placenta-US, were constructed from magnetic resonance and ultrasound images. Compared with several state-of-the-art methods, DECSDF-Net achieved superior segmentation performance on both datasets. Specifically, the proposed method obtained Dice/MCC/Jaccard values of 0.8838/0.8775/0.7923 on Placenta-MR and 0.8778/0.8642/0.7826 on Placenta-US. Moreover, ablation studies confirmed the effectiveness of each designed module. These results demonstrate that DECSDF-Net provides accurate and robust placenta segmentation, showing promising potential for automated prenatal assessment and computer-aided diagnosis.

Introduction

The placenta is a vital temporary organ during pregnancy and plays an essential role in maintaining maternal-fetal interactions, nutrient exchange, oxygen transport, and endocrine regulation. Abnormalities in placental structure and function are closely associated with various pregnancy complications, including fetal growth restriction, preeclampsia, placental insufficiency, and adverse perinatal outcomes. With the continuous development of prenatal imaging techniques and the increasing emphasis on maternal and fetal health monitoring, accurate assessment of placental morphology and location has become increasingly important in clinical practice. However, the placenta exhibits considerable variations in shape, size, position, and tissue characteristics among different individuals, which increases the difficulty of reliable evaluation. In particular, unclear boundaries, heterogeneous intensity distributions, and interference from surrounding tissues often make manual delineation time-consuming and highly dependent on clinical expertise. At present, placental assessment mainly relies on imaging examinations such as ultrasound (US) and magnetic resonance imaging (MRI), which provide valuable information for pregnancy monitoring and clinical decision-making. Therefore, accurate and efficient placenta segmentation is of great significance for quantitative placental analysis, early identification of pregnancy-related abnormalities, and the development of computer-aided prenatal diagnosis systems.

Traditional placenta segmentation mainly relies on manual delineation by experienced radiologists or sonographers. However, manual annotation is not only labor-intensive and time-consuming but also suffers from subjective variations among observers, which limits its efficiency and reproducibility in large-scale clinical applications. To reduce the dependence on manual annotation, early studies explored conventional image processing techniques for automatic placenta segmentation, including region growing, active contour models, and atlas-based approaches. Although these methods can achieve acceptable

segmentation results in relatively simple cases, their performance is often affected by the complex characteristics of placental images. Specifically, ultrasound images are frequently degraded by speckle noise, low contrast, and unclear boundaries, while magnetic resonance images may exhibit substantial variations in placental morphology, intensity distribution, and anatomical location. Furthermore, traditional approaches generally depend on handcrafted features, predefined rules, or strong prior knowledge, making them difficult to adapt to diverse placental appearances and different imaging conditions. Therefore, more robust and adaptive segmentation methods are required to improve the accuracy and reliability of automated placenta segmentation.

With the rapid advancement of deep learning techniques, convolutional neural networks (CNNs) have become an important approach in medical image segmentation due to their powerful feature learning capabilities. Among these methods, the U-Net architecture¹ and its variants have been widely adopted in biomedical image analysis owing to their encoder-decoder design and skip connections, which enable effective integration of semantic information and spatial details from different feature levels. For placenta segmentation, several deep learning-based methods have been developed to automatically extract placental regions from ultrasound and magnetic resonance images. For example, Han et al.² proposed a U-Net and Separable U-Net-based method for automatic placenta segmentation in MRI images, demonstrating the effectiveness of deep learning in this task. Looney et al.³ developed a deep learning-based framework for fully automated 3D ultrasound placenta segmentation, enabling automatic estimation of first-trimester placental volume. Zimmer et al.⁴ introduced a multi-task learning framework for ultrasound placenta segmentation, improving segmentation performance under challenging imaging conditions. Huang et al.⁵ investigated MRI-based placenta segmentation using a 3D U-Net 3+ architecture for automatic delineation of the placenta and uterine cavity. Saito et al.⁶ proposed PlaNet-S, an ensemble framework for MRI placenta segmentation by combining U-Net and SegNeXt. Although the above methods have achieved promising results, several challenges remain unresolved. First, most existing approaches are designed for a single imaging modality, making it difficult to evaluate their adaptability across different imaging characteristics. Second, placenta images often exhibit large variations in morphology, unclear boundaries, and heterogeneous intensity distributions, which limit the ability of conventional encoder-decoder networks to accurately preserve fine structures. Third, existing methods generally lack effective mechanisms for modeling dependencies among multi-scale features, resulting in insufficient interaction between detailed spatial information and high-level semantic representations.

In recent years, advanced deep learning architectures have further promoted the development of medical image segmentation. Among them, attention-based networks have been widely explored to improve feature representation by modeling spatial and channel-wise dependencies. For example, Wang et al.⁷ introduced an auto-dynamic convolution strategy combined with location attention to enhance feature representation in medical image segmentation. Ji et al.⁸ developed a grouped multi-scale vision transformer that improves multi-scale feature representation through efficient cross-scale interactions. Chen et al.⁹ combined a hybrid CNN-transformer architecture with dual attention and cross-layer fusion to enhance feature representation and boundary localization. Liu et al.¹⁰ proposed a complementary CNN-transformer fusion strategy with boundary perception, further improving global context modeling and boundary representation. Beyond attention-based and hybrid architectures, dual-encoder and multi-branch encoder designs have recently attracted increasing attention in medical image segmentation. Compared with conventional single encoder-decoder networks, these architectures can learn complementary feature representations through parallel feature extraction pathways, thereby improving the ability to characterize complex anatomical structures. For example, Yang et al.¹¹ proposed a dual-encoder interactive fusion network that facilitates information exchange between different encoding branches to enhance feature representation in medical image segmentation. Wang et al.¹² introduced an enhanced dual-encoder framework, which improves segmentation performance by strengthening feature extraction and fusion between multiple encoder pathways. More recently, Atabansi et al.¹³ developed DCM-Net, a dual-encoder CNN-Mamba architecture with cross-branch fusion, combining local feature extraction and long-range dependency modeling for robust medical image segmentation. However, existing dual-encoder approaches mainly focus on improving feature extraction through parallel branches, while the effective modeling of cross-scale dependencies among hierarchical features remains insufficient. Moreover, maintaining accurate boundary information during feature fusion is still challenging for medical images with ambiguous structures.

To address the aforementioned challenges, we propose a dual-encoder cross-scale dependency fusion network (DECSDF-Net) for placenta segmentation using MR and ultrasound images. The proposed network employs a dual-encoder architecture to learn complementary representations and introduces cross-scale dependency modeling to enhance hierarchical feature interaction. Moreover, attention-guided feature enhancement and progressive multi-scale fusion strategies are incorporated to improve boundary preservation and feature reconstruction, leading to more accurate placenta segmentation. The main contributions of this study are summarized as follows:

- (1) A dual-encoder cross-scale dependency fusion network, namely DECSDF-Net, is proposed for automatic placenta segmentation, which effectively exploits complementary representations from different encoding pathways and improves feature learning for complex placenta structures.

- (2) A GCSDM is developed to enhance cross-scale feature interaction by adaptively modeling dependencies among

hierarchical representations, reducing the semantic gap between low-level spatial details and high-level contextual information.

(3) A CDCAM, AGERFB, and PMSFFM are introduced to improve spatial feature representation, contextual information extraction, and multi-scale feature reconstruction, respectively, resulting in more accurate boundary delineation and segmentation performance.

(4) Two placenta segmentation databases, including Placenta-MR and Placenta-US, are constructed to comprehensively evaluate the proposed method under different imaging modalities. Extensive experiments demonstrate that DECSDF-Net achieves superior segmentation performance compared with existing state-of-the-art methods.

