applied sciences

MDPI

Article

DSAFG-UNet: Dual-Branch Semantic Adaptive Fusion Gate UNet for Infrared Sea Surface Image Segmentation

Jiacheng Bu, Zhifei Wu *, Lei Chen, Kai Wu and Yiming Zhang

Naval University of Engineering, Wuhan 430000, China; m24111103@nue.edu.cn (J.B.); m26111101@nue.edu.cn (L.C.); m26111102@nue.edu.cn (K.W.); m25111101@nue.edu.cn (Y.Z.)
* Correspondence: m24111104@nue.edu.cn; Tel.: +86-13135693568

Featured Application

The proposed DSAFG-UNet can be applied to infrared sea surface image segmentation in unmanned surface vessels and maritime surveillance systems, providing structured sea-region information for downstream tasks such as wave-direction detection, target tracking, and autonomous navigation.

Abstract

Infrared sea surface image segmentation is a fundamental task in marine perception, but its performance degrades significantly due to faint wave textures, blurred sea–sky boundaries, and low contrast. To address these issues, this paper proposes a Dual-branch Semantic Adaptive Fusion Gate UNet (DSAFG-UNet). The method replaces the static concatenation in the skip connections of standard UNet with a Dual-branch Semantic Adaptive Fusion Gate (DSAF-Gate) module, which adaptively fuses global semantic and local edge information to dynamically filter skip-connection features. Meanwhile, a multi-scale edge loss and a multi-scale context module are introduced to complement the gating mechanism from the perspectives of boundary supervision and global perception, respectively. Experiments show that DSAFG-UNet achieves an mIoU of 91.96% and a sea-class F1-Score of 99.40%, outperforming gating-based models, multi-scale context models, and attention-based models on multiple accuracy metrics, with good stability and certain generalization capability on a limited external sample, although its computational efficiency still needs improvement. In summary, the proposed model achieves strong overall performance in infrared sea surface segmentation and provides an effective solution for sea-region segmentation.

Keywords: infrared sea surface image; semantic segmentation; attention gating; multi-scale edge loss; multi-scale context

Check for updates

Academic Editor: Atsushi Mase

Received: 19 September 2026
Revised: 17 October 2026
Accepted: 20 October 2026
Published: 24 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Research Background and Significance

With the continuous growth of the demand for maritime transportation, marine resource exploration, and disaster emergency response, all-day, all-weather, and high-precision perception and monitoring of the marine environment has become an important topic in information processing. Benefiting from passive imaging, independence from illumination, and a certain ability to penetrate haze, infrared imaging has become a core means for sea surface target detection and scene understanding at night and under complex weather conditions. However, infrared images generally suffer from low contrast, strong

noise interference, and indistinct gray-level differences between targets and backgrounds, which poses challenges for subsequent automated analysis.

Semantic segmentation of sea surface images is a key component of the marine visual perception system. It assigns each pixel a semantic label such as sky, sea, or obstacle, converting the raw pixel array into a structured scene description. This information is the basic input for downstream tasks such as wave-direction detection, target tracking, situation assessment, and autonomous navigation decision-making, and its accuracy and robustness directly determine the reliability of the entire marine perception system.

The application of traditional image segmentation methods to infrared sea surface scenes faces three challenges. First, wave regions lack stable gray-level gradients and texture features in the infrared band, so general segmentation models can hardly establish effective discriminative cues. Second, the sea surface environment is highly dynamic under the influence of fog, clouds, illumination, and sea states, and the infrared appearance of the same scene varies significantly at different times, which places high demands on the domain adaptation and generalization capabilities of models. Third, acquiring and annotating infrared sea surface segmentation datasets is costly, and the limited sample size can hardly support the effective training and adaptation of general models.

These challenges are clearly reflected in the segmentation results of existing methods. As shown in Figure 1, sea regions are misclassified as sky, the segmentation is severely fragmented, and the boundaries deviate significantly from the true contours. Therefore, this task cannot simply reuse general segmentation solutions and requires dedicated network design and optimization for infrared sea surface scenes.



Figure 1. Typical failure cases of infrared sea surface image segmentation. The green, blue, and red regions represent the sky, sea, and obstacle classes, respectively.

1.2. Related Work

1.2.1. Traditional Image Segmentation

Before the prevalence of deep learning, image segmentation mainly relied on low-level features such as gray levels, edges, and textures. Threshold segmentation is one of the most direct methods. The Otsu algorithm is widely used because of its high computational efficiency and good performance in low-noise, high-contrast scenes [1]; however, it adapts poorly to complex infrared images with unimodal or multimodal histograms. To address this problem, an information-energy minimization thresholding method was proposed, which constructs a one-dimensional gray-level histogram reflecting the spatial positions of

pixels and selects the threshold under the criterion of cross-domain information-energy minimization, improving the adaptability and accuracy of segmentation [2]. Clustering-based methods and active contour models can handle blurred boundaries more flexibly [3], but they are computationally expensive and sensitive to initial conditions. Edge detection and region growing methods can hardly separate targets from backgrounds under low contrast and strong noise. For image sequences, background modeling and density clustering methods combined with spatiotemporal information can exploit inter-frame information to compensate for the insufficient quality of single frames [4]. Overall, traditional methods are still valuable in scenarios with limited computing resources or strict real-time requirements, but their limited generalization to complex scenes has led to their gradual replacement by deep learning methods.

1.2.2. Deep Learning-Based Image Segmentation

Deep learning automatically learns high-level image features through multilayer neural networks and significantly surpasses classical methods in segmentation accuracy and robustness. The Fully Convolutional Network (FCN) first introduced end-to-end training into pixel-level prediction tasks and established the basic encoder–decoder framework. Among CNN architectures, UNet and its variants fuse shallow details with deep semantic information through skip connections and have been widely applied in medical and infrared image segmentation [5,6]. The DeepLab series captures multi-scale context through atrous convolution and Atrous Spatial Pyramid Pooling (ASPP), enlarging the receptive field while maintaining resolution, and has shown good performance in infrared scenes such as photovoltaic panels and power equipment [7,8]. Instance segmentation models such as Mask R-CNN and YOLACT jointly handle object detection and pixel-level segmentation and have been applied in multi-object scenarios such as power inspection [9,10]. Multi-scale context methods such as PSPNet aggregate features from different receptive fields and improve adaptability to scale variations. CNN models combined with attention mechanisms, such as CBAM-UNet, further improve segmentation accuracy by adaptively adjusting feature responses in both channel and spatial dimensions.

In recent years, Transformer architecture has been introduced into the field of segmentation owing to the global modeling capability of its self-attention mechanism. SegFormer adopts a hierarchical encoder and a lightweight MLP decoder and outperforms traditional CNN methods on multiple benchmark datasets [11]. Swin Transformer reduces computational complexity through the shifted-window attention mechanism and achieves performance gains over traditional CNNs in tasks such as long-wave infrared sea–sky-line detection [12]. However, the high computational complexity of Transformers and the training instability caused by scarce infrared data limit their application in resource-constrained scenarios. Visual foundation models represented by the Segment Anything Model (SAM) have also attracted attention, but the significant modality gap between infrared and natural images, together with the relative scarcity of infrared data, constitutes the main challenge for directly applying or adapting foundation models [13]. Lightweight networks such as SMALNet introduce SAM to provide auxiliary supervision signals, significantly reducing computation while maintaining accuracy [14], which provides a feasible path for adapting foundation models to infrared scenes.

1.2.3. Infrared Image Segmentation

Dedicated segmentation studies for infrared images mainly focus on common challenges such as low contrast, strong noise, and small target detection. In the industrial field, fault diagnosis of power equipment is an important application of infrared segmentation; improved marker-controlled watershed algorithms [15] and improved butterfly-

optimization Kapur entropy algorithms [16] are used to extract equipment contours from non-uniform backgrounds, and such methods are usually customized for specific equipment shapes and fault modes. For infrared small target detection, the TEP-SAM framework modulates SAM through temporal-emerged prompts and achieves non-interactive segmentation in image sequences with extremely low signal-to-noise ratios [17].

The above studies cover multiple directions such as industry, medicine, and small target detection, but most methods are designed for specific infrared scenes and have limited generality, and they are particularly weak on the specific task of infrared sea surface segmentation. Sea surface images are characterized by faint wave textures, blurred sea–sky boundaries, and low contrast with the background, and these characteristics significantly degrade the performance of general infrared segmentation methods. In particular, although existing gating mechanisms in skip connections can adapt to the input, they mostly adopt a single-path structure and rely on a single type of feature response. The dual-branch structure proposed in this paper further enriches this line of research by jointly exploiting global semantic and local edge information. Meanwhile, strategies such as multi-scale context and edge constraints are mostly stacked as independent modules and lack collaborative design with the gating mechanism.

In the maritime domain, several dedicated datasets have been developed, including MaSTr1325 [18], MODS [19], LaRS [20], and WaSR [21]. These works focus on visible-light maritime scene understanding and provide valuable benchmarks for USV navigation. Although these datasets also provide pixel-level annotations for sky, water, and obstacles, they are designed for visible-light imagery. The key difference between these works and the present study lies in the image modality: the applicability of these visible-light methods and benchmarks to infrared sea surface segmentation has not been fully explored. In contrast, MassMIND [22] is a thermal long-wave infrared maritime dataset that also provides pixel-level annotations for sky, water, and obstacles, making it a relevant study for infrared maritime perception. These observations further motivate the present study and point out the design direction of the DSAFG-UNet model proposed in Section 2.

1.3. Research Content and Innovations

To address the problems in existing infrared sea surface segmentation methods, this paper proposes a Dual-branch Semantic Adaptive Fusion Gate UNet (DSAFG-UNet). The innovations of this method are reflected in the following two aspects.

First, a Dual-branch Semantic Adaptive Fusion Gate module (DSAF-Gate) is designed to replace the static fusion in the skip connections of standard UNet. The module contains a global semantic branch and a local edge branch: the global branch captures scene-level semantics through pooling and channel recalibration, while the local branch extracts spatial edge responses through the Sobel operator. The two branches are adaptively fused and then concatenated with the decoder features, enabling the network to dynamically adjust its dependence on semantic and edge information according to the input content and to better distinguish sea wave noise from target edges.

Second, the complete DSAFG-UNet architecture is constructed. On the basis of the DSAF-Gate module, a multi-scale edge loss and a multi-scale context module are introduced into the framework, which complement the gating mechanism from the two dimensions of boundary accuracy and global consistency, forming a combined gate-edge-context framework.

2. Materials and Methods

2.1. Overall Architecture of DSAFG-UNet

Standard UNet consists of three parts: an encoder, a decoder, and skip connections [23], as shown in Figure 2. In this architecture, the skip connections directly concatenate encoder features and decoder features. Such an indiscriminate fusion inevitably introduces interference such as sea wave noise and uneven illumination while the decoder receives shallow detail information. For infrared sea surface images, wave regions lack stable gray-level gradients and texture features in the infrared band, and their visual responses are highly similar to background noise, which makes it difficult for direct-concatenation skip connections to distinguish sea surface targets from background interference. To this end, based on standard UNet, this paper designs the DSAFG-UNet. By introducing a gating mechanism into the skip connections, the network can adaptively select and enhance useful semantic features according to the input content and simultaneously suppress irrelevant responses such as sea wave noise. Its structure is shown in Figure 3.

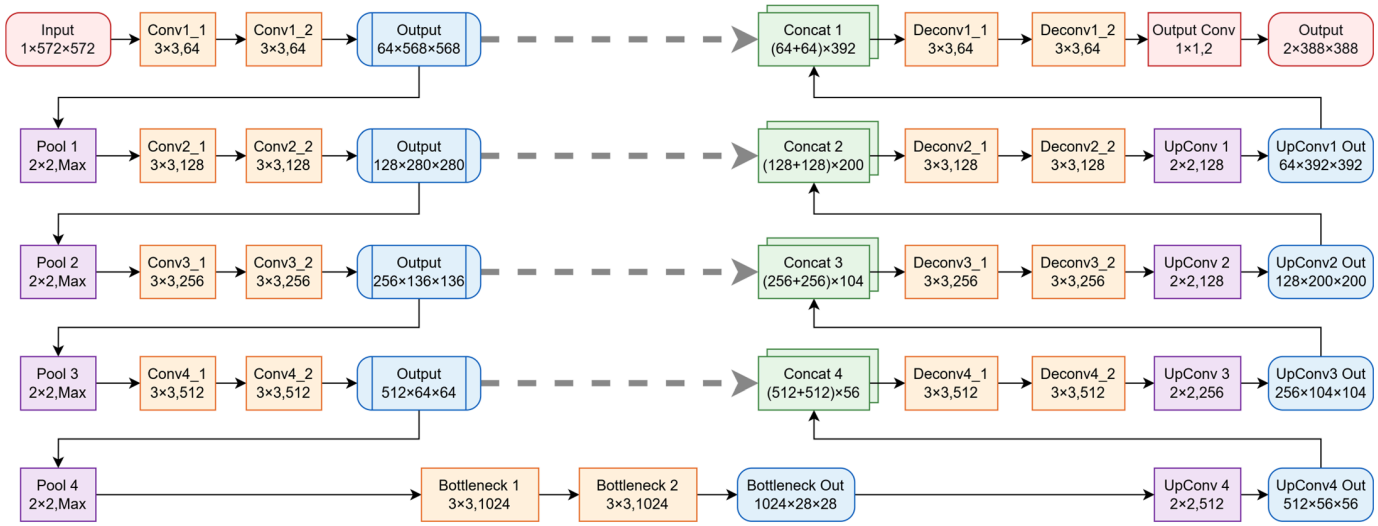


Figure 2. Standard UNet network architecture.

While keeping the main structure of UNet unchanged, DSAFG-UNet replaces the direct concatenation at each skip connection with a DSAF-Gate module (yellow region in Figure 3). This gating module adaptively filters the skip-connection features through a dual-branch structure, so that the decoder obtains filtered shallow detail information at every level. In addition, a multi-scale context module (orange-red region) is introduced at the bottleneck layer, which uses parallel convolutions with different dilation rates to capture multi-scale receptive field features and thus enhances the recognition of large uniform sea regions and large-scale wave structures. Furthermore, a multi-scale edge loss is introduced during training as an auxiliary supervision signal (purple region), which imposes edge constraints on the decoder output to strengthen the characterization of sea surface target contours. The encoder extracts multi-scale features level by level, and the decoder restores the spatial resolution level by level. After each decoder upsampling, the DSAF-Gate module gates the skip-connection features and concatenates the filtered features with the upsampled features, which are then fed into the decoder convolutional layers. Finally, a 1×1 convolution outputs the pixel-level classification results.

Unlike the original UNet, which uses unpadded convolutions, all convolutional layers in the encoder and decoder are set to use padding = 1, so that the spatial size of feature maps remains unchanged before and after convolution. This simplifies the size matching between corresponding decoder and encoder layers.

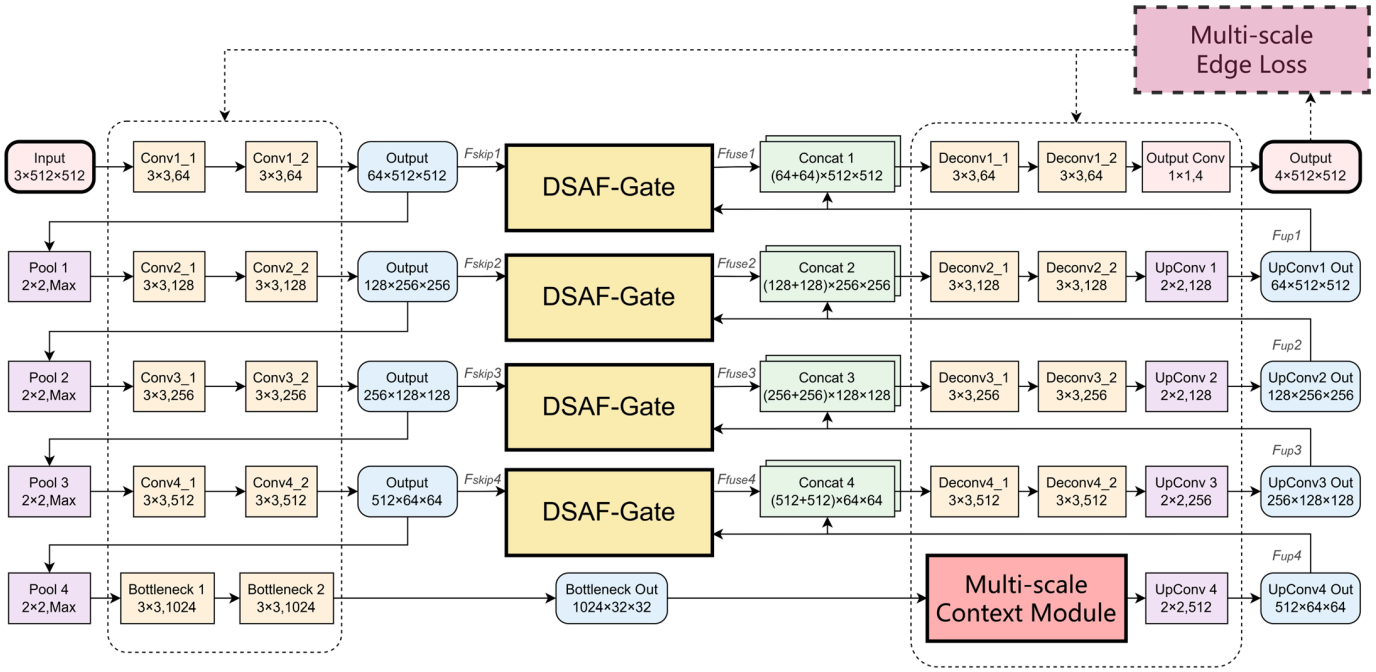


Figure 3. DSAFG-UNet network architecture.

2.2. DSAF-Gate Module

The DSAF-Gate module builds a learnable gating mechanism at the skip connections, drawing on ideas from SE-style channel attention [24], scSE-style spatial attention [25], and SKNet-style adaptive branch fusion [26], as well as the edge-guided gating design in Gated-SCNN [27]. Unlike scSE, which simply combines channel-wise and spatial attention through direct addition, the proposed module explicitly decouples the attention into a global semantic branch and a local edge branch. Furthermore, we introduce a dynamic Softmax-based fusion mechanism (α and β) to adaptively adjust the reliance between these two distinct feature sources based on the input content. It adaptively evaluates the semantic importance of each part of the encoder features according to the input content, assigns higher weights to sea surface target regions to enhance their responses, and assigns lower weights to wave noise regions to suppress their propagation, passing the weighted features to the decoder.

The DSAF-Gate module takes the shallow features F_{skip} from the encoder and the deep features F_{up} upsampled by the current decoder layer as inputs. The two inputs are processed in parallel by a global semantic branch and a local edge branch, and the outputs of the two branches are then adaptively fused. The final output is the gated skip-connection feature F_{gate}, and the architecture of the module is shown in Figure 4. This process can be formulated as:

$$F_{gate} = G(F_{skip}, F_{up}) \tag{1}$$

where $G(\cdot)$ denotes the mapping function of the DSAF-Gate module.

The global semantic branch aims to extract scene-level semantic information from deep features and provide high-level semantic guidance for generating the gating weights. It first applies global average pooling to F_{up}, compressing the spatial dimensions to 1×1 to obtain a global context feature vector. The vector is then transformed by two 1×1 convolutions: the first compresses the number of channels to $1/16$ of the original (i.e., $C/16$), and the second restores it to the original number of channels C. Finally, Sigmoid activation outputs the channel attention weights $W_{glo} \in R^{C\times1\times1}$, which reflects the importance of different

channels at the global semantic level. The weighted features output by the global semantic branch can be expressed as:

$$F_{glo} = F_{skip} \odot W_{glo} \tag{2}$$

where $\odot$ denotes channel-wise multiplication.

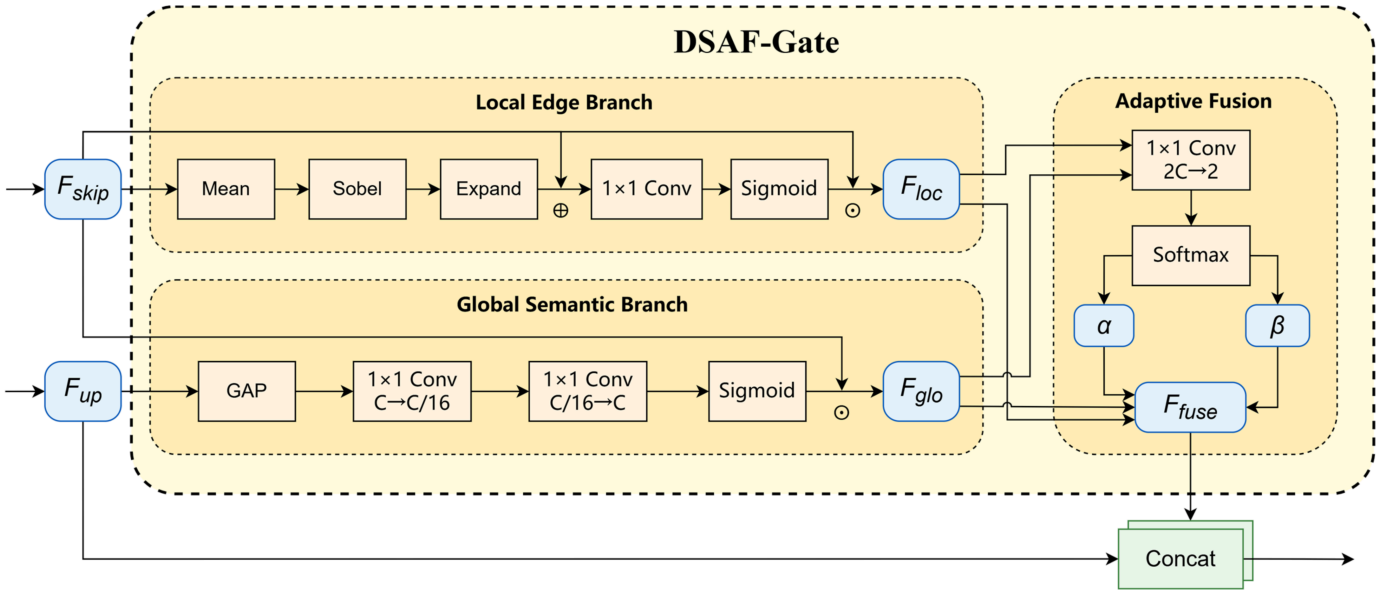


Figure 4. Architecture of the DSAF-Gate module.

The local edge branch aims to extract spatial structure information from shallow features and provide edge position guidance for generating the gating weights. It first averages the shallow features F_{skip} along the channel dimension to obtain a single-channel spatial response map. The Sobel operator is then used to extract the horizontal and vertical edge responses of this map, and the gradient magnitude is taken as the edge saliency map. The saliency map is expanded to the same number of channels as F_{skip} and then added element-wise to the original shallow features to enhance the response intensity of edge regions. A 1×1 convolution is applied to integrate the channel-wise edge responses. Then, we average the integrated features along the channel dimension and apply a Sigmoid activation to generate the final spatial local edge weights $W_{loc} \in R^{1\times H\times W}$. This weight map has the exact same spatial resolution (H × W) as the input shallow feature F_{skip}, but with a single channel. It reflects the likelihood that different spatial positions contain target edges. The weighted features output by the local edge branch can be expressed as:

$$F_{loc} = F_{skip} \odot W_{loc} \tag{3}$$

The adaptive fusion module dynamically integrates the outputs of the global semantic branch and the local edge branch. The outputs of the two branches F_{glo} and F_{loc} are concatenated along the channel dimension and passed through a 1×1 convolution to generate a two-channel fusion weight map. After Softmax normalization, the two channels of the weight map are multiplied with the corresponding branch outputs and summed:

$$F_{fuse} = \alpha \cdot F_{glo} + \beta \cdot F_{loc} \tag{4}$$

where α and β are the two channels of the fusion weight map, satisfying $\alpha + \beta = 1$. This adaptive fusion mechanism enables the network to dynamically adjust its reliance on global semantic and local edge information according to the input content, so that it relies more

on the global semantic branch in texture-uniform regions and more on the local branch in regions with complex boundaries.

Finally, the fused feature F_{fuse} and the decoder upsampled feature F_u are concatenated along the channel dimension, forming the complete input feature of the current decoder layer, which is then fed into the subsequent convolutional layers. This design enables the decoder to preserve target edge information while effectively suppressing background noise such as sea waves at every level.

2.3. Auxiliary Optimization Strategies

On the basis of the adaptive filtering of skip-connection features by the DSAF-Gate module, two auxiliary optimization strategies are further introduced, which enhance the training process in terms of global context perception and edge-supervision constraints, respectively, forming a complementary relationship with DSAF-Gate.

2.3.1. Multi-Scale Edge Loss

Semantic segmentation requires not only correct category prediction for target regions but also high consistency between segmentation boundaries and true target contours. For infrared sea surface images, the boundary between sea and sky is often a blurred transition, and the edges of obstacle targets are difficult to locate precisely because of low contrast. Standard UNet applies only pixel-level supervision to the decoder output through the cross-entropy loss and does not explicitly constrain the spatial accuracy of segmentation boundaries. To further strengthen the characterization of target contours, a multi-scale edge loss is introduced during training. It is combined with the cross-entropy loss as a weighted sum and back-propagated jointly, imposing edge-level supervision on the decoder output.

The segmentation prediction map of size $4 \times 512 \times 512$ is used to compute the main segmentation loss, i.e., the cross-entropy loss L_{ce}. It is also fed into the edge extraction branch, where the Sobel operator computes the predicted edge map. The corresponding edge ground truth is extracted from the training labels. The difference between the predicted edge map and the edge ground truth is measured by the binary cross-entropy loss to obtain the edge loss L_{edge}. Its structure is shown in Figure 5. The total loss is defined as the weighted sum of the two losses:

$$L_{total} = L_{ce} + \lambda \cdot L_{edge} \tag{5}$$

where λ is the weight coefficient of the edge loss, which adjusts the relative contribution of the two losses.

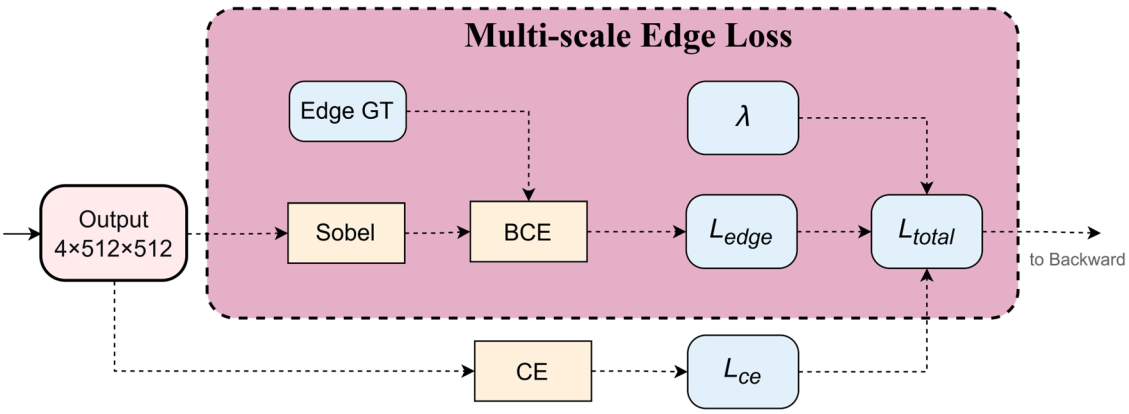


Figure 5. Architecture of the multi-scale edge loss.