Methods

Network architecture of DECSDF-Net

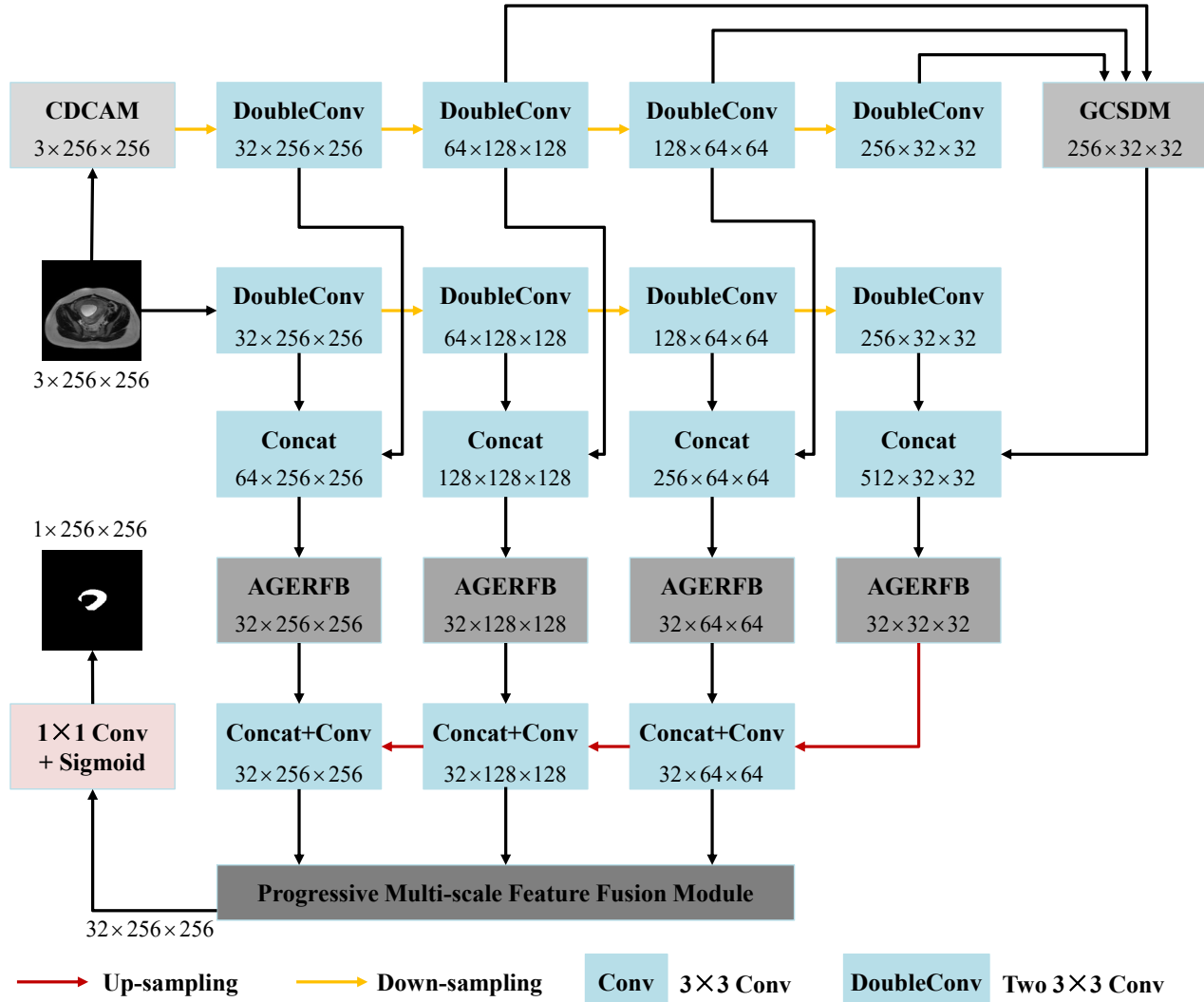


Figure 1. Network architecture of DECSDF-Net

Although U-Net is widely used for medical image segmentation, its single-encoder design and direct skip connections may limit cross-scale feature interaction and boundary preservation. Therefore, we propose DECSDF-Net for placenta segmentation using MR and ultrasound images, as shown in Fig. 1. The proposed DECSDF-Net adopts a dual-encoder and progressive-decoder architecture. Specifically, the input image is first processed by the CDCAM to enhance directional spatial information, while the original image is simultaneously fed into a parallel encoding branch. The two encoders progressively extract hierarchical features through multiple DoubleConv stages, allowing complementary representations to be learned from different feature pathways. To strengthen the interaction among hierarchical features, multi-level features from the main encoder are further introduced into the GCSDM. GCSDM aligns features from different scales and uses high-level semantic

information to guide the refinement of shallower features, thereby improving cross-scale dependency modeling. Subsequently, corresponding features from the two encoding branches are concatenated and refined through convolutional operations. The fused features at different levels are then processed by the AGERFB, which combines multi-scale receptive-field extraction with channel and spatial attention to enhance contextual representation. These refined features are progressively delivered to the decoder through skip connections. During decoding, high-level semantic features are gradually upsampled and fused with higher-resolution encoder features to recover spatial details and object boundaries. Finally, the outputs from multiple decoder stages are integrated by the PMSFFM. Through progressive alignment and aggregation of multi-resolution features, PMSFFM further enhances complementary information across decoder levels. The fused representation is then mapped to the final placenta segmentation result through a 1×1 convolution followed by a Sigmoid activation.

Structure of cross-directional coordinate attention module

Accurate placenta segmentation requires effective spatial localization, particularly for regions with ambiguous boundaries and complex surrounding tissues. Although conventional coordinate attention can encode positional information along the horizontal and vertical directions, the two directional features are generally modeled independently. To enhance the interaction between them, we design a cross-directional coordinate attention module, as illustrated in Fig. 2. Given an input feature $X \in R^{C \times H \times W}$ ($C = 3, H = W = 256$), average pooling is first performed along the width and height directions to obtain two coordinate-aware representations:

$$F_h(c, h) = \frac{1}{W} \sum_{w=1}^W X(c, h, w), \quad F_w(c, w) = \frac{1}{H} \sum_{h=1}^H X(c, h, w) \quad (1)$$

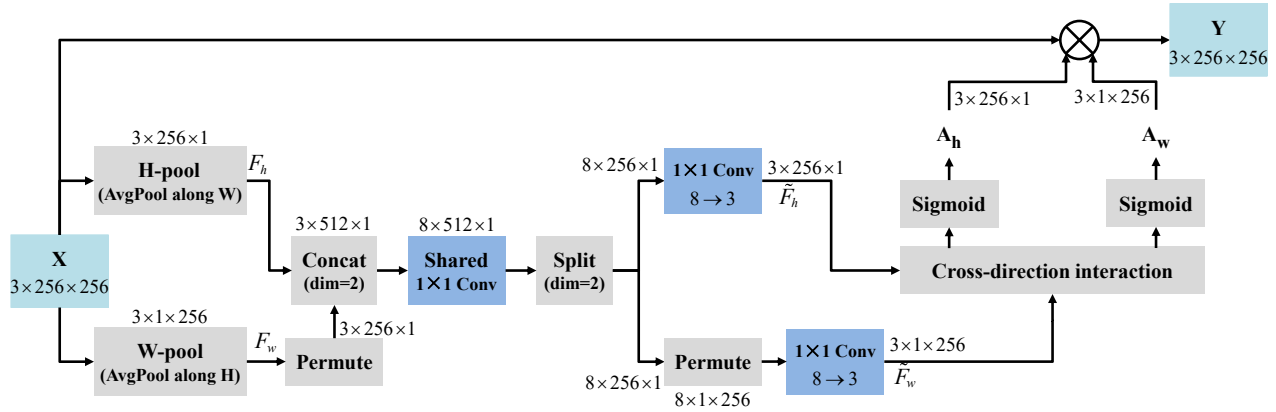


Figure 2. Structure of cross-directional coordinate attention module

The two directional features are then aligned and concatenated, followed by a shared 1×1 convolution to compress the channel dimension. The resulting representation is subsequently split into horizontal and vertical branches, and independent 1×1 convolutions are used to recover the original channel dimension, producing F_h and F_w . In the implementation, the channel dimension is compressed from 3 to 8 and then restored to 3.

To enable explicit interaction between the two spatial directions, global contextual information extracted from one branch is introduced into the other branch. The enhanced directional features are formulated as:

$$\tilde{F}_h = F_h + \tanh(\alpha)C_w, \quad \tilde{F}_w = F_w + \tanh(\beta)C_h \quad (2)$$

where $C_h = \text{Mean}(F_h)$ and $C_w = \text{Mean}(F_w)$ denote the global contextual descriptors of the two directions, while α and β are learnable coefficients controlling the interaction strength. The $\tanh(\cdot)$ function constrains the interaction weights within a bounded range, improving training stability.

Finally, the enhanced directional features are converted into attention maps by the Sigmoid function and applied to the original input:

$$Y = X \otimes A_h \otimes A_w \quad (3)$$

where $A_h = \sigma(\tilde{F}_h)$, $A_w = \sigma(\tilde{F}_w)$, and $\otimes$ denotes element-wise multiplication. Through this lightweight cross-direction interaction strategy, CDCAM preserves coordinate-aware positional information while strengthening communication between horizontal and vertical features, thereby improving the representation of placental boundaries and spatial structures.

Structure of guided cross-scale dependency module

Features extracted from different encoder stages contain complementary information, with shallow layers preserving richer spatial details and deep layers providing stronger semantic representations. However, directly concatenating these multi-scale features may introduce semantic inconsistencies and cannot explicitly model their dependencies. To address this issue, we design a guided cross-scale dependency module, as illustrated in Fig. 3. Given three hierarchical features X_1, X_2, X_3 , independent 1×1 convolutions are first employed to unify their channel dimensions. Since X_2 and X_3 are extracted from shallower encoder stages and have higher spatial resolutions, they are further down-sampled by factors of 2 and 4 to match the spatial resolution of X_1 . The aligned features are expressed as:

$$\begin{aligned} F_1 &= \text{Conv}_{1 \times 1}(X_1) \\ F_2 &= \text{Down2}(\text{Conv}_{1 \times 1}(X_2)) \\ F_3 &= \text{Down4}(\text{Conv}_{1 \times 1}(X_3)) \end{aligned} \quad (4)$$

where $\text{Conv}_{1 \times 1}$ denotes a 1×1 convolution, while Down2 and Down4 represent $2 \times$ and $4 \times$ down-sampling operations using nearest-neighbor interpolation.