The edge loss is active only during training and is not computed during inference, so it does not increase the inference cost. With the edge supervision signal, the network is encouraged to keep the sea region connected during training, which reduces the fragmentation of the sea region under low-contrast infrared conditions.

2.3.2. Multi-Scale Context Module

In standard UNet, the bottleneck layer extracts features only through two consecutive 3×3 convolutions, so its receptive field is relatively limited and it struggles to capture the global context of large uniform sea regions and large-scale wave structures. To this end, a multi-scale context module is introduced at the bottleneck layer. This module is inspired by the atrous spatial pyramid pooling (ASPP) design in DeepLab [28], but differs from ASPP in that it does not include the image-pooling branch and instead uses only parallel dilated convolutions with different dilation rates. It enlarges the receptive field without changing the spatial resolution, thereby enhancing the network's perception of multi-scale scene structures.

The multi-scale context module takes the $1024 \times 32 \times 32$ feature map output by the bottleneck layer as input and contains three parallel branches: branch 1 uses a 1×1 convolution to extract basic features; branch 2 uses a 3×3 convolution with a dilation rate of 2 (padding = 2) to capture medium-scale context; and branch 3 uses a 3×3 convolution with a dilation rate of 4 (padding = 4) to capture large-scale context. The three branches all output 1024 channels, and their outputs are concatenated along the channel dimension into a 3072-channel feature map, which is then passed through a 1×1 convolution to restore 1024 channels, ensuring that the output has the same number of channels and the same spatial resolution as the input. Its structure is shown in Figure 6.

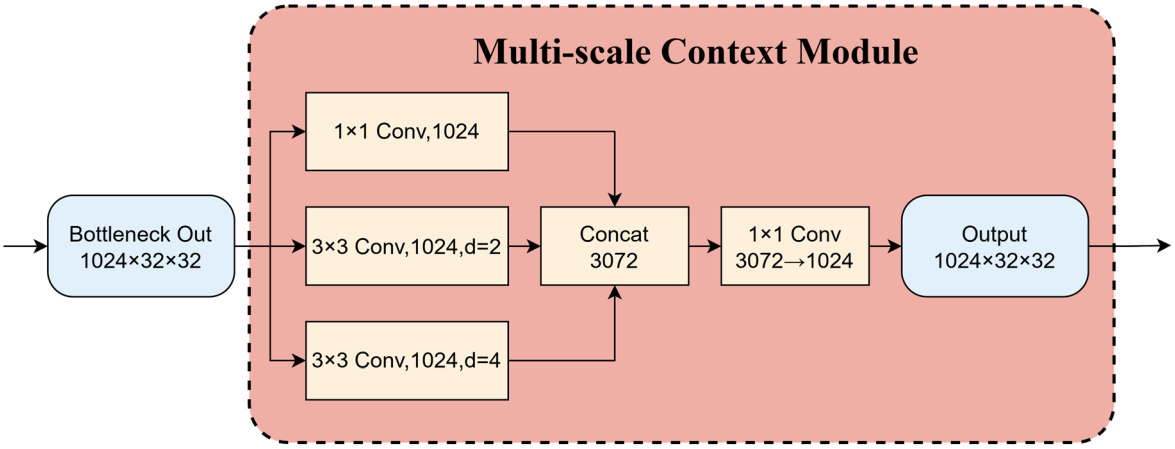


Figure 6. Architecture of the multi-scale context module.

The three branches have different effective receptive fields: the 1×1 convolution branch focuses on local features, the branch with a dilation rate of 2 covers a medium-scale range, and the branch with a dilation rate of 4 covers a large-scale range. By concatenating the outputs of the three branches along the channel dimension, the network can exploit context information at different scales simultaneously, improving the recognition of large-area sea regions and large-scale wave structures. All three branches keep the spatial resolution of 32×32, so the output feature map can be directly fed into the decoder for subsequent upsampling.

2.4. Dataset and Preprocessing

The infrared sea surface image dataset used in this study consists of two parts. The first part is the public "Short-Wave Infrared Ship Target Instance Segmentation and Multi-Target

Tracking Dataset" released by the infrared perception team of Shandong University, which contains infrared images under various sea states [29,30]. The second part is infrared sea surface image data collected by the authors. The two parts cover typical marine scenes including daytime, nighttime, thin fog, and thick clouds. After integration and cleaning, a total of 2760 infrared images are obtained. This distribution covers the three main scene-variation factors considered in this study: sea state, presence or absence of obstacles, and fog conditions. Figure 7 shows the distribution of the dataset over these three main scene-variation factors.

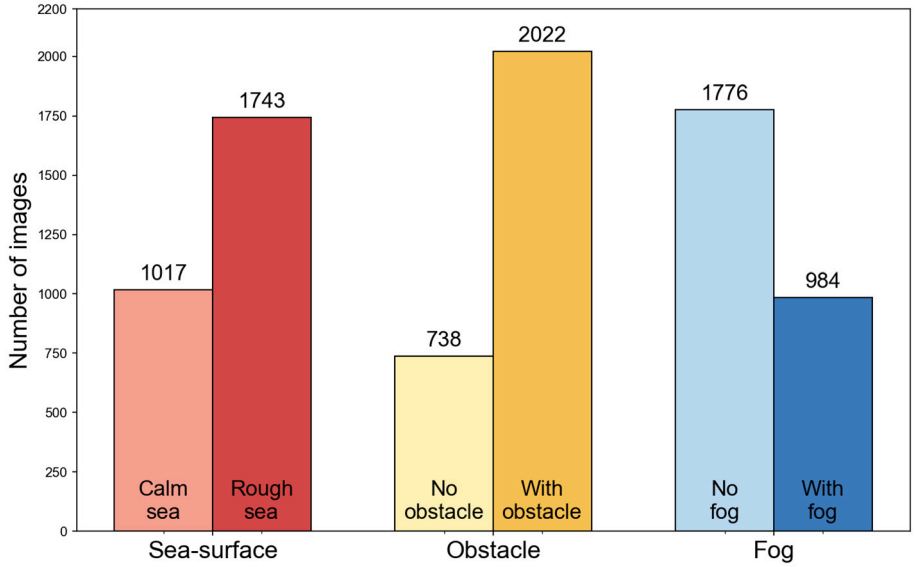


Figure 7. Scene distribution of the dataset.

To improve the contrast of infrared images and enhance the recognizability of sea surface textures, three preprocessing operations are applied to the input images before model training.

1. WCE-Net (Wavelet-domain CNN Enhancement Network) is used to enhance the raw infrared images. This method was proposed in the authors' previous work [31]. It decomposes the input image into one low-frequency structural sub-band and three directional high-frequency texture sub-bands through the discrete wavelet transform, and then applies a lightweight convolutional neural network to enhance the three high-frequency sub-bands in a targeted manner to highlight the directional features of wave textures. Finally, the inverse wavelet transform reconstructs the enhanced high-frequency sub-bands and the original low-frequency sub-band into the enhanced image, as shown in Figure 8. Unlike spatial-domain enhancement methods, WCE-Net separates the underlying structure and texture information of the image in the wavelet domain, effectively enhancing the faint wave textures in infrared images while preserving the overall structure, thus providing inputs with clearer texture information for downstream sea surface segmentation.
2. All images are resized to a spatial resolution of 512 × 512. The raw images may vary in size at acquisition; unified resizing eliminates the influence of size differences on model training and matches the input size expected by the network design.
3. The images are first scaled from [0, 255] to [0, 1] and then normalized with a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225]. These statistics are consistent with those of the ImageNet dataset and have been widely validated in a large number of vision tasks; they help maintain gradient stability during training and thus accelerate convergence.

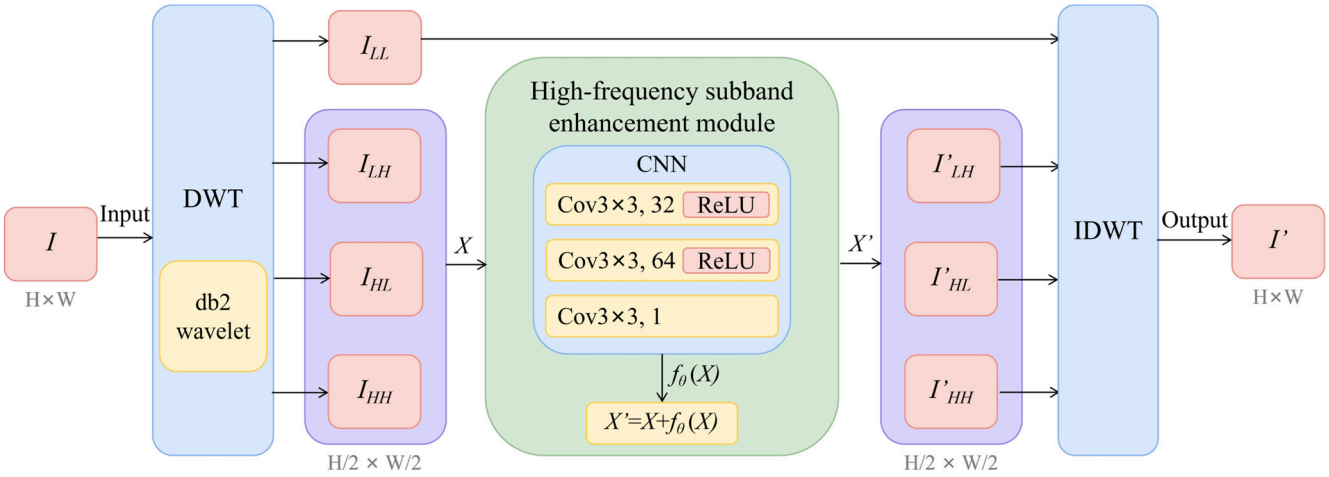


Figure 8. Architecture of WCE-Net.

Figure 9 compares the enhanced image, the original image, and the semantic label. In the semantic label shown on the right of Figure 9, different classes are annotated with different colors, while the gaps between adjacent classes are left unlabeled. All images are annotated manually pixel by pixel, and four classes are defined: 0 denotes the unlabeled background region, which includes the gaps along class boundaries, and is masked out during training and evaluation and excluded from loss computation and accuracy statistics; 1 denotes sky; 2 denotes sea; and 3 denotes obstacles, including sea surface targets such as ships, buoys, and reefs. Although the labels in the right image of Figure 9 are displayed in the form of polygons for visualization, the annotation itself is performed at the pixel level: every pixel enclosed by a given polygon is assigned the corresponding class label, and the gaps not enclosed by any polygon correspond to class 0. Because the boundaries in infrared sea surface images are inherently blurred and the annotation is performed manually, it is difficult to define a precise boundary between adjacent classes, and the annotated regions therefore leave a certain gap along the boundaries.



Figure 9. Comparison of the enhanced image (**left**), the original image (**middle**), and the semantic label (**right**). In the semantic label, the red, yellow, and green outlines represent the sky, obstacle, and sea regions, respectively.

The dataset was split at the frame level. Since the dataset contains multiple scenes and each scene provides only a limited number of frames, a scene-wise split would make the training, validation, and test sets differ too much in scene composition. A frame-level random split was therefore adopted, so that every subset contains frames from all scenes. The dataset is divided into training, validation, and test sets at a ratio of approximately

7:1.5:1.5. All quantitative results in this paper are reported on the test set, with the model checkpoint selected on the validation set.

2.5. Evaluation Metrics

The infrared sea surface segmentation dataset contains four classes of annotations. Class 0 is the unlabeled background region, i.e., the part of the image that does not participate in training or evaluation; class 1 is sky; class 2 is sea; and class 3 is obstacles such as ships, buildings, and reefs. During training and evaluation, class 0 is completely masked out and excluded from loss computation and accuracy statistics. To comprehensively evaluate model performance, the following evaluation metrics are adopted.

1. Best Validation Loss. It evaluates the convergence performance of the model. During training, the cross-entropy loss on the validation set is computed after each epoch. The cross-entropy loss is defined as:

$$L = -\frac{1}{N}\sum_{k=1}^{N}\sum_{c=1}^{C} y_{k,c} log(p_{k,c}) \quad (6)$$

where N is the total number of pixels, $y_{k,c}$ is the one-hot encoding of the true class label of the k-th pixel, $p_{k,c}$ is the probability predicted by the model that the pixel belongs to class c. The lowest loss value on the validation set is taken as the best validation loss of the model, which is used as the criterion for early stopping and model selection.

2. Pixel Accuracy (PA). It computes the proportion of correctly predicted pixels to the total number of pixels. Let n_{ij} denote the number of pixels whose true class is i and which are predicted as class j. When $i = j$, the classification is correct. The PA is calculated as:

$$PA = \frac{\sum_{i=1}^{C} n_{ii}}{\sum_{i=1}^{C}\sum_{j=1}^{C} n_{ij}} \quad (7)$$

The numerator is the total number of correctly classified pixels and the denominator is the total number of pixels. PA is the most intuitive measure of segmentation accuracy, but it can be inflated when the class distribution is severely imbalanced: even if the minority classes are poorly predicted, PA may remain high.