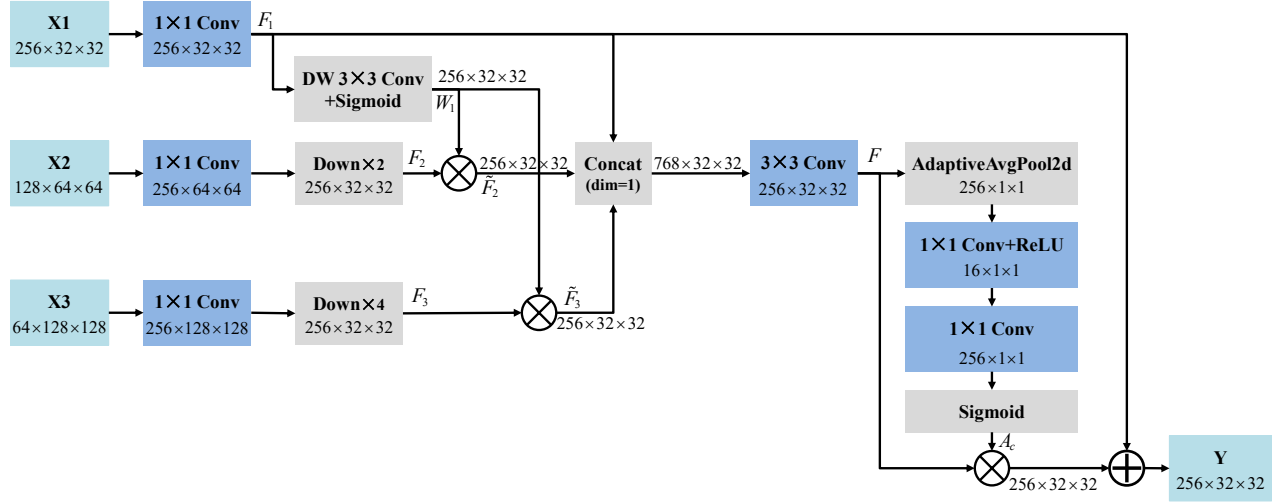


Figure 3. Structure of guided cross-scale dependency module

After feature alignment, the deepest representation F_1 is employed as the guidance feature. A lightweight depth-wise 3×3 convolution followed by Sigmoid activation is used to generate a guidance map W_1 . The generated W_1 is then applied independently to F_2 and F_3 , producing two guided features:

$$\tilde{F}_2 = F_2 \otimes W_1, \quad \tilde{F}_3 = F_3 \otimes W_1 \quad (5)$$

In this way, the high-level semantic information contained in F_1 selectively enhances informative responses in the two shallower representations while suppressing less relevant features.

Subsequently, $F_1, \tilde{F}_2$ and $\tilde{F}_3$ are concatenated along the channel dimension and further fused through a 3×3 convolution:

$$F = \text{Conv}_{3 \times 3}(\text{Concat}(F_1, \tilde{F}_2, \tilde{F}_3)) \quad (6)$$

This operation integrates semantic and spatial information from different encoder levels into a unified representation. To further refine the fused feature, a lightweight channel attention mechanism is introduced. Global average pooling is first applied to F , followed by two 1×1 convolutions with ReLU and Sigmoid activations to generate the channel attention map A_c . The refined representation is then combined with the guidance feature F_1 through a residual connection:

$$Y = F \otimes A_c + F_1 \quad (7)$$

The residual connection preserves the original high-level semantic information while facilitating stable feature propagation. Through the combination of scale alignment, semantic guidance, cross-scale fusion, channel refinement, and residual learning, GCSDM establishes explicit dependencies among hierarchical features and improves the integration of fine spatial details with high-level semantic information, thereby providing more discriminative representations for subsequent placenta segmentation.

Structure of attention-guided enhanced receptive field block

Placental regions often exhibit irregular shapes, blurred boundaries, and considerable scale variations, which make it difficult for conventional convolutional blocks to capture sufficient contextual information. To improve multi-scale feature representation and enhance discriminative responses, we design an attention-guided enhanced receptive field block, as illustrated in Fig. 4. AGERFB first employs three parallel branches to extract features with different receptive fields. The first branch uses a 1×1 convolution to preserve local information, while the other two branches combine 1×1 convolution, Group Normalization, and 3×3 convolutions with dilation rates of 1 and 2. The outputs of the three branches are concatenated and further integrated through a 3×3 convolution. Meanwhile, a residual pathway based on a 1×1 convolution is introduced to preserve the original feature information and facilitate stable feature propagation. After residual fusion, the resulting feature is simultaneously fed into the channel attention module (CAM) and spatial attention module (SAM) to generate channel-wise and spatial-wise attention maps. The original feature is then independently multiplied by the two attention maps, producing a channel-enhanced representation and a spatial-enhanced representation. The two refined features are subsequently concatenated along the channel dimension and fused through a 1×1 convolution. Through the combination of multi-branch receptive-field extraction, residual learning, and parallel channel-spatial attention refinement, AGERFB captures contextual information at different spatial scales while selectively enhancing informative channel and spatial responses. This facilitates more effective representation of complex placental structures and ambiguous boundaries for subsequent decoding and segmentation.

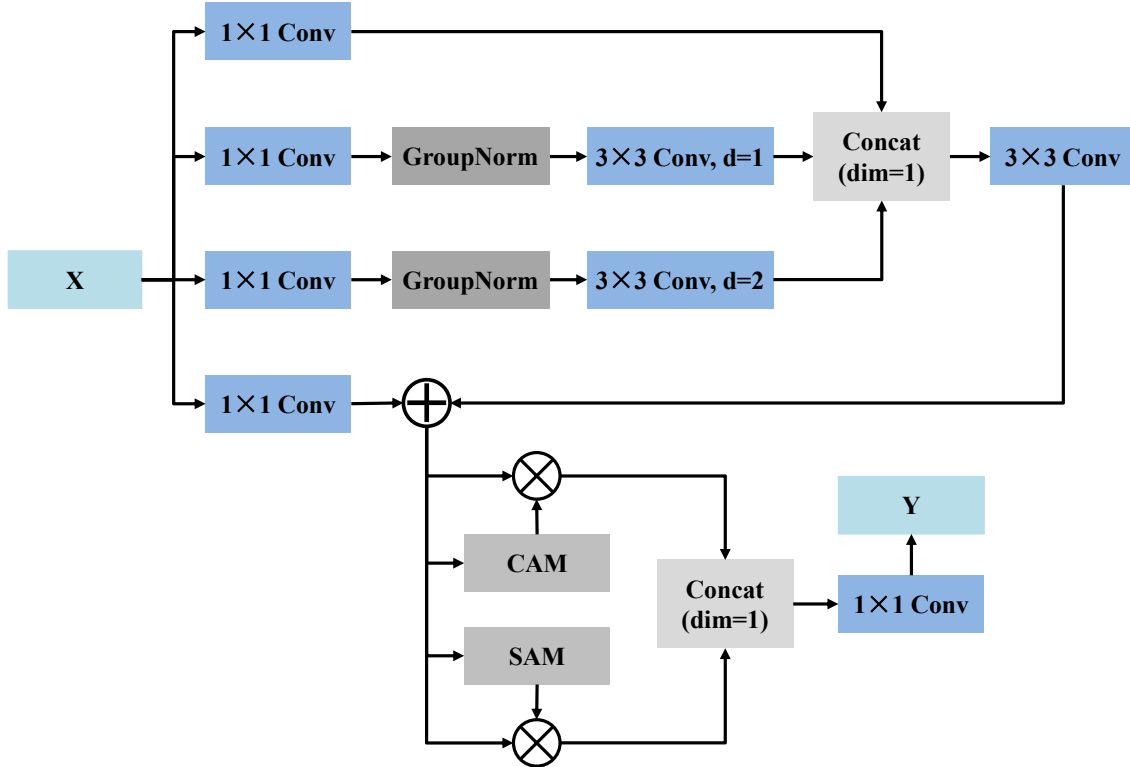


Figure 4. Structure of attention-guided enhanced receptive field block

As shown in Fig. 5, the channel attention module^{14,15} receives the feature X and employs adaptive average pooling and adaptive max pooling in parallel to aggregate complementary global channel information. The two pooled descriptors are then passed through a shared bottleneck transformation composed of two 1×1 convolutions and a ReLU activation. The first

convolution reduces the channel dimension from C_{in} to C_{in}/r , while the second restores it to C_{in} . Subsequently, the outputs of the two pooling branches are combined by element-wise addition and activated using the Sigmoid function to generate the channel attention map A_c . This attention map is multiplied with the original feature to emphasize informative channels and suppress redundant responses, thereby producing the channel-enhanced representation.

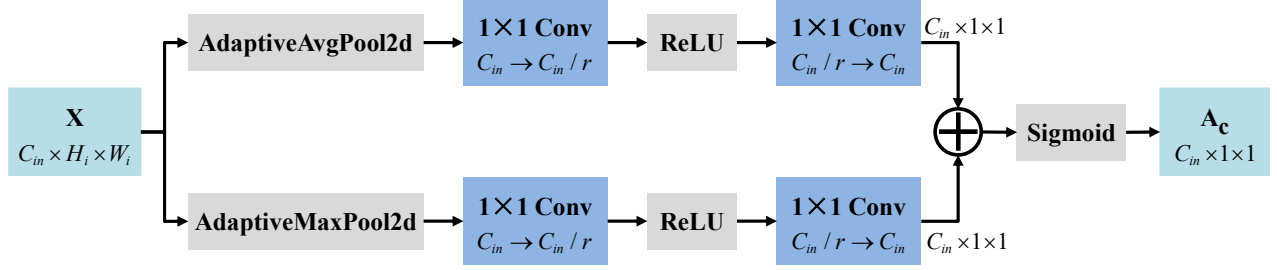


Figure 5. Structure of channel attention module

In parallel, as shown in Fig. 6, the spatial attention module^{16,17} focuses on identifying informative spatial regions. Average pooling and max pooling are first performed along the channel dimension to obtain two spatial descriptors, which are concatenated and processed by a convolution followed by a Sigmoid activation to generate the spatial attention map A_s . The original feature is then multiplied by A_s to obtain the spatial-enhanced representation. Finally, the channel-enhanced and spatial-enhanced features are concatenated and fused through a 1×1 convolution.