3. Mean Intersection over Union (mIoU). It is one of the core metrics in semantic segmentation. It first computes the ratio of the intersection to the union between the prediction and the ground truth for each class, i.e., the Intersection over Union (IoU), and then averages the results over all classes. The IoU of class i is calculated as:

$$IoU_i = \frac{n_{ii}}{\sum_{j=1}^{C} n_{ij} + \sum_{j=1}^{C} n_{ji} - n_{ii}} \quad (8)$$

The mIoU is calculated as:

$$mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU_i \quad (9)$$

Compared with PA, mIoU computes the metric for each class independently and then averages the results, making it more robust to class imbalance. It is widely accepted as the primary evaluation criterion for semantic segmentation.

4. F1-Score. It is the harmonic mean of precision and recall at the class level, measuring how finely the model distinguishes a specific class. For class i, the precision and recall are calculated as:

$$Precision_i = \frac{n_{ii}}{\sum_{j=1}^{C} n_{ji}} \quad (10)$$

$$Recall_i = \frac{n_{ii}}{\sum_{j=1}^{C} n_{ij}} \tag{11}$$

The F1-Score is the harmonic mean of the two:

$$F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{12}$$

Precision measures the proportion of predicted positives that are actually positive, reflecting the model's sensitivity to false detections; recall measures the proportion of actual positives that are correctly predicted, reflecting its sensitivity to missed detections. F1-Score accounts for both and thus better reflects the fine-grained performance of the model. The F1-Score is reported separately for the sky, sea, and obstacle classes.

5. In addition, the number of parameters (Params), the floating-point operations (FLOPs), and the inference time per image are reported to evaluate the computational efficiency and practical deployability of each model. The number of parameters is the total number of trainable parameters, measured in millions (M), which reflects the storage cost of the model. FLOPs refer to the number of floating-point operations required for one forward inference, measured in units of 10^9, denoted as GFLOPs, which reflects the theoretical computational complexity of the model. The inference time per image is the average time required for the full inference pipeline on a 512 × 512 input image, including image reading, resizing, enhancement, the model forward pass, postprocessing, and result saving, measured in seconds (s). This reflects the actual running speed of the model in a practical setting. Together, these three metrics provide a multi-faceted view of model efficiency.

3. Results

This section evaluates DSAFG-UNet through ablation and comparative experiments. The ablation experiments use six configurations to verify the effectiveness and complementarity of the DSAF-Gate module, the multi-scale edge loss, and the multi-scale context module. In the comparative experiments, DSAFG-UNet is compared with six mainstream semantic segmentation methods to assess its overall performance and generalization ability.

3.1. Experimental Setup

All models in this paper are implemented with the PaddlePaddle framework and trained on a single NVIDIA Tesla V100 GPU with 32 GB of GPU memory. All models are trained from scratch, and no pretrained weights are used. Inference tests are performed on a local computer equipped with an NVIDIA GeForce MX350 GPU with 2 GB of GPU memory.

The main hyperparameters used for model training are listed in Table 1.

Table 1. Main hyperparameters.

Parameter	Value	Parameter	Value
Optimizer	Adam	Maximum training epochs	100
Initial learning rate	5×10^{-5}	Training batch size	4
Learning rate schedule	Cosine annealing	Validation batch size	2

An early stopping strategy based on the validation loss is used during training: training stops once the validation loss no longer decreases, so as to avoid overfitting. The weight coefficient λ of the multi-scale edge loss is set to 0.3 based on preliminary experiments. The weight coefficient λ of the multi-scale edge loss is set to 0.3 based on preliminary experiments. Since the edge loss is a component of module B, this weight is applied only in the groups that include module B, i.e., Groups 03, 05, and 06 in the ablation

experiments and Group 06 (DSAFG-UNet) in the comparative experiments. The other comparison methods do not include this edge loss.

3.2. Ablation Experiments

3.2.1. Ablation Experimental Design

The ablation experiments comprise six configurations, as listed in Table 2. For brevity, A, B, and C denote the DSAF-Gate module, the multi-scale edge loss, and the multi-scale context module, respectively, and each group is named in the form "UNet + module code". The full model, UNet + A + B + C, is the DSAFG-UNet proposed in this paper. All groups are trained with the same hyperparameters described in Section 3.1 and evaluated on the same validation and test sets.

Table 2. Configurations of the ablation experiments.

Group	Configuration	DSAF-Gate	Multi-Scale Edge Loss	Multi-Scale Context Module
01	UNet			
02	UNet + A	√		
03	UNet + A + B	√	√	
04	UNet + A + C	√		√
05	UNet + B + C		√	√
06	UNet + A + B + C	√	√	√

The symbol √ indicates that the corresponding module is included in the configuration.

Among the six configurations, Group 01 is the baseline model, which measures the baseline performance of standard UNet on infrared sea surface segmentation. Group 02 introduces only module A to verify the independent contribution of DSAF-Gate as the core innovation. Groups 03 and 04 add B and C, respectively, on the basis of A to examine the gain of each auxiliary strategy on its own. Group 05 uses only B and C without A to test whether the auxiliary strategies can still work when the core gating module is absent. Group 06 is the full model proposed in this paper to evaluate the combined effect of the three modules.

3.2.2. Ablation Training Process

In the ablation experiments, the training process of every group follows a trend similar to that of the full model, so only one training run of DSAFG-UNet is shown here. Figure 10 shows the training curves of one representative run of DSAFG-UNet, in which the loss, mIoU, PA, and F1-Score are computed on the validation set. As can be seen, the validation loss drops rapidly in the early stage of training, levels off after about 10 epochs, and finally stabilizes at around 0.05. The mIoU, the PA, and the per-class F1-Score increase rapidly within the first 10 epochs and then grow slowly until they reach their respective plateaus, indicating that the training process converges quickly. No obvious abnormal fluctuation appears in any metric during this run, and the training process is relatively stable.

In Figure 10, the curve of the sea class consistently lies above those of the sky and obstacle classes in both the IoU and the F1-Score plots, which indicates that the model segments sea regions with the highest accuracy. Since accurate sea-region segmentation is the goal of this work, the model attains its highest fidelity on the sea class, which meets the task requirement.

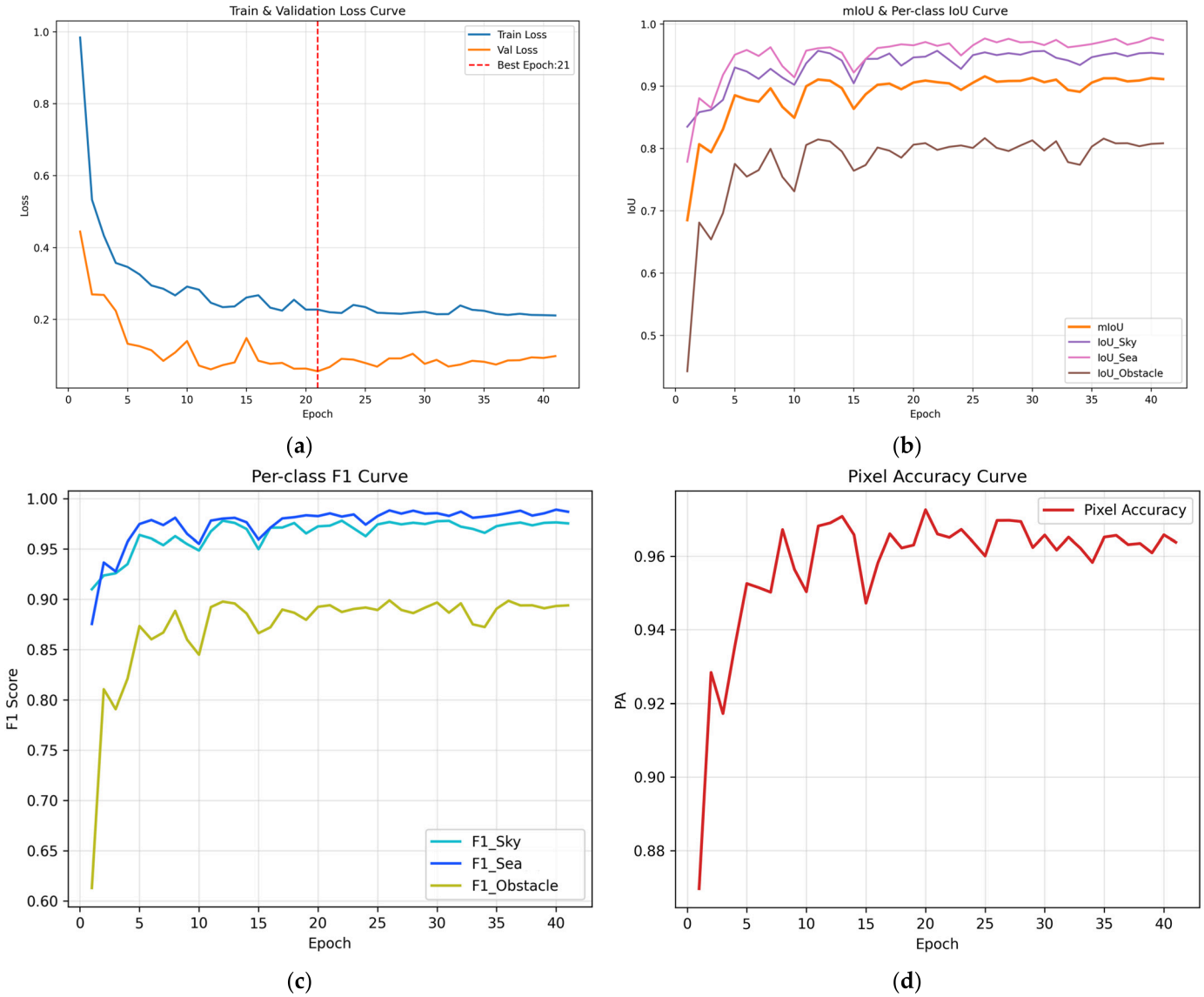


Figure 10. Training curves of DSAFG-UNet: (**a**) training loss and validation loss; (**b**) mIoU and per-class IoU; (**c**) F1-Score of the sky, sea, and obstacle classes; (**d**) pixel accuracy (PA).

It should be noted that the training loss in the figure is slightly higher than the validation loss. This is because module B is enabled only during training: the training loss contains both the cross-entropy loss and the edge loss, whereas only the cross-entropy loss is computed during validation. The two values are therefore not directly comparable. The same phenomenon appears in the other groups that include B.

3.2.3. Ablation Segmentation Performance

The quantitative results of the six ablation groups on the test set are given in Table 3, reported as mean ± standard deviation over 10 repeated runs for each group. Table 3 shows that the full model attains the lowest best validation loss among the six groups, which confirms its advantage in convergence performance. For the sea-class F1-Score, the full model achieves the highest value among the six groups, while its PA is slightly lower than that of Group 04, with only a marginal difference. Overall, the full model remains at a high level on all accuracy metrics.

Table 3. Ablation experiment results.

Group	Configuration	Best Epoch	Best Validation Loss	mIoU	PA	F1_Sea
01	UNet	31.2 ± 12.0	0.0697 ± 0.0060	0.9049 ± 0.0061	0.9721 ± 0.0027	0.9641 ± 0.0185
02	UNet + A	37.0 ± 6.9	0.0564 ± 0.0023	0.9176 ± 0.0073	0.9863 ± 0.0011	0.9886 ± 0.0047
03	UNet + A + B	22.0 ± 5.2	0.0524 ± 0.0014	0.9170 ± 0.0037	0.9857 ± 0.0023	0.9693 ± 0.0378
04	UNet + A + C	**19.7 ± 3.2**	0.0521 ± 0.0018	0.9192 ± 0.0041	**0.9866 ± 0.0008**	0.9912 ± 0.0041
05	UNet + B + C	21.7 ± 7.1	0.0565 ± 0.0047	0.9115 ± 0.0039	0.9792 ± 0.0016	0.9818 ± 0.0059
06	UNet + A + B + C	21.8 ± 6.1	**0.0506 ± 0.0036**	**0.9196 ± 0.0031**	0.9858 ± 0.0017	**0.9940 ± 0.0020**

Group	Configuration	F1_Sky	F1_Obstacle	Inference time per image (s)	Parameters (M)	GFLOPs
01	UNet	0.9883 ± 0.0074	0.8933 ± 0.0207	**11.39 ± 0.14**	**31.06**	**218.91**
02	UNet + A	0.9803 ± 0.0139	0.8974 ± 0.0135	13.54 ± 0.15	31.45	223.39
03	UNet + A + B	**0.9890 ± 0.0221**	**0.9084 ± 0.0258**	13.28 ± 0.47	31.45	223.39
04	UNet + A + C	0.9813 ± 0.0122	0.8961 ± 0.0010	14.13 ± 0.51	54.53	247.02
05	UNet + B + C	0.9758 ± 0.0044	0.8996 ± 0.0136	11.89 ± 0.17	54.13	242.53
06	UNet + A + B + C	0.9792 ± 0.0004	0.8959 ± 0.0096	14.44 ± 0.28	54.53	247.02

Bold values indicate the best value in each column.

To further verify that this advantage is statistically reliable, a paired t-test was conducted over 10 repeated runs, with $p < 0.05$ considered significant. The results show that DSAFG-UNet achieves significantly higher mIoU than the standard UNet baseline ($p < 0.001$) and Group 05 ($p < 0.001$). The differences between DSAFG-UNet and Groups 02, 03, and 04 are not statistically significant, which is expected, since these groups already contain the core DSAF-Gate module. This further confirms that module A is the core of the proposed method, while modules B and C serve as auxiliary components.