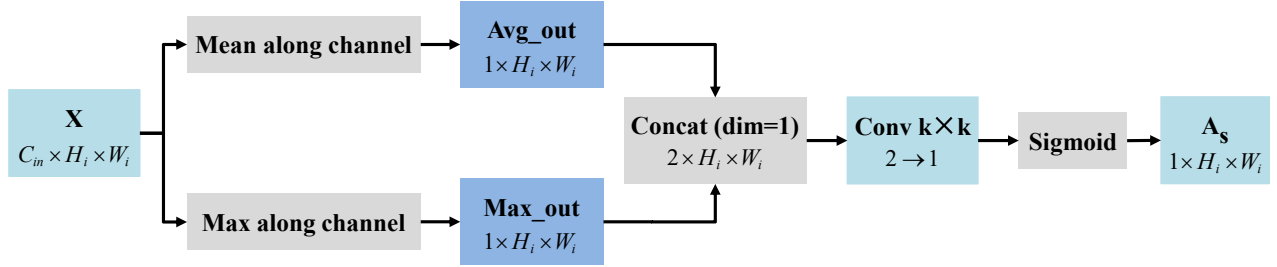


Figure 6. Structure of spatial attention module

Structure of progressive multi-scale feature fusion module

Given three decoder features X_1 , X_2 , and X_3 , PMSFFM first upsamples the lowest-resolution feature X_1 by a factor of 2 using nearest-neighbor interpolation and then applies a 3×3 convolution for feature refinement. The resulting feature is concatenated with X_2 along the channel dimension, thereby incorporating intermediate-scale spatial information. The fused feature is subsequently upsampled again by a factor of 2 and compressed through another 3×3 convolution. At the second fusion stage, the upsampled feature is concatenated with the highest-resolution feature X_3 . A final 3×3 convolution is then employed to integrate the concatenated representation and restore the channel dimension, producing the output feature Y . As illustrated in Fig. 7, the feature resolution is progressively recovered from 64×64 to 128×128 and finally to 256×256 , while the output channel number remains 32. Different from one-step multi-scale aggregation, PMSFFM adopts a progressive fusion strategy, allowing semantic information from deeper decoder layers to be gradually combined with higher-resolution spatial details. This hierarchical aggregation improves information interaction across decoder stages and facilitates the recovery of fine placental structures and boundary information, providing a more discriminative representation for the final segmentation prediction.

Loss function

During placenta segmentation, the foreground region and background often occupy markedly different proportions, which may cause the optimization process to be dominated by background pixels. To alleviate this problem, DECSDF-Net is trained using the Dice loss^{18,19}, which focuses on the degree of overlap between the predicted mask and the reference annotation rather than treating each pixel independently. The Dice coefficient is expressed as:

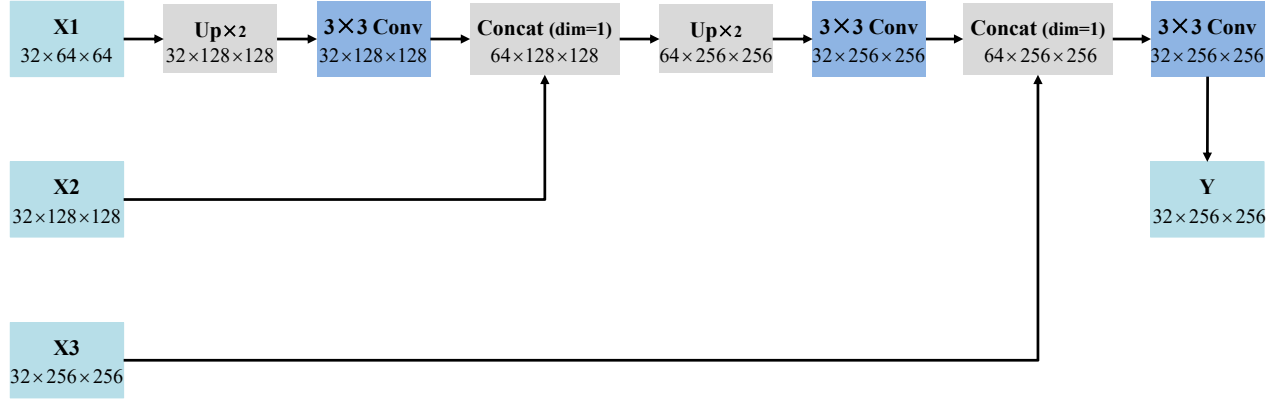


Figure 7. Structure of progressive multi-scale feature fusion module

$$L_{dice}(y, p) = 1 - \frac{2 \sum_{i=1}^N p_i y_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N p_i} \quad (8)$$

where N represents the total number of pixels, p_i and y_i denote the predicted probability and ground-truth label of the i -th pixel.

Experiments and discussion

Dataset

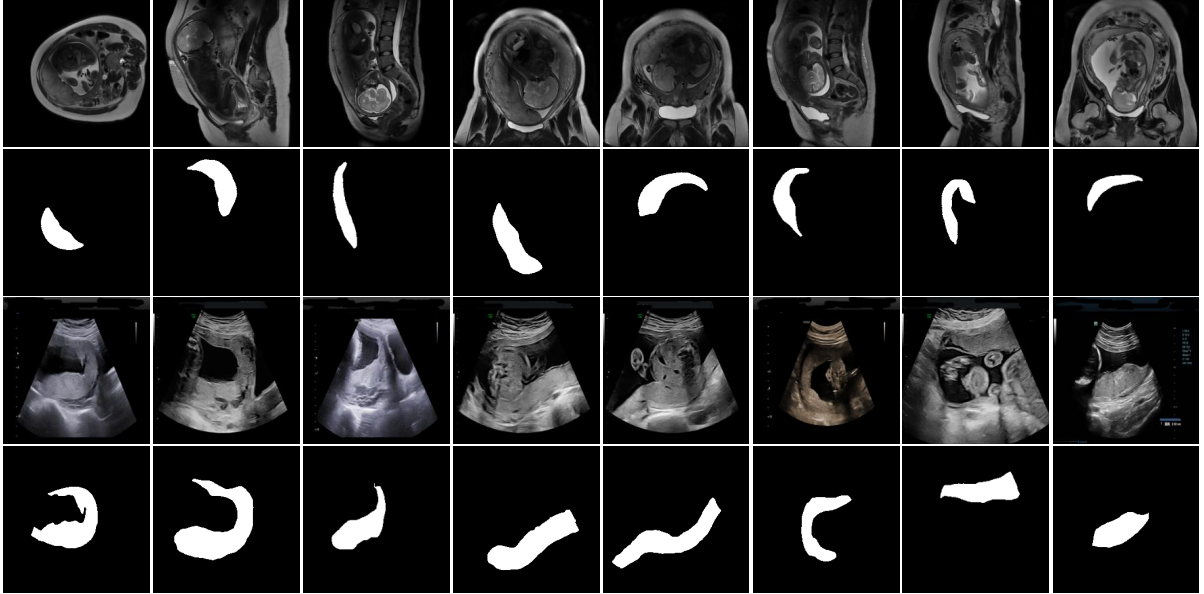


Figure 8. Representative images and corresponding annotations from the Placenta-MR and Placenta-US datasets. The first and second rows show Placenta-MR images and annotations, while the third and fourth rows show Placenta-US images and annotations.

To comprehensively evaluate the performance of the proposed DECSDF-Net, two in-house placenta segmentation datasets, namely Placenta-MR and Placenta-US, were collected from Deyang People's Hospital. The two datasets were acquired using different imaging modalities and therefore exhibit distinct visual characteristics. Placenta-MR consists of MR images, which provide relatively clear anatomical structures and soft-tissue contrast, whereas Placenta-US contains US images that present

more complex appearance variations, including speckle noise, low contrast, and indistinct tissue boundaries. All images in both datasets were manually delineated by experienced clinicians to obtain pixel-level placenta masks, which were used as reference annotations for supervised training and quantitative evaluation. To ensure independent model development and assessment, each dataset was divided into training, validation, and test subsets. Specifically, Placenta-MR contains 1,546 images, of which 928 were used for training, 309 for validation, and 309 for testing. Placenta-US contains 1,557 images, including 935 training images, 311 validation images, and 311 test images. The validation subsets were employed for monitoring model performance and selecting the optimal model during training, while the test subsets were used exclusively for the final performance evaluation. By incorporating placenta images from both MR and ultrasound examinations, the two datasets cover different imaging appearances and structural characteristics of the placenta, providing a more comprehensive basis for evaluating the effectiveness and adaptability of DECSDF-Net. Representative images and their corresponding manual annotations from Placenta-MR and Placenta-US are presented in Fig. 8.

The Placenta-MR and Placenta-US datasets from Deyang People’s Hospital are private. All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by Deyang People’s Hospital. Informed consent was obtained from all subjects and their legal guardian.