Figure 11 compares the segmentation results of the six ablation models on five test images. The first row shows the five original test images, rows 2 to 7 show the results of Groups 01 to 06, respectively, and each column corresponds to the same test image. The first four columns (a–d) are four images randomly drawn from the test set, covering scenes such as thin fog, obstacles, and blurred boundaries. The fifth column (e) is an independent image that does not belong to the dataset and was not used in training; it is used to evaluate the generalization ability of the models. Selected evaluation metrics of each model on each image are labeled below the corresponding result, and the best value for each image is marked in red.

Figure 11 shows that, on the four images from the dataset, all ablation models built on UNet exhibit good basic segmentation ability, although the groups still differ in detail. Groups 01 to 05 all show slight fragmentation, which makes the same semantic class spatially discontinuous; this is particularly obvious near the boundary with obstacles. In contrast, the results of Group 06 are more coherent, the semantic regions are well connected, and no obvious fragmentation appears. In terms of the quantitative metrics, Group 06 outperforms the other configurations on all four metrics for most images, which agrees with its visual performance.

On the external image (e), the segmentation quality of every model decreases relative to that on the images from the dataset, which indicates that generalization to unseen scenes is limited. DSAFG-UNet nevertheless maintains relatively good results, with more complete connectivity in the sea region, showing that it generalizes better than standard UNet. However, the overall drop in segmentation quality on the external image also indicates that the current training data are limited in scale and that the adaptability of the model to new scenes still needs improvement.

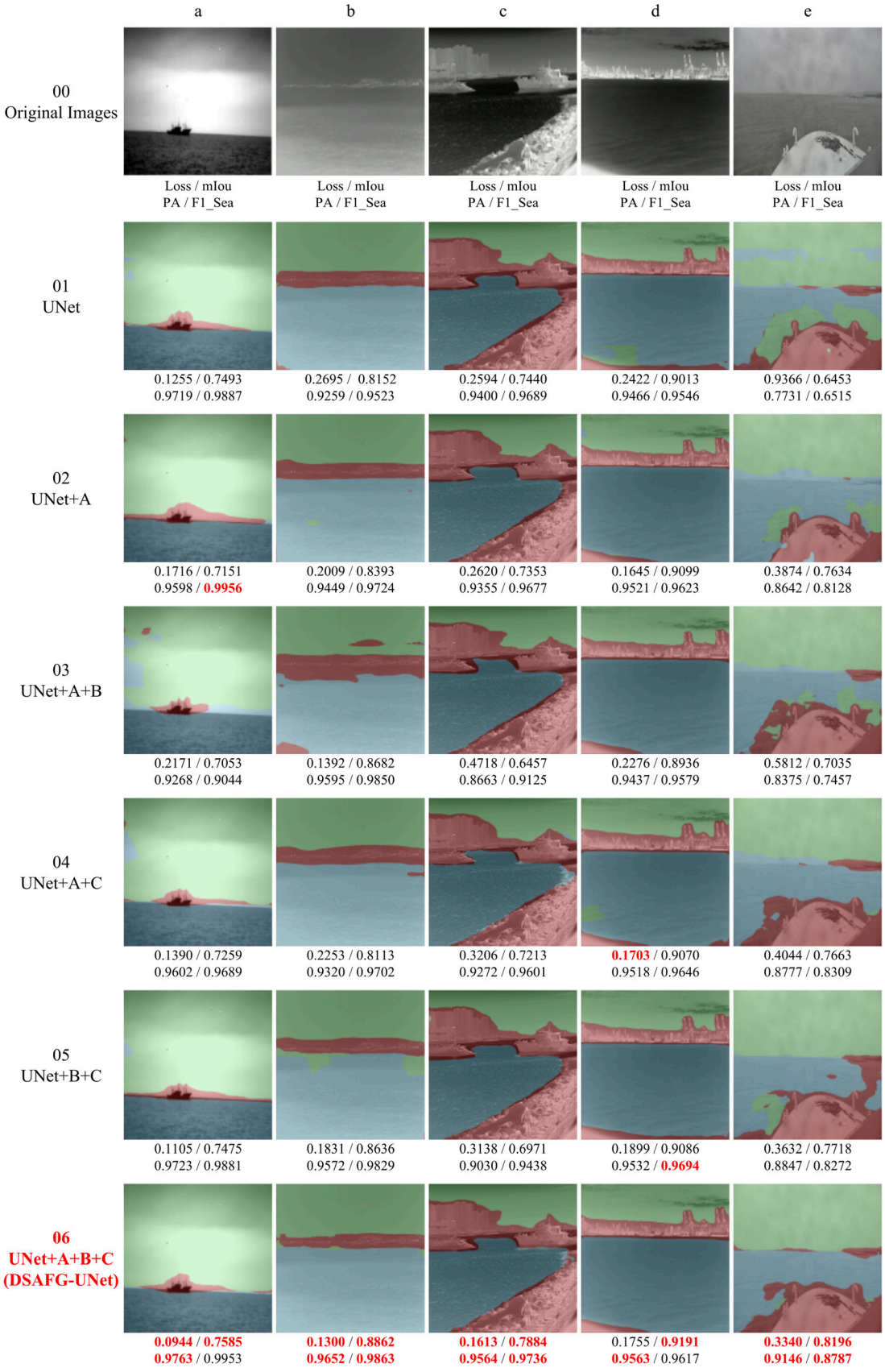


Figure 11. Comparison of segmentation results of the ablation experiments. (**a**–**d**) Four different test images randomly selected from the test set; (**e**) An independent external image not used in training. The green, blue, and red regions represent the sky, sea, and obstacle classes, respectively.

3.3. Comparative Experiments

3.3.1. Comparative Experimental Design

This section compares DSAFG-UNet with several mainstream semantic segmentation methods to verify its overall performance advantage on infrared sea surface segmentation. The comparison models are as follows. Groups 07 (UNet + GateSkip) and 08 (UNet + CBAM) use the traditional attention gate [32] and a general attention mechanism [33], respectively, and serve to verify the superiority of the proposed DSAF-Gate over traditional gating and general attention. Group 09 (DeepLabv3+) is a general segmentation method based on atrous convolution [28] and represents multi-scale context methods. Groups 10 (SegFormer-B0) and 11 (SegFormer-B4) are lightweight and heavyweight Transformer architectures, respectively [11]. Group 12 (SegNeXt) is a CNN architecture based on convolutional attention [34]. Groups 10, 11, and 12 are used together to assess the applicability of attention-based methods to this task, where Groups 10 and 11 represent self-attention and Group 12 represents convolutional attention.

All seven models are trained on the same training and validation sets. As Transformer-based architectures are more sensitive to the learning rate, the initial learning rate of Groups 10, 11, and 12 is set to 6×10^{-4}, as determined by preliminary experiments. The remaining models are trained with the hyperparameters described in Section 3.1, so that all methods can be compared under reasonably comparable configurations.

3.3.2. Comparative Training Process

Figure 12 shows the curves of selected evaluation metrics during the training of the seven comparison models, and the best run of each model is selected. Because early stopping terminates the models at different epochs, the loss curves in Figure 12a are extended horizontally to the right with dashed lines after training ends for ease of comparison.

Figure 12a shows that DeepLabv3+ converges fastest, followed by DSAFG-UNet. The loss curves of the other five methods almost always lie above those of DSAFG-UNet. The SegFormer models converge slowly: their loss remains high in the early and middle stages of training and only decreases to an average level in the late stage. Figure 12b–d show consistent overall trends. The training curves of the SegFormer models fluctuate strongly throughout, improve slowly with periodic drops, and finally undergo training collapse. The curves of the other five groups are stable overall, rising gradually and then leveling off, and they reach similar final levels.

3.3.3. Comparative Segmentation Performance

The quantitative results of the seven comparison models on the test set are given in Table 4, reported as mean $\pm$ standard deviation over 10 repeated runs for each model. DSAFG-UNet achieves the best values on several accuracy metrics and shows good recognition capability for sea regions.

When the loss curve of the best run in Figure 12a is considered together with the average results over multiple runs in Table 4, DSAFG-UNet performs best not only in a single best run but also on average over multiple runs, attaining the lowest validation loss. This indicates that its convergence performance is consistent.

To further verify that this advantage is statistically reliable, a paired t-test was conducted over 10 repeated runs, with $p < 0.05$ considered significant. The results show that DSAFG-UNet achieves significantly higher mIoU than UNet + GateSkip and UNet + CBAM ($p < 0.05$), and the improvements over DeepLabv3+, SegFormer-B0, SegFormer-B4, and SegNeXt are even more significant ($p < 0.001$). This confirms that the superiority of DSAFG-UNet is not caused by random fluctuations but is statistically significant.

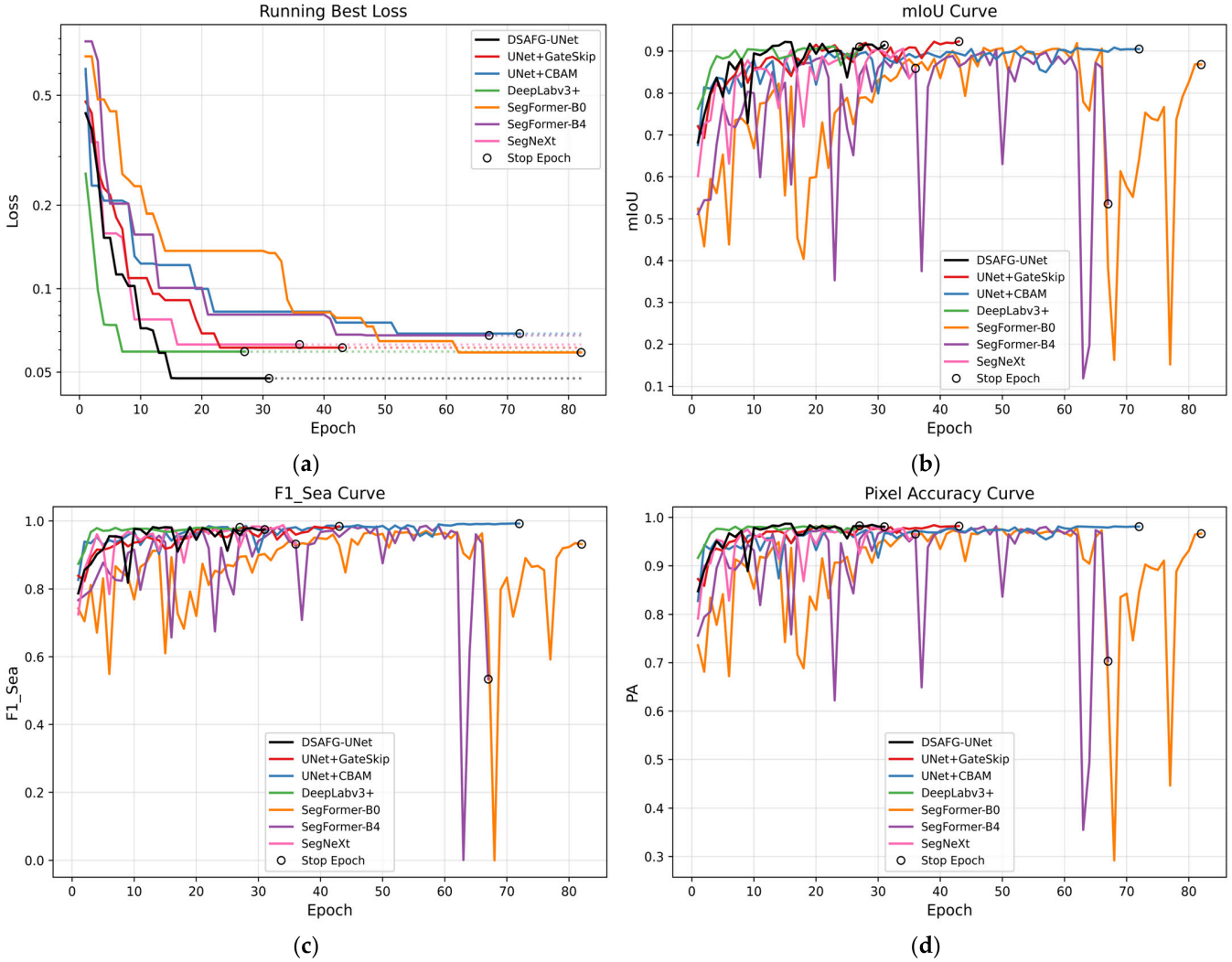


Figure 12. Training curves of the comparative experiments: (**a**) historical best loss of the seven models (dashed lines indicate the horizontal extension after early stopping); (**b**) mIoU of the seven models; (**c**) F1-Score of the sea class for the seven models; (**d**) PA of the seven models.

Figure 13 further presents the segmentation results of the seven comparison models on the five test images, providing an intuitive view of the segmentation ability of each model. The first row shows the five original test images, rows 2 to 8 show the results of Groups 06 to 12, and the best run of each model is selected.

Figure 13 shows that all seven models are capable of basic segmentation on the four images from the dataset, and the differences between the models are small. For image (a), which has an open field of view and no obstacles, the results of the models are almost identical and the differences are negligible. For image (c), which contains obstacles with complex contours, DSAFG-UNet preserves the contours clearly better than the other methods. According to the evaluation metrics of each image, the proposed method is superior on most metrics, and the main problem of the other models is that they misclassify the sea–sky line as an obstacle.

Table 4. Comparative experiment results.