Evaluation metrics

The performance of DECSDF-Net was quantitatively evaluated using three widely adopted segmentation metrics, namely the Dice coefficient^{20,21}, Matthews correlation coefficient (MCC)^{22,23}, and Jaccard index^{24,25}.

$$Dice = \frac{2TP}{2TP + FN + FP} \quad (9)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (10)$$

$$Jaccard = \frac{TP}{TP + FN + FP} \quad (11)$$

Implementation details

All experiments were conducted in PyTorch on a workstation with an NVIDIA RTX 4090 GPU (24 GB). The proposed DECSDF-Net was optimized using Adam^{26,27} with an initial learning rate of 0.001, a mini-batch size of 16, and a maximum of 200 epochs. Prior to training, images from both Placenta-MR and Placenta-US were uniformly resized to 256×256 pixels to standardize the network input and reduce computational overhead. Figure illustrates the evolution of Dice, MCC, and Jaccard during the training of DECSDF-Net on the Placenta-MR and Placenta-US datasets. For both datasets, the three metrics increase rapidly during the early training stage, indicating that the network can quickly learn effective placental representations. As training proceeds, the growth rate gradually decreases and the curves enter a relatively stable stage. Although several local fluctuations can be observed in the validation curves during the intermediate epochs, particularly for Jaccard, these variations gradually diminish as the optimization progresses. On Placenta-MR, the validation Dice, MCC, and Jaccard curves stabilize at approximately 0.87, 0.87, and 0.78. A similar convergence pattern is observed on Placenta-US, where the corresponding validation metrics remain around 0.87, 0.85, and 0.77 after convergence. Meanwhile, the training curves continue to improve smoothly and eventually approach their upper limits. Overall, the consistent trends observed across the two imaging modalities demonstrate that DECSDF-Net exhibits stable optimization behavior and reliable convergence for placenta segmentation.

Ablation study

To investigate the contribution of each component in DECSDF-Net, a series of ablation experiments were conducted on the Placenta-MR dataset. As reported in Table 1, the dual-encoder architecture without the proposed enhancement modules was adopted as the baseline, achieving Dice, MCC, and Jaccard scores of 0.8496, 0.8425, and 0.7393. The remaining modules were then introduced progressively to examine their effects on segmentation performance. After incorporating PMSFFM, the Dice, MCC, and Jaccard values increased to 0.8548, 0.8475, and 0.7474, corresponding to improvements of 0.0052, 0.0050, and 0.0081 over the baseline. This result indicates that progressively aggregating decoder features at different resolutions facilitates the integration of semantic information and spatial details. With the further addition of CDCAM, the three metrics reached 0.8562, 0.8494, and 0.7495. Although the improvement at this stage is relatively moderate, the consistent gains suggest that cross-directional coordinate attention contributes to more effective spatial localization and feature recalibration. When GCSDM was introduced, performance improved more noticeably, with Dice, MCC, and Jaccard reaching 0.8633, 0.8562, and 0.7601.

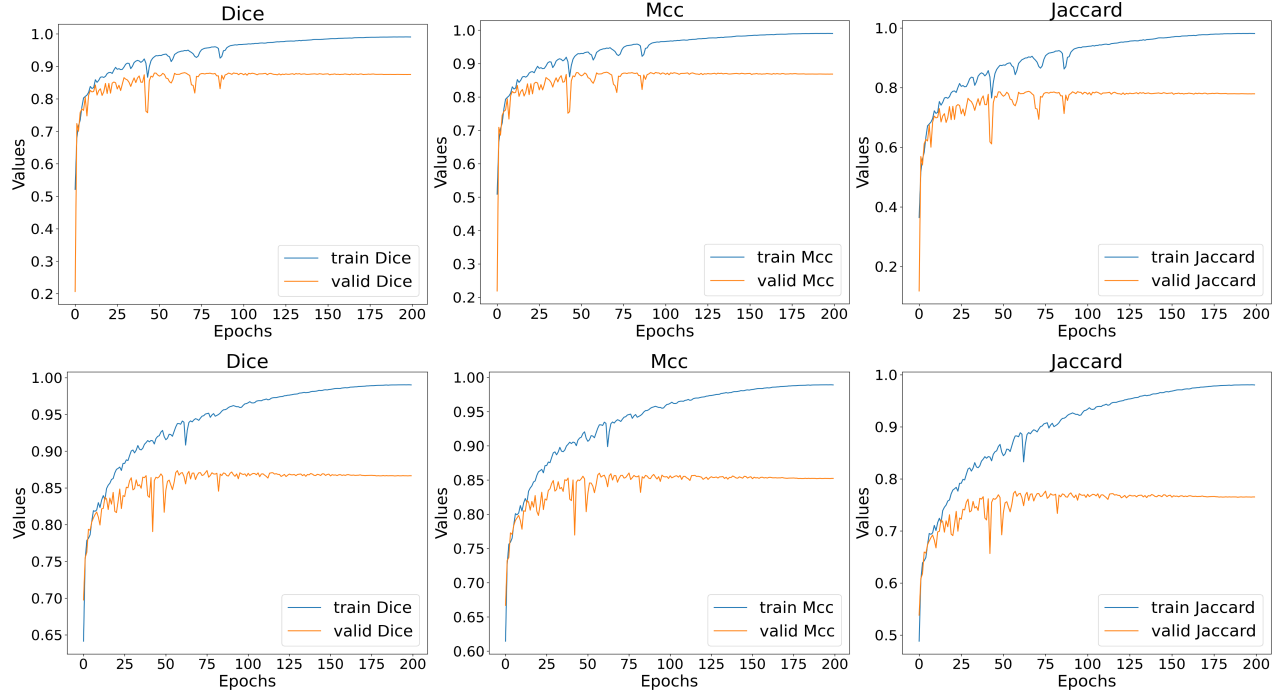


Figure 9. Dice, MCC, and Jaccard curves of DECSDF-Net on the Placenta-MR and Placenta-US datasets. The upper and lower rows represent the results for Placenta-MR and Placenta-US.

Compared with the preceding configuration, the corresponding gains were 0.0071, 0.0068, and 0.0106. This demonstrates the benefit of using deep semantic information to guide the interaction and fusion of hierarchical encoder features. Finally, adding AGERFB produced the best results, yielding a Dice score of 0.8838, an MCC of 0.8775, and a Jaccard index of 0.7923. Relative to the configuration without AGERFB, these values increased by 0.0205, 0.0213, and 0.0322, respectively, showing that multi-receptive-field extraction combined with channel and spatial attention substantially strengthens feature discrimination and boundary representation. Overall, the complete DECSDF-Net improves Dice, MCC, and Jaccard by 0.0342, 0.0350, and 0.0530, respectively, compared with the dual-encoder baseline. The progressive performance gains across the ablation settings confirm that PMSFFM, CDCAM, GCSDM, and AGERFB provide complementary contributions to placenta segmentation, while their joint integration leads to the strongest overall performance.

Method	Dice	MCC	Jaccard
Baseline (Dual-encoder)	0.8496	0.8425	0.7393
Baseline+PMSFFM	0.8548	0.8475	0.7474
Baseline+PMSFFM+CDCAM	0.8562	0.8494	0.7495
Baseline+PMSFFM+CDCAM+GCSDM	0.8633	0.8562	0.7601
Baseline+PMSFFM+CDCAM+GCSDM+AGERFB	0.8838	0.8775	0.7923

Table 1. Ablation study on the Placenta-MR dataset.

The qualitative results of the ablation experiments are illustrated in Fig. 10. The baseline dual-encoder network can roughly identify the placental region, but its predictions still contain evident false-positive areas, incomplete structures, and irregular boundaries in several challenging cases. After introducing PMSFFM, the predicted masks become more continuous and the overall placental shape is better preserved, indicating that progressive fusion of decoder features helps recover spatial details. The addition of CDCAM further improves the localization of the target region and reduces some background interference, although small isolated regions and local boundary deviations can still be observed. With the incorporation of GCSDM, the segmentation results show better structural consistency and fewer spurious responses, particularly in samples with complex placental morphology or weak contrast. This suggests that guided cross-scale interaction enables deeper semantic information to assist the refinement of features from different encoder levels. Finally, the complete model with AGERFB produces masks that are visually closest to the ground-truth annotations across most cases. The predicted placental regions are generally more

complete, the boundaries are smoother, and isolated false-positive components are further suppressed. These visual observations are consistent with the quantitative results in Table 1 and further demonstrate that PMSFFM, CDCAM, GCSDM, and AGERFB progressively improve the segmentation capability of DECSDF-Net.