Group	Configuration	Best Epoch	Best Validation Loss	mIoU	PA	F1_Sea
06	DSAFG-UNet	21.8 ± 6.1	**0.0506 ± 0.0036**	**0.9196 ± 0.0031**	**0.9858 ± 0.0017**	**0.9940 ± 0.0020**
07	UNet + GateSkip	29.7 ± 16.1	0.0646 ± 0.0041	0.9074 ± 0.0061	0.9819 ± 0.0024	0.9815 ± 0.0112
08	UNet + CBAM	39.0 ± 18.4	0.0741 ± 0.0054	0.9070 ± 0.0185	0.9781 ± 0.0030	0.9843 ± 0.0097
09	DeepLabv3+	**7.7 ± 3.1**	0.0621 ± 0.0025	0.8857 ± 0.0150	0.9808 ± 0.0010	0.9827 ± 0.0063
10	SegFormer-B0	43.7 ± 15.9	0.0639 ± 0.0067	0.8574 ± 0.0216	0.9493 ± 0.0034	0.9654 ± 0.0308
11	SegFormer-B4	48.0 ± 19.5	0.0744 ± 0.0058	0.8853 ± 0.0089	0.9368 ± 0.0045	0.9786 ± 0.0978
12	SegNeXt	10.7 ± 4.7	0.0886 ± 0.0227	0.8498 ± 0.0183	0.9435 ± 0.0088	0.9671 ± 0.0127

Group	Configuration	F1_Sky	F1_Obstacle	Inference time per image (s)	Parameters (M)	GFLOPs
06	DSAFG-UNet	0.9792 ± 0.0004	**0.8959 ± 0.0096**	14.44 ± 0.28	54.53	247.02
07	UNet + GateSkip	**0.9836 ± 0.0056**	0.8832 ± 0.0127	12.60 ± 0.34	31.40	223.23
08	UNet + CBAM	0.9752 ± 0.0131	0.8891 ± 0.0183	13.46 ± 1.36	31.19	218.91
09	DeepLabv3+	0.9804 ± 0.0107	0.8437 ± 0.0321	3.99 ± 0.23	19.42	71.83
10	SegFormer-B0	0.9576 ± 0.0215	0.8376 ± 0.0567	**3.32 ± 0.06**	**8.52**	**9.78**
11	SegFormer-B4	0.9670 ± 0.0275	0.8648 ± 0.2081	16.94 ± 0.35	69.89	54.59
12	SegNeXt	0.9403 ± 0.0303	0.8411 ± 0.0164	4.19 ± 0.18	20.67	19.76

Bold values indicate the best value in each column.

On the external image (e), the segmentation quality of every model decreases relative to that on the images from the dataset. This is consistent with the general observation that model performance degrades under distribution shift, that is, when the test scene differs from the training distribution. DSAFG-UNet performs best on this image and is numerically superior to the other models, which indicates stronger generalization. The problems of the other models are concentrated in obstacle regions, where sky or sea areas inside the obstacles are misclassified and the obstacle boundaries expand outward.

In addition, a comparison under similar model sizes further supports the effectiveness of DSAF-Gate. Group 02 (31.45 M) has a comparable parameter count to Group 07 (31.40 M) and Group 08 (31.19 M). Despite containing only module A, Group 02 achieves higher mIoU and PA than both comparison models, indicating that the proposed DSAF-Gate is more effective than the traditional attention gate and the general attention mechanism under comparable model sizes. Moreover, a paired t-test on mIoU over the 10 runs confirms that the improvement of Group 02 over Group 07 and Group 08 is statistically significant (p = 0.00001 and p = 0.020, respectively).

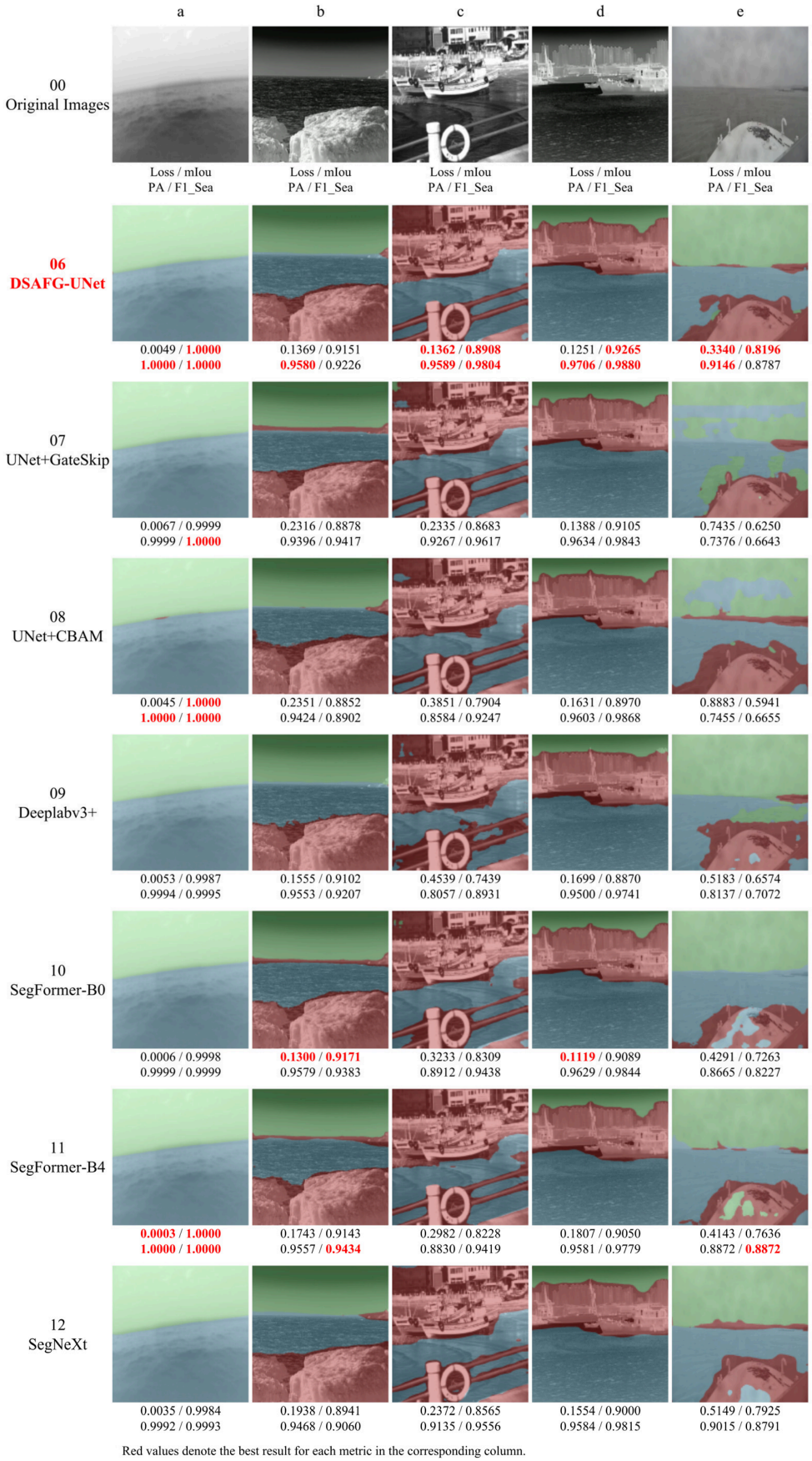


Figure 13. Comparison of segmentation results of the comparative experiments. (**a**–**d**) Four different test images randomly selected from the test set; (**e**) An independent external image not used in training. The green, blue, and red regions represent the sky, sea, and obstacle classes, respectively.

4. Discussion

4.1. Contribution of Each Module

DSAF-Gate (module A) is the core innovation of this paper. In Figure 14, the sea-class F1-Score of Group 02 deviates positively from the baseline, improving by about 2.5%, and mIoU and PA also deviate positively. This shows that the gain of the module is concentrated mainly in the sea class and raises the overall accuracy at the same time. However, the module brings no obvious contribution to the sky and obstacle classes, whose metrics change by less than 1%. This agrees with the design goal of adaptively adjusting the weights of semantic and edge information through a dual-branch structure, and indicates that the module effectively suppresses the interference of sea wave noise with the skip-connection features; the reduction in fragmentation in the sea region in Figure 11 is consistent with this. In terms of efficiency, Table 3 shows that DSAF-Gate is the main source of the increase in inference time. Groups 01 and 05, which do not contain this module, have similar inference times, whereas Groups 02, 03, 04, and 06, which contain it, all take noticeably longer by about 1.5–3 s. The module thus trades longer inference time for higher accuracy.

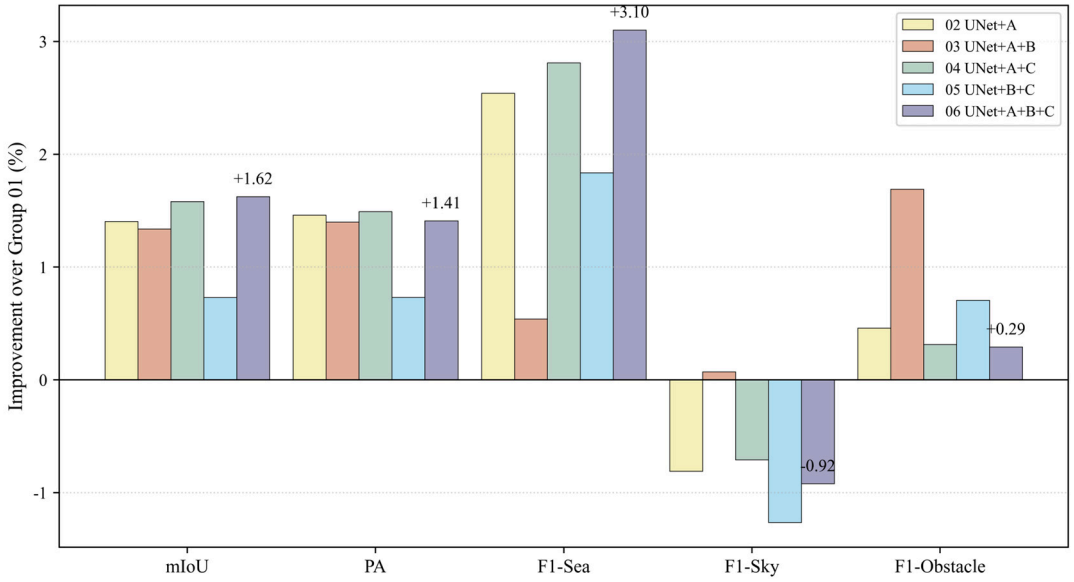


Figure 14. Improvement rate of each accuracy metric in the ablation experiments relative to the baseline.

To provide a more intuitive understanding of the adaptive filtering mechanism of the DSAF-Gate, the W_{loc} at different decoder levels are visualized in Figure 15. In the shallow layer (Gate 1), the spatial response is predominantly in the medium-weight range. This is because the local edge branch is overwhelmed by dense, high-frequency textures (such as fine wave ripples and obstacle textures), making it difficult to distinguish meaningful edges. As the network deepens to Gate 2 and Gate 3, the response becomes increasingly selective and semantically discriminative. Regions lacking high-frequency textures exhibit mixed low-to-medium weight responses, whereas highly textured regions are densely mapped with high-weight responses.

It is worth noting that the evolution does not follow a simple linear gradient; a distinct reversal occurs at Gate 4, which is not merely a result of resolution reduction. At this deepest level, as the feature resolution is significantly reduced, the high-frequency textures are abstracted, and the internal regions shift to low-to-medium weights. Consequently, the structural boundaries that were previously masked by dense textures emerge as highly

salient high-weight strips. This reversal demonstrates that the deepest gate performs a critical semantic abstraction, discarding detailed textures and locking solely onto macroscopic semantic transitions. Moreover, this boundary response is the cumulative effect of the multi-level gate: the shallower gates progressively filter out and distinguish high-frequency background noise, enabling the deepest gate to focus exclusively on the most stable and critical semantic transitions.

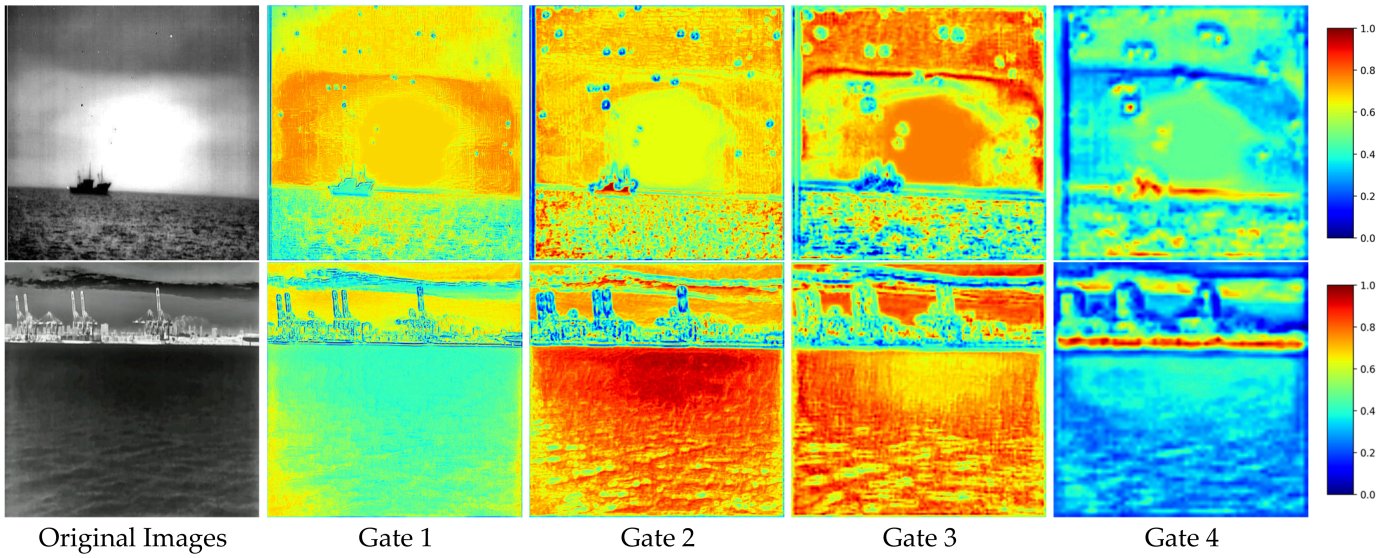


Figure 15. Visualization of the W_{loc} across different decoder levels. The columns from left to right represent the original image and the W_{loc} maps at Gate 1, Gate 2, Gate 3, and Gate 4, respectively.