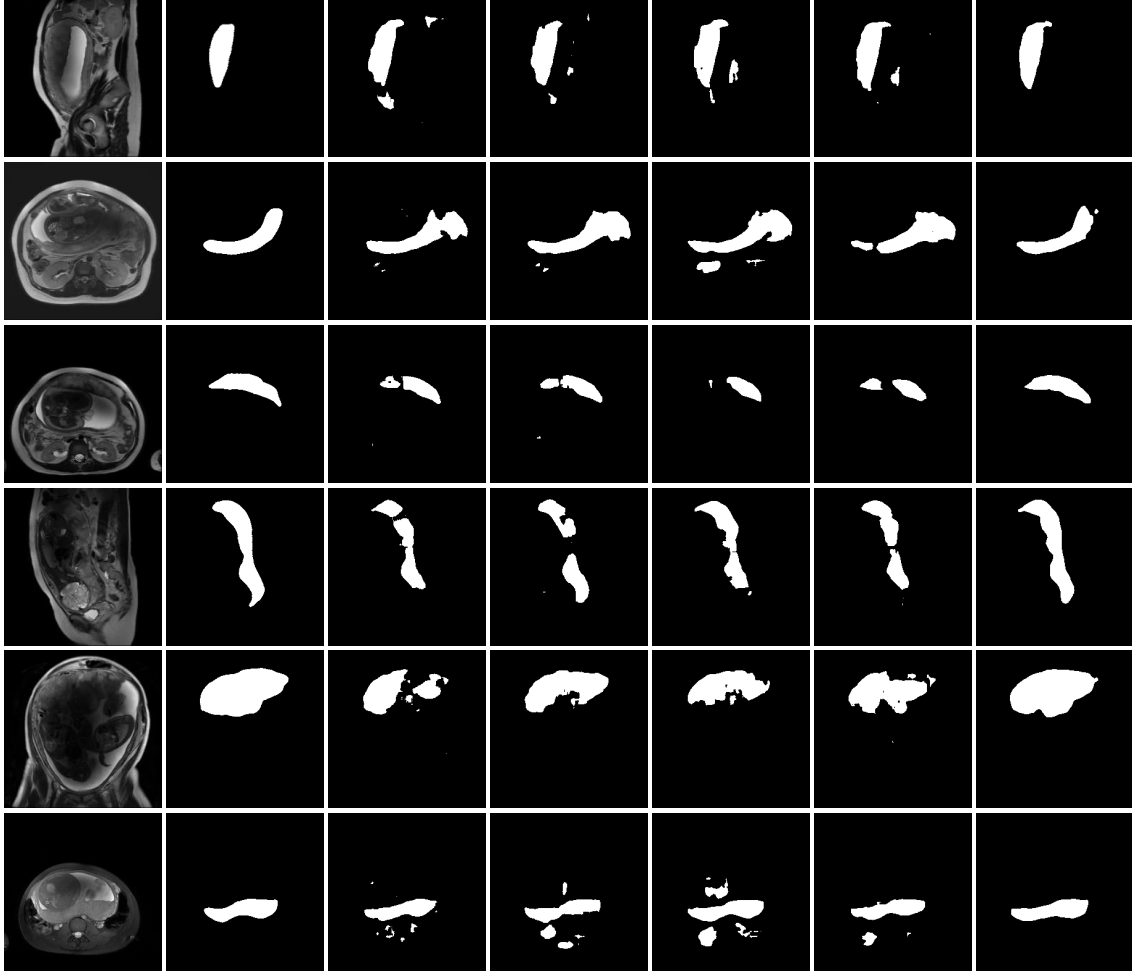


Figure 10. Visualization of ablation results on the Placenta-MR dataset. The first and second columns show the original images and ground-truth masks, while the subsequent columns present the segmentation results obtained by Baseline (Dual-encoder), Baseline+PMSFFM, Baseline+PMSFFM+CDCAM, Baseline+PMSFFM+CDCAM+GCSDM, and Baseline+PMSFFM+CDCAM+GCSDM+AGERFB.

Comparison experiments

To further assess the segmentation capability of DECSDF-Net, seven representative methods were compared on the Placenta-MR dataset, including DualA-Net²⁸, CKDNet²⁹, A-Net³⁰, BMANet³¹, LMFR-Net³², PMFSNet³³, and DSNet³⁴. As shown in Table 2, noticeable performance differences can be observed among these models. Among the existing methods, A-Net achieves the strongest overall performance, with Dice, MCC, and Jaccard values of 0.8709, 0.8640, and 0.7722. DSNet follows closely, obtaining 0.8693, 0.8624, and 0.7692. CKDNet achieves 0.8567 of Dice, 0.8493 of MCC, and 0.7501 of Jaccard, while DualA-Net obtains 0.8549, 0.8488, and 0.7475, respectively. PMFSNet reaches corresponding scores of 0.8520, 0.8443, and 0.7428. LMFR-Net [32] records 0.8409 of Dice, 0.8324 of MCC, and 0.7261 of Jaccard, whereas BMANet yields the lowest results among the evaluated methods, with values of 0.8184, 0.8086, and 0.6934. The proposed DECSDF-Net achieves the highest scores for all three metrics, reaching 0.8838 for Dice, 0.8775 for MCC, and 0.7923 for Jaccard. Compared with A-Net, the best-performing competing method, DECSDF-Net improves Dice, MCC, and Jaccard by 0.0129, 0.0135, and 0.0201. In comparison with BMANet, which obtains the lowest performance among the evaluated methods, the corresponding improvements reach 0.0654, 0.0689, and 0.0989. These results indicate that DECSDF-Net provides the most accurate overall segmentation performance on the Placenta-MR dataset and maintains a clear advantage across all three quantitative evaluation

criteria.

Method	Dice	MCC	Jaccard
DualA-Net ²⁸	0.8549	0.8488	0.7475
CKDNet ²⁹	0.8567	0.8493	0.7501
A-Net ³⁰	0.8709	0.8640	0.7722
BMANet ³¹	0.8184	0.8086	0.6934
LMFR-Net ³²	0.8409	0.8324	0.7261
PMFSNe ³³	0.8520	0.8443	0.7428
DSNet ³⁴	0.8693	0.8624	0.7692
DECSDf-Net	0.8838	0.8775	0.7923

Table 2. Comparison experiments on the Placenta-MR dataset.

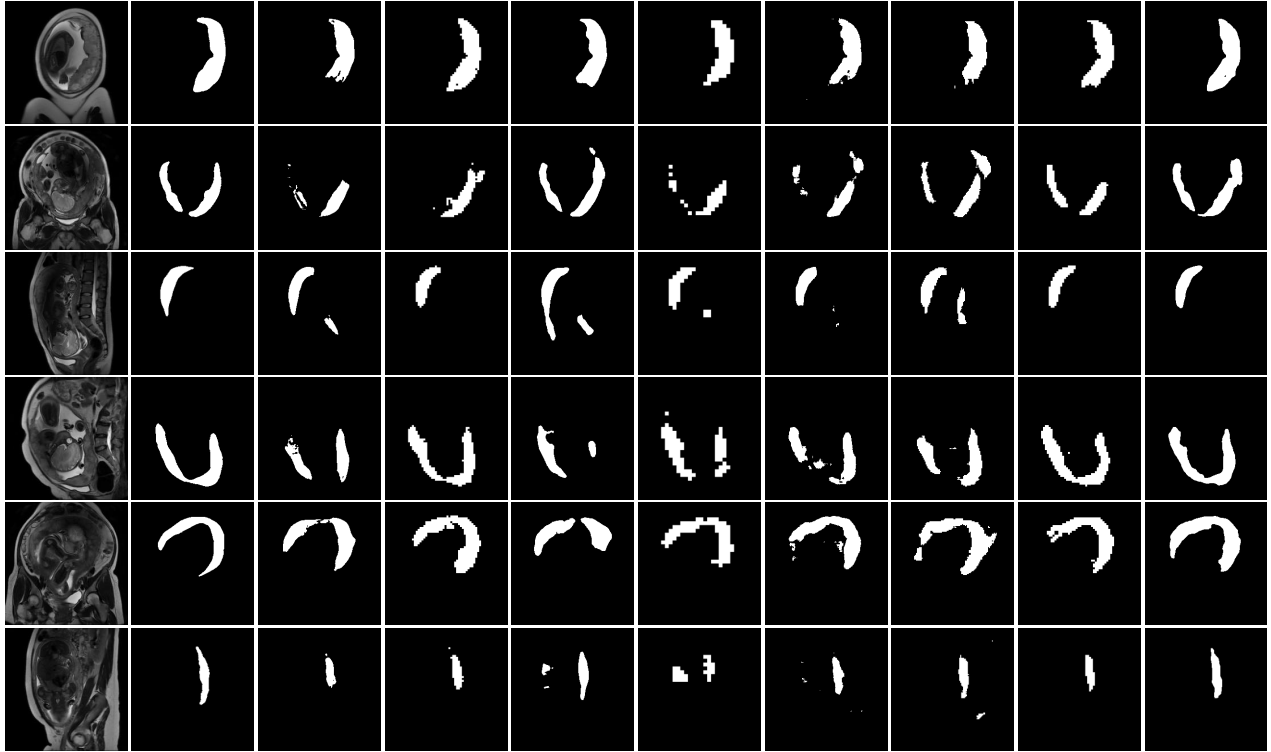


Figure 11. Visual comparison of segmentation results on the Placenta-MR dataset. The first two columns represent the original images and ground-truth masks, followed by the results of DualA-Net, CKDNet, A-Net, BMANet, LMFR-Net, PMFSNet, DSNet, and DECSDf-Net.

Figure 11 provides a qualitative comparison of different methods on the Placenta-MR dataset. DualA-Net can identify the main placental regions in most cases, but incomplete structures and local false-positive responses are still visible in several samples. CKDNet also captures the approximate placental location, although some predictions exhibit fragmented regions and noticeable boundary deviations. A-Net performs relatively well and produces more continuous masks with better agreement with the ground truth. However, small missing regions and local contour errors can still be observed in some challenging cases. BMANet shows more obvious under-segmentation and discontinuous predictions, especially when the placental morphology is complex or the boundary is unclear. LMFR-Net and PMFSNet are able to preserve the overall placental structure in most samples, but they still produce several isolated false-positive areas and local shape distortions. DSNet provides comparatively stable segmentation results and maintains good structural continuity, although slight boundary deviations remain in some cases. In contrast, the proposed DECSDf-Net generates predictions that are generally closer to the ground-truth masks. It better preserves elongated and curved placental structures, reduces isolated false-positive regions, and produces more complete and regular boundaries. These visual observations are consistent with the quantitative results in Table 2, further demonstrating the

effectiveness of DECSDF-Net for placenta segmentation in MR images.

Method	Dice	MCC	Jaccard
DualA-Net ²⁸	0.8709	0.8564	0.7716
CKDNet ²⁹	0.8584	0.8430	0.7526
A-Net ³⁰	0.8723	0.8583	0.7741
BMANet ³¹	0.8499	0.8332	0.7395
LMFR-Net ³²	0.8530	0.8367	0.7440
PMFSNe ³³	0.8597	0.8443	0.7541
DSNet ³⁴	0.8732	0.8593	0.7753
DECSDF-Net	0.8778	0.8642	0.7826

Table 3. Comparison experiments on the Placenta-US dataset.

Table 3 summarizes the quantitative comparison on the Placenta-US dataset. Compared with the results on Placenta-MR, the performance differences among several competing methods are relatively smaller, indicating that a number of models can achieve competitive segmentation accuracy on ultrasound images. Among the existing approaches, DSNet obtains the best overall performance, with Dice, MCC, and Jaccard scores of 0.8732, 0.8593, and 0.7753. A-Net follows closely, achieving 0.8723, 0.8583, and 0.7741, while DualA-Net also shows strong performance with corresponding values of 0.8709, 0.8564, and 0.7716. PMFSNet reaches Dice, MCC, and Jaccard scores of 0.8597, 0.8443, and 0.7541, whereas CKDNet obtains 0.8584, 0.8430, and 0.7526. LMFR-Net achieves 0.8530 Dice, 0.8367 MCC, and 0.7440 Jaccard. BMANet shows the lowest performance among the compared methods, with values of 0.8499, 0.8332, and 0.7395. The proposed DECSDF-Net achieves the highest values for all three metrics, reaching 0.8778 Dice, 0.8642 MCC, and 0.7826 Jaccard. Compared with DSNet, the strongest competing method, DECSDF-Net improves Dice, MCC, and Jaccard by 0.0046, 0.0049, and 0.0073. Compared with BMANet, which obtains the lowest scores, the corresponding improvements increase to 0.0279, 0.0310, and 0.0431. Although the margin over the strongest competing methods is relatively modest, DECSDF-Net consistently ranks first across all evaluation criteria, demonstrating its stable segmentation capability on the Placenta-US dataset.

Figure 12 shows the visual comparison of different methods on the Placenta-US dataset. Compared with MR images, ultrasound images generally contain stronger speckle noise, weaker tissue contrast, and less distinct placental boundaries, making accurate segmentation more challenging. DualA-Net can identify the main placental regions, but local shape distortion and incomplete predictions are observed in several samples. CKDNet shows more obvious fragmentation in some cases, particularly when the placenta has an irregular or curved appearance. A-Net provides relatively complete segmentation results and preserves the overall morphology well, although local boundary deviations remain. BMANet tends to produce relatively coarse or incomplete masks in several ultrasound samples. LMFR-Net and PMFSNet can recover most of the target area, but isolated false-positive regions and local over-segmentation are still visible, especially in images with strong background interference. DSNet exhibits relatively stable performance and generates masks with good structural consistency. Nevertheless, slight missing regions and contour inaccuracies can still be found in several challenging examples. In comparison, DECSDF-Net produces predictions that more closely match the ground-truth masks across different placental shapes. It effectively suppresses scattered false-positive regions, maintains better continuity of the placental structure, and provides more accurate boundaries in low-contrast ultrasound images. In particular, for samples affected by strong speckle noise or irregular target morphology, the proposed method shows more stable segmentation results than the competing approaches. These qualitative observations further support the quantitative results in Table 3 and demonstrate the adaptability of DECSDF-Net to the complex appearance characteristics of placenta ultrasound images.

Complexity analysis

To further evaluate the computational efficiency of different segmentation methods, GFLOPs, parameter number, and inference speed were compared on the Placenta-MR dataset. The results are summarized in Table 4. Considerable differences can be observed among the evaluated networks in terms of computational cost and model size. Among the compared methods, PMFSNet and A-Net require the fewest floating-point operations, with 0.59 and 0.60 GFLOPs. LMFR-Net contains only 0.07 M parameters and achieves the highest inference speed of 325.25 FPS, demonstrating its lightweight characteristics. DualA-Net also maintains relatively low complexity, requiring 5.47 GFLOPs and 2.58 M parameters while reaching 147.57 FPS. In contrast, CKDNet, BMANet, and DSNet contain substantially more parameters, reaching 59.34 M, 28.40 M, and 29.33 M. In particular, DSNet has the highest computational cost at 32.72 GFLOPs. The proposed DECSDF-Net requires 17.96 GFLOPs and contains 4.60 M parameters, while achieving an inference speed of 160.05 FPS. Although its computational cost is higher than that of several lightweight networks, its parameter number remains considerably lower than those of CKDNet,

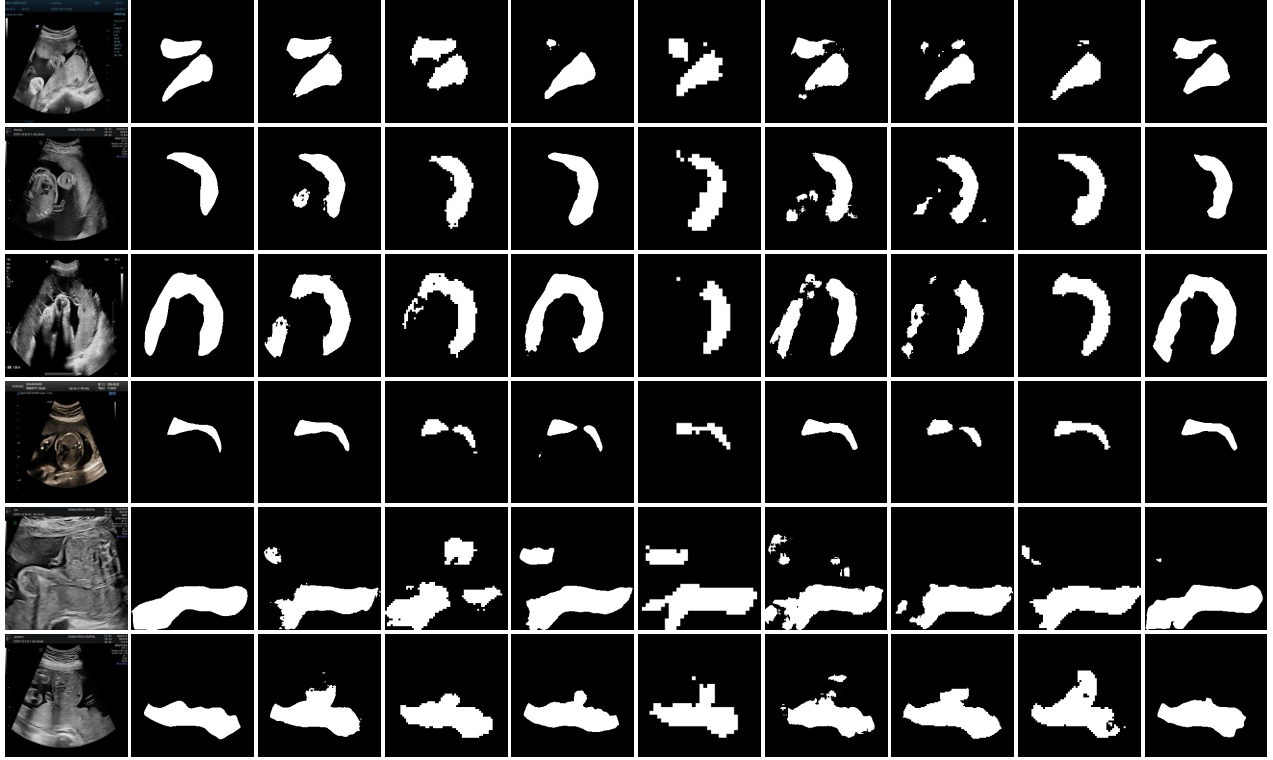


Figure 12. Visual comparison of segmentation results on the Placenta-US dataset. The first two columns represent the original images and ground-truth masks, followed by the results of DualA-Net, CKDNet, A-Net, BMANet, LMFR-Net, PMFSNet, DSNet, and DECSDF-Net.

BMANet, and DSNet. Meanwhile, DECSDF-Net maintains a high processing speed that is suitable for efficient inference. Combined with the superior segmentation accuracy reported in Tables 2 and 3, these results indicate that DECSDF-Net achieves a favorable balance between segmentation performance, model complexity, and inference efficiency rather than pursuing minimum computational cost alone.

Method	GFLOPs	Params(M)	FPS
DualA-Net ²⁸	5.47	2.58	147.57
CKDNet ²⁹	16.58	59.34	74.20
A-Net ³⁰	0.60	0.39	166.99
BMANet ³¹	9.19	28.40	38.13
LMFR-Net ³²	3.49	0.07	325.25
PMFSNe ³³	0.59	0.33	144.27
DSNet ³⁴	32.72	29.33	175.31
DECSDF-Net	17.96	4.60	160.05

Table 4. Complexity analysis on the Placenta-MR dataset.

Optimizer selection

To determine a suitable optimization strategy for DECSDF-Net, five commonly used optimizers, including Adam, Adamax, RMSprop, Rprop, and SGD, were evaluated on the Placenta-MR dataset. As shown in Table 5, the choice of optimizer has a noticeable influence on segmentation performance. Among the tested optimizers, Adam achieves the best overall results, with Dice, MCC, and Jaccard scores of 0.8838, 0.8775, and 0.7923. RMSprop performs very similarly, reaching 0.8831 Dice, 0.8770 MCC, and 0.7912 Jaccard, but remains slightly lower across all three metrics. Adamax also provides competitive results, with corresponding values of 0.8784, 0.8719, and 0.7837. In comparison, Rprop and SGD achieve lower segmentation accuracy, particularly in terms of Jaccard, with scores of 0.7694 and 0.7542. Overall, Adam provides the most favorable and

consistent performance among the evaluated optimization methods. Its adaptive parameter update mechanism enables stable optimization of DECSDF-Net and leads to the highest overlap between the predicted placental regions and the ground-truth masks. Therefore, Adam was selected as the optimizer for all experiments.