The multi-scale edge loss (module B) contributes not to accuracy but to boundary quality. In Figure 14, mIoU and PA of Group 03 are close to those of Group 02, but the F1-Score of the sea class is lower than that of Group 02. The positive deviation of the sea class narrows markedly, from about 2.5% to about 0.5%, while the sky and obstacle classes show positive deviations instead. With stronger boundary supervision, the model sacrifices part of its pixel-level accuracy near boundaries in order to keep them smooth, which changes the class assignment of some boundary pixels; boundary quality improves while sea-class accuracy fluctuates slightly. The enlarged local comparison in Figure 16 shows that the groups with the multi-scale edge loss (03 and 06) have smoother edges than the corresponding groups without it (02 and 04), which show clear fragmentation. This improvement in boundary quality is of practical value for applications that require subsequent in-depth processing of sea regions.

The multi-scale context module (module C) improves the overall metrics by enlarging the receptive field of the bottleneck layer through parallel convolutions with multiple dilation rates, which strengthens the global perception of sea regions. In Figure 14, Group 04 deviates positively from the baseline on all three metrics, namely mIoU, PA, and the F1-Score of the sea class, and the deviations are larger than those of Group 02. Calculations based on the data in Table 3 show that the module adds about 23 M parameters, making it the main source of parameter growth, and also increases the inference time to some extent.

Among the three modules, which complement each other, DSAF-Gate is the foundation. In Figure 14, Group 05, which does not contain module A, shows lower mIoU and PA than Groups 02, 04, and 06, although it still performs better than the UNet baseline (Group 01). Its sea-class F1-Score also decreases relative to Group 02. This indicates that, without the core gating module, the two auxiliary strategies cannot fully exploit their potential, and their combination performs worse than configurations that include DSAF-Gate. Group 06, in contrast, reaches a high level on most accuracy metrics and performs best overall. This

indicates that DSAF-Gate is the basis for the three modules to work together: DSAF-Gate filters the useful information in the skip connections, the multi-scale edge loss helps keep the sea region connected on this basis, and the multi-scale context module enhances global perception at the bottleneck layer.

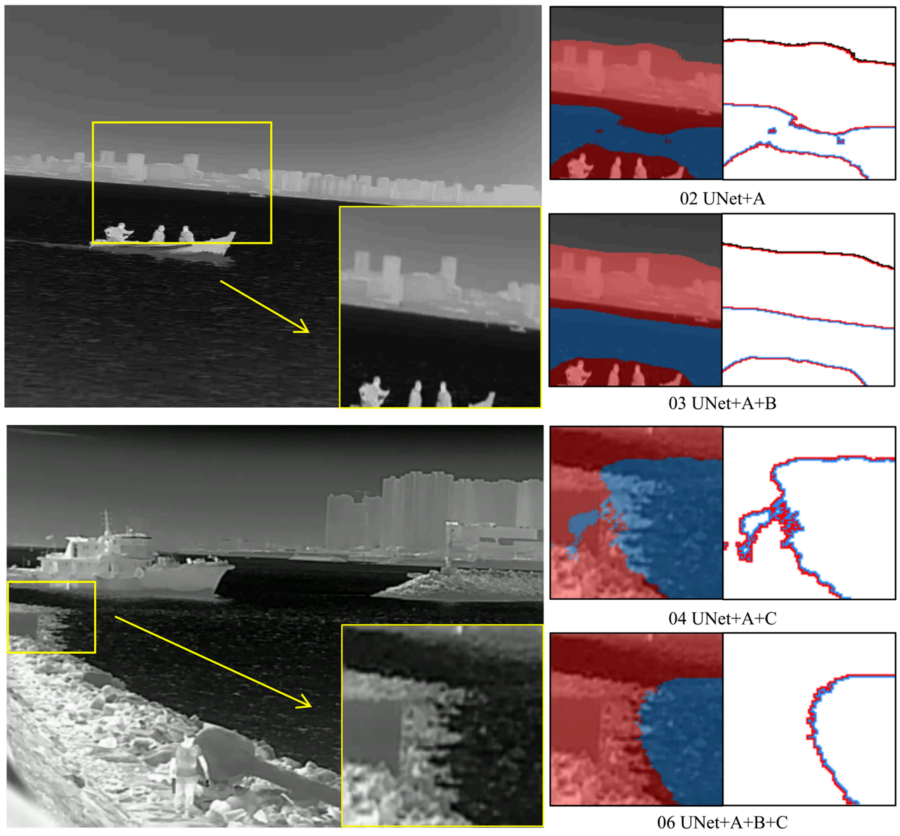


Figure 16. Effect of the multi-scale edge loss on boundary quality. The red and blue regions represent the segmentation results of the obstacle and sea classes, respectively.

4.2. Performance Comparison

Figure 17 presents the ranking of the seven models on six metrics as a heat map, where the Total column is the sum of the individual rankings and a smaller value indicates a better overall ranking. Figure 17 shows that DSAFG-UNet has the smallest Total value and thus the best overall performance. Groups 06 to 09 rank in the upper middle, and Groups 07 and 08 rank below Group 06 on most metrics, which indicates that, for the infrared sea surface segmentation task, neither the traditional attention gate nor general attention mechanisms outperform the DSAF-Gate proposed in this paper. Groups 09 and 10 are the two fastest models, with inference times of 3.99 s and 3.32 s, respectively, but both rank low in mIoU, so their advantage is concentrated in efficiency. Groups 10 to 12 rank clearly lower overall for two main reasons. First, self-attention lacks the local inductive bias of convolution and struggles to learn local features adequately when the training data are limited. Second, in our setting, the core difficulty of this task is the local discrimination of wave textures, which calls for mechanisms that emphasize local receptive fields and edge sensitivity, whereas the global modeling of attention does not match this requirement; the design of the local edge branch in DSAF-Gate targets exactly this point.

Figure 18 shows the relationship between accuracy and efficiency of the models, where the marker size indicates the number of parameters. The models fall into clearly separated regions. DSAFG-UNet is at the top, reaching the highest accuracy at a relatively high processing time. SegFormer-B4 has the most parameters among the seven models but only

ranks in the middle in accuracy, so simply enlarging a model does not improve accuracy. The models in the lower-left region have inference times of about 4 s but mIoU values below 0.90, showing that high efficiency comes at the cost of a clear drop in accuracy. No model achieves both the highest accuracy and the best efficiency, so the two cannot be achieved simultaneously in this task. Under this trade-off, the proposed method prioritizes accuracy and accepts a certain efficiency cost in exchange for higher segmentation quality.

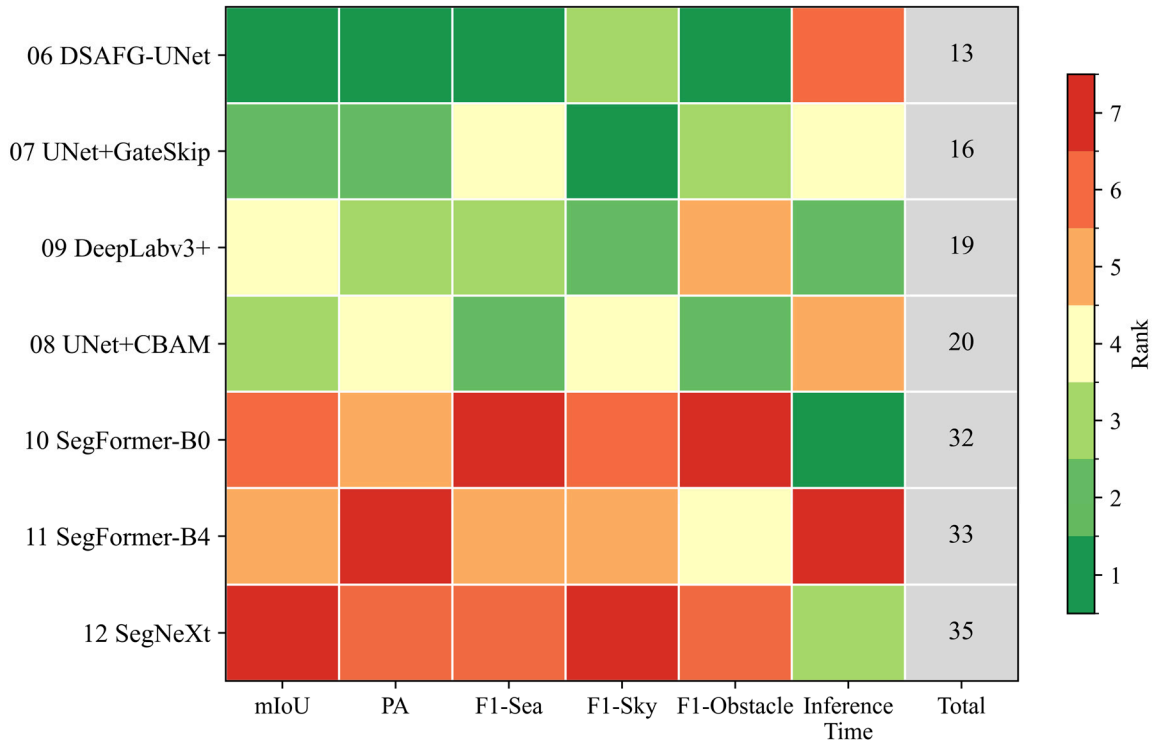


Figure 17. Heat map of metric rankings in the comparative experiments.

4.3. Stability and Generalization

As for model stability, Figure 19 shows that DSAFG-UNet has short error bars on the first five of the accuracy metrics, which indicates that it is reasonably stable while achieving the highest accuracy. In comparison, the error bars of SegFormer-B0, SegFormer-B4, and SegNeXt are clearly longer. Of all metrics, SegFormer-B4 shows the largest standard deviation for the sea-class and obstacle-class F1-Scores, and this model varies most strongly between runs, which is consistent with the collapse of the SegFormer models in the late stage of training. The error bars of UNet+GateSkip, UNet + CBAM, and DeepLabv3+ are of medium length. These results show that attention-based methods are not only less accurate on infrared sea surface segmentation, but also more sensitive to random factors during training. DSAFG-UNet is more stable across runs.

As for model generalization, the external image (e) in Figures 11 and 13 shows that DSAFG-UNet suffers the smallest performance drop and keeps the sea region relatively well connected, which may be related to the ability of its gating mechanism to suppress wave noise. Even when the scene changes, DSAF-Gate can still select effective semantic and edge information through adaptive weights and thus maintain correct recognition of the sea region. On the external image (e), DSAFG-UNet still achieves an F1-Score of 87.87% for the sea class, about 10 percentage points lower than the result of the same run on the dataset, which suggests a certain generalization capability of the model on external samples. Due to the limited number of external images available, this observation should be regarded as preliminary and cannot fully support a general conclusion on generalization.

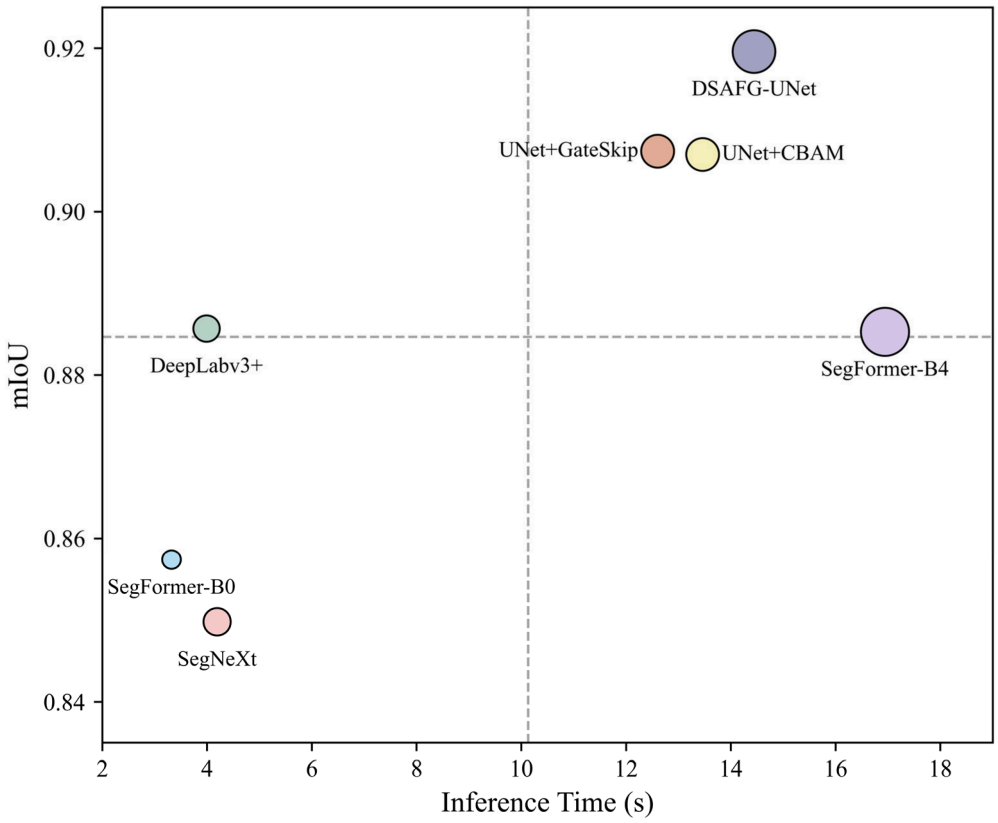


Figure 18. Distribution of the models in terms of accuracy and efficiency.

4.4. Limitations

In terms of the model, first, the predicted obstacle boundaries expand outward in a wrapping manner, and the boundary shape is not fine enough. During training, the model tends to preserve the integrity of the sea region and therefore adopts a conservative decision strategy for obstacle boundaries. Second, the model has no advantage in inference efficiency and runs relatively slowly, which makes deployment difficult in scenarios with strict real-time requirements.

In terms of experimental conditions, first, the dataset is limited in size, containing 2760 infrared images, and lacks scene diversity. Second, the dataset is annotated manually pixel by pixel, so some pixels at transition regions always remain unlabeled, as shown by the semantic label in Figure 9. These pixels do not participate in training, which leaves the model without sufficient supervision near boundaries and also biases the evaluation metrics. Third, the inference time of DSAFG-UNet, 14.44 s, is measured with the hardware configuration described in Section 3.1; since experimental conditions are affected by many factors, this value reflects the efficiency under specific conditions rather than the theoretical upper and lower limits of the model.

In terms of scene coverage, the dataset used in this paper covers common marine scenes such as daytime, nighttime, thin fog, and thick clouds, but contains few samples of extreme weather, and the model performance under such conditions may be affected. However, the practical demand for sea surface detection in extreme weather is usually low, so the impact of this limitation should be weighed against the actual application scenario.

In addition, the dataset was split at the frame level. Although this ensures that every subset covers all scenes, it may introduce a certain degree of information overlap between the training and test sets, which could affect the absolute values of the reported metrics; however, the relative comparison trends remain valid, since all models are trained and evaluated under the same split. This limitation stems from the limited scale and scene

diversity of the available dataset. Moreover, the generalization ability of the model was only verified on a limited number of external images, and a systematic cross-source or cross-dataset evaluation was not conducted due to the limited availability of public infrared sea surface datasets. These will be addressed in future work.

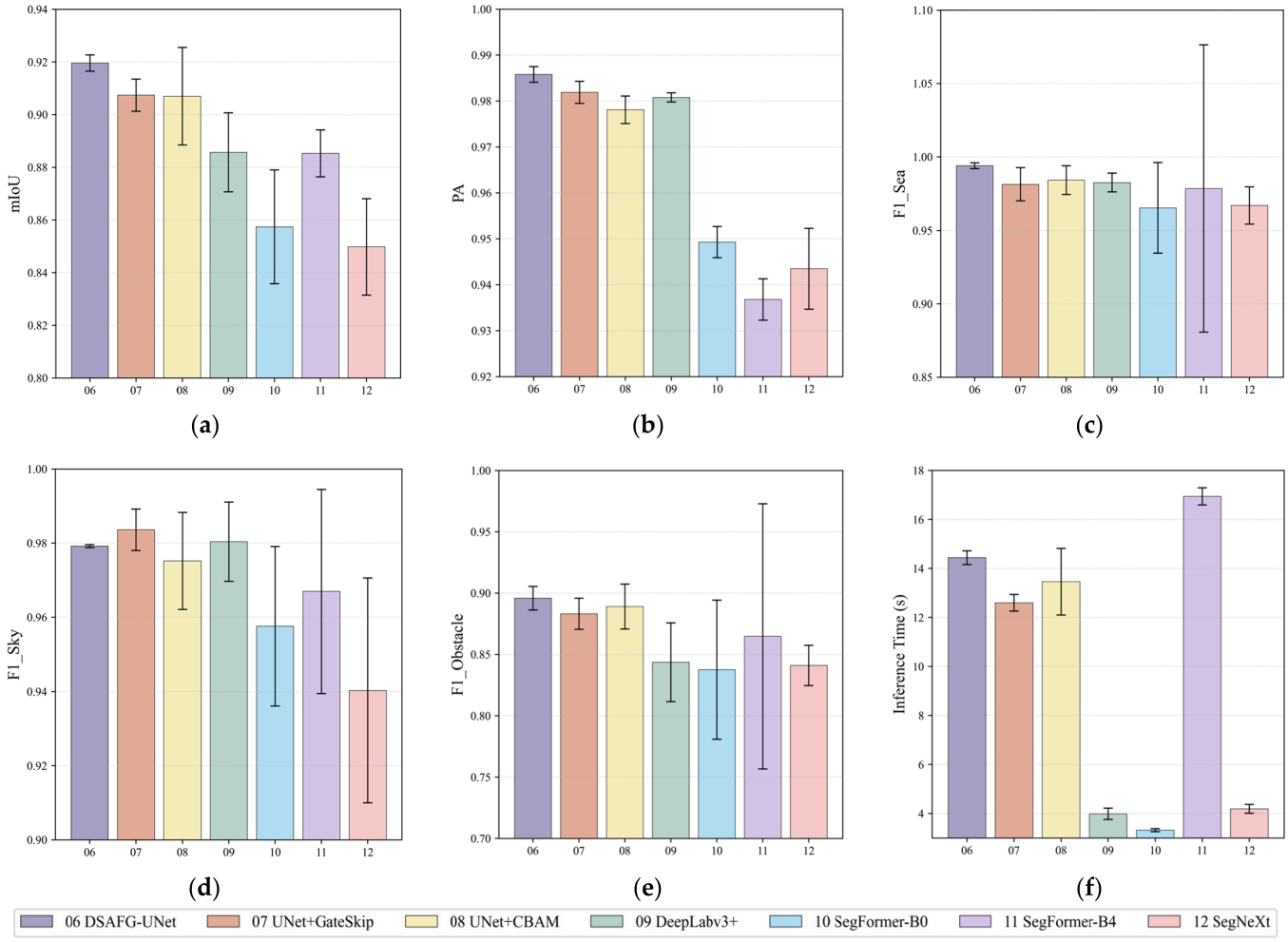


Figure 19. Mean and standard deviation of the models on six metrics: (**a**) mIoU; (**b**) PA; (**c**) F1-Score of the sea class; (**d**) F1-Score of the sky class; (**e**) F1-Score of the obstacle class; (**f**) inference time per image.

5. Conclusions

To address the difficulties of infrared sea surface image segmentation, such as faint wave textures, blurred sea–sky boundaries, and low contrast, as well as the shortcomings of existing methods, whose gating mechanisms mostly adopt a single-path structure and whose auxiliary strategies are mostly stacked independently, this paper proposes DSAFG-UNet. By combining the DSAF-Gate module, the multi-scale edge loss, and the multi-scale context module, the method achieves adaptive filtering of skip-connection features, improved connectivity of the sea region, and enhanced global perception. The study also reveals several limitations of the method, including the wrapping outward expansion of obstacle boundaries, the lack of advantage in inference efficiency, the limited size and annotation quality of the dataset, and the dominance of ordinary sea states in scene coverage.

Future work can proceed in the following directions. For the boundary problem, the design of the edge branch can be further optimized, or finer-grained boundary supervision can be introduced. For the efficiency problem, lightweight design, knowledge distillation,

or model pruning can reduce the number of parameters and the inference time. For the limited size and annotation quality of the dataset, the data scale can be expanded and semi-supervised or weakly supervised annotation strategies can be introduced. For scene coverage, samples of extreme sea states and complex weather can be added to improve the adaptability of the model in diverse environments. In addition, how to better balance accuracy and efficiency is another issue that requires further study.

Author Contributions: Conceptualization, J.B. and Z.W.; methodology, J.B.; software, J.B.; validation, Y.Z.; formal analysis, J.B.; investigation, J.B.; resources, Z.W.; data curation, J.B.; writing—original draft preparation, J.B.; writing—review and editing, L.C.; visualization, K.W.; supervision, Z.W.; project administration, Z.W.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Fund of China, grant number 2022-SKJJ-B-033.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Shandong University Infrared Perception Team and are available at https://lit.qd.sdu.edu.cn/info/1069/1900.htm (accessed on 26 May 2026) or from hwgztd@sdu.edu.cn with the permission of Shandong University Infrared Perception Team.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASPP	Atrous Spatial Pyramid Pooling
CBAM	Convolutional Block Attention Module
CNN	Convolutional Neural Network
DSAF-Gate	Dual-branch Semantic Adaptive Fusion Gate
DSAFG-UNet	Dual-branch Semantic Adaptive Fusion Gate UNet
FCN	Fully Convolutional Network
FLOPs	Floating-Point Operations
GFLOPs	Giga Floating-Point Operations
GPU	Graphics Processing Unit
IoU	Intersection over Union
mIoU	Mean Intersection over Union
MLP	Multilayer Perceptron
PA	Pixel Accuracy
SAM	Segment Anything Model
WCE-Net	Wavelet-domain CNN Enhancement Network

References

1. Deng, C.H.; Li, M.; Deng, K.J.; Zeng, R.F.; Cao, S.F. Infrared image segmentation and target extraction techniques. *Comput. Program. Ski. Maint.* **2025**, *9*, 153–155. [CrossRef]
2. Ma, J.W.; Zou, Y.B.; Lu, X.C. Information energy threshold segmentation method for ship infrared images. *J. Three Gorges Univ. (Nat. Sci.)* **2026**, *48*, 104–112. [CrossRef]
3. Wu, X.J.; Yu, X.Z.; Wang, P.; Zhao, H.; Li, T.C. Superpixel-Based Improved Fuzzy C-Means Clustering for Electrical Equipment Infrared Image Segmentation. *Infrared Technol.* **2025**, *47*, 235–242.
4. Wang, X.; Xu, S.W.; Dong, K.J.; Jin, W.Q. Infrared gas image segmentation method based on background modeling and density clustering. *Infrared Technol.* **2024**, *46*, 1355–1361.

5. Yu, Z.; Wei, Z.; Wang, M.H.; Cui, J.; Tan, J.; Xu, Y. Quantitative evaluation of meibomian gland dysfunction via deep learning-based infrared image segmentation. *Front. Artif. Intell.* **2025**, *8*, 1642361. [CrossRef] [PubMed]
6. Tian, Z.; Liu, H.; Ji, Q.; Wang, S.; Li, Q.; Yi, D.; Liu, X. Near-infrared vascular image segmentation using multi-stage enhancement and TAU-Net. *Infrared Phys. Technol.* **2025**, *148*, 105875. [CrossRef]
7. Ye, L.; Wang, X.; An, A.; Guan, H. DeepLab-YOLO: A method for detecting hot-spot defects in infrared image PV panels by combining segmentation and detection. *J. Real.-Time Image Process.* **2024**, *21*, 52. [CrossRef]
8. Deng, C.Z.; Luo, B.J.; Fu, T.; Gong, M.Q.; Liu, M.Z. Infrared Image Segmentation Algorithm of Power Equipments Based on Improved DeepLabv3+ Algorithm. *Infrared Technol.* **2026**, *48*, 176–183.
9. Ning, X.F.; Li, L.; Wei, W.; Zhang, Y.; Liu, J. Overheating recognition method for composite insulators based on infrared image segmentation. *Laser Infrared* **2025**, *55*, 723–731. [CrossRef]
10. Chen, S.S. Optimization research on infrared image instance segmentation algorithm for switching equipment based on Mask R-CNN model. *Electr. Ind.* **2025**, *7*, 51–55.
11. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Virtual Event, USA, 6–14 December 2021; pp. 12077–12090. [CrossRef]
12. Li, C.; Cai, C.; Zhou, W.; Wu, K. A sea-sky-line detection method for long wave infrared image based on improved Swin transformer. *Infrared Phys. Technol.* **2024**, *138*, 105125. [CrossRef]
13. Zhang, M.; Zhang, C.; Zhang, Q.; Li, Y.; Gao, X.; Zhang, J. Unleashing the power of generic segmentation models: A simple baseline for infrared small target detection. *arXiv* **2024**, arXiv:2409.04714. [CrossRef]
14. Ding, K.; Xiang, S.; Pan, C. SMALNet: Segment anything model aided lightweight network for infrared image segmentation. *Infrared Phys. Technol.* **2024**, *142*, 105540. [CrossRef]
15. Wang, Z.; Liu, L. Infrared image segmentation of power equipment based on improved watershed algorithm. *Infrared Technol.* **2025**, *47*, 484–492.
16. Ru, C.H.; Fan, J.H.; Zhao, Z.; Shen, X.F. Power equipment infrared image segmentation model based on Kapur entropy and improved butterfly optimization algorithm. *J. Comput. Appl. Softw.* **2025**, *42*, 167–176. [CrossRef]
17. Xing, Y.; Chu, D.; Zhang, S.; Xu, D. Temporal-emerged prompting for segment anything in multiframe infrared small target detection. *arXiv* **2026**, arXiv:2606.27655. [CrossRef]
18. Bovcon, B.; Muhovič, J.; Perš, J.; Kristan, M. The MaSTr1325 Dataset for Training Deep USV Obstacle Detection Models. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 3431–3438. [CrossRef]
19. Bovcon, B.; Muhovič, J.; Vranac, D.; Mozetič, D.; Perš, J.; Kristan, M. MODS—A USV-Oriented Object Detection and Obstacle Segmentation Benchmark. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 13403–13418. [CrossRef]
20. Žust, L.; Perš, J.; Kristan, M. LaRS: A Diverse Panoptic Maritime Obstacle Detection Dataset and Benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 20304–20314. [CrossRef]
21. Bovcon, B.; Kristan, M. WaSR—A Water Segmentation and Refinement Maritime Obstacle Detection Network