Optimizer	GFLOPs	Params(M)	FPS
Adam	0.8838	0.8775	0.7923
Adamax	0.8784	0.8719	0.7837
RMSprop	0.8831	0.8770	0.7912
Rprop	0.8694	0.8632	0.7694
SGD	0.8594	0.8520	0.7542

Table 5. Optimizer selection on the Placenta-MR dataset.

Conclusion

In this study, we proposed DECSDF-Net, a dual-encoder cross-scale dependency fusion network for placenta segmentation in MR and ultrasound images. The network combines complementary feature extraction from two encoder branches with several targeted modules to improve the representation of placental structures. Specifically, CDCAM enhances directional spatial information, GCSDM strengthens dependency modeling and interaction among hierarchical features, AGERFB improves multi-scale receptive-field representation through channel and spatial attention, and PMSFFM progressively integrates decoder features to recover detailed structural information. Extensive experiments were conducted on two in-house datasets, Placenta-MR and Placenta-US. DECSDF-Net achieved Dice, MCC, and Jaccard scores of 0.8838, 0.8775, and 0.7923 on Placenta-MR, and 0.8778, 0.8642, and 0.7826 on Placenta-US. Ablation experiments confirmed the complementary contributions of the proposed components, while comparisons with several representative segmentation networks demonstrated the competitive performance of DECSDF-Net under both imaging modalities. In addition, the model contains only 4.60 M parameters and reaches an inference speed of 160.05 FPS, indicating a favorable balance between segmentation accuracy and computational efficiency. Although encouraging results were obtained, several aspects can be further explored. The current study is based on datasets collected from a single clinical center, and the variability of imaging devices, acquisition protocols, and patient populations has not yet been fully investigated. Future work will therefore focus on multi-center validation and more diverse clinical data to further assess model robustness and generalization. In addition, lightweight architectural optimization and more effective boundary-aware modeling will be investigated to further improve the applicability of DECSDF-Net in practical placenta assessment and computer-aided prenatal imaging.

References

1. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
2. Han, M. *et al.* Automatic segmentation of human placenta images with u-net. *IEEE Access* **7**, 180083–180092 (2019).
3. Looney, P. *et al.* Fully automated, real-time 3d ultrasound segmentation to estimate first trimester placental volume using deep learning. *JCI insight* **3**, e120178 (2018).
4. Zimmer, V. A. *et al.* Placenta segmentation in ultrasound imaging: Addressing sources of uncertainty and limited field-of-view. *Med. image analysis* **83**, 102639 (2023).
5. Huang, J. *et al.* Deep learning based automatic segmentation of the placenta and uterine cavity on prenatal mr images. In *Proceedings of SPIE—the International Society for Optical Engineering*, vol. 12465, 124650N (2023).
6. Yamamoto, S. *et al.* Planet-s: Automatic semantic segmentation of placenta. *arXiv preprint arXiv:2312.11580* (2023).
7. Wang, Y., Wang, H. & Zhang, F. A medical image segmentation model with auto-dynamic convolution and location attention mechanism. *Comput. Methods Programs Biomed.* **261**, 108593 (2025).
8. Ji, Z., Chen, Z. & Ma, X. Grouped multi-scale vision transformer for medical image segmentation. *Sci. Reports* **15**, 11122 (2025).
9. Chen, J., Liang, Z. & Lu, X. A dual attention and cross layer fusion network with a hybrid cnn and transformer architecture for medical image segmentation. *Sci. reports* **15**, 35707 (2025).
10. Liu, X., Tian, J., Huang, S. & Shen, W. Enhancing medical image segmentation via complementary cnn-transformer fusion and boundary perception. *Front. Comput. Sci.* **7**, 1677905 (2025).

11. Yang, H., Fan, Y. & Yang, P. Medical image segmentation network based on dual-encoder interactive fusion. *Appl. Sci.* **15**, 3785 (2025).
12. Wang, D., Sun, Y., Chen, H. & Zhao, X. Image segmentation network based on enhanced dual encoder. *Sci. Reports* **15**, 35983 (2025).
13. Atabansi, C. C. *et al.* Dcm-net: dual-encoder cnn-mamba network with cross-branch fusion for robust medical image segmentation. *BMC medical imaging* **25**, 395 (2025).
14. Cai, S., Jiang, Y., Xiao, Y., Zeng, J. & Zhou, G. Transumobilenet: Integrating multi-channel attention fusion with hybrid cnn-transformer architecture for medical image segmentation. *Biomed. Signal Process. Control.* **107**, 107850 (2025).
15. Sun, L. *et al.* Kcam-senet: Speech enhancement network with kan-based channel attention module. *Speech Commun.* **182**, 103425 (2026).
16. Cheng, Z., Qu, A. & He, X. Contour-aware semantic segmentation network with spatial attention mechanism for medical image. *The Vis. Comput.* **38**, 749–762 (2022).
17. Lin, C., Hu, X., Zhan, Y. & Hao, X. Mobilenetv2 with spatial attention module for traffic congestion recognition in surveillance images. *Expert. Syst. with Appl.* **255**, 124701 (2024).
18. Luo, H., Zhou, D., Cheng, Y. & Wang, S. Mpeda-net: A lightweight brain tumor segmentation network using multi-perspective extraction and dense attention. *Biomed. Signal Process. Control.* **91**, 106054 (2024).
19. Yuan, H., Chen, L. & He, X. Mmunet: Morphological feature enhancement network for colon cancer segmentation in pathological images. *Biomed. Signal Process. Control.* **91**, 105927 (2024).
20. Wan, M. *et al.* Mhaed-net: a lightweight multiscale hybrid attention encoder-decoder network for the efficient segmentation of industrial forging images. *The J. Supercomput.* **81**, 1001 (2025).
21. Selvaraj, A. & Nithiyaraj, E. Cedrrn: A convolutional encoder-decoder residual neural network for liver tumour segmentation. *Neural Process. Lett.* **55**, 1605–1624 (2023).
22. Wan, M. *et al.* Virus diffusion algorithm: a novel metaheuristic algorithm for segmenting the spun crack. *Meas. Sci. Technol.* **36**, 045601 (2025).
23. Shang, Z., Yu, C., Huang, H. & Li, R. Dcnet: A lightweight retinal vessel segmentation network. *Digit. Signal Process.* **153**, 104651 (2024).
24. Abdel-Nabi, H., Ali, M. Z. & Awajan, A. A joint multi-decoder dual-attention u-net framework for tumor segmentation in whole slide images. *J. King Saud Univ. Comput. Inf. Sci.* **35**, 101835 (2023).
25. Yang, M.-Y., Shen, Q.-L., Xu, D.-T., Sun, X.-L. & Wu, Q.-B. Striped wrinet: automatic wrinkle segmentation based on striped attention module. *Biomed. Signal Process. Control.* **90**, 105817 (2024).
26. Liu, Y., Deng, H. & Fu, J. Dcm-net: A novel dual-branch cnn–mamba cross-layer feature fusion network for medical image segmentation. *Biomed. Signal Process. Control.* **114**, 109267 (2026).
27. Meng, W., Liu, S. & Wang, H. Afc-unet: Attention-fused full-scale cnn-transformer unet for medical image segmentation. *Biomed. Signal Process. Control.* **99**, 106839 (2025).
28. Doc, Y. Z. & Doc, S. W. Duala-net: A generalizable and adaptive network with dual-branch encoder for medical image segmentation. *Comput. Methods Programs Biomed.* **243**, 107877 (2024).
29. Jin, Q., Cui, H., Sun, C., Meng, Z. & Su, R. Cascade knowledge diffusion network for skin lesion diagnosis and segmentation. *Appl. soft computing* **99**, 106881 (2021).
30. Chen, B. *et al.* A-net: An a-shape lightweight neural network for real-time surface defect segmentation. *IEEE Transactions on Instrumentation Meas.* **73**, 1–14 (2023).
31. Wu, Z. *et al.* Bmanet: Boundary-guided multi-level attention network for polyp segmentation in colonoscopy images. *Biomed. Signal Process. Control.* **105**, 107524 (2025).
32. Zhang, W., Qu, S. & Feng, Y. Lmfr-net: lightweight multi-scale feature refinement network for retinal vessel segmentation. *Pattern Analysis Appl.* **28**, 44 (2025).
33. Zhong, J. *et al.* Pmfsnet: Polarized multi-scale feature self-attention network for lightweight medical image segmentation. *Comput. Methods Programs Biomed.* **261**, 108611 (2025).
34. Guo, Z. *et al.* Dsnet: A novel way to use atrous convolutions in semantic segmentation. *IEEE Transactions on Circuits Syst. for Video Technol.* **35**, 3679–3692 (2024).

Acknowledgements

Thanks to the editorial team and all the anonymous reviewers who helped us improve the quality of this paper.

Author contributions statement

Conceptualization, methodology, and writing by S.S. and Y.X.; validation and experiments by S.S. and H.F.; writing review and editing, Y.X. and H.F. All authors have read and agreed to the published version of the manuscript.

Data availability

The Placenta-MR and Placenta-US datasets generated and/or analyzed during the current study are not openly available in the public domain, but may request access to these datasets by contacting the original authors.

Funding

This work was supported by the Deyang Municipal Science and Technology Plan Project under Grant 2025SYCX251.

Conflict of Interests

The authors have no conflict of interest to declare.

Additional information

Correspondence and requests for materials should be addressed to Y.X.

Ethics statement

Ethical approval for this study was obtained from the Medical Ethics Committee of Deyang People's Hospital (Approval No. 2025-04-058-K01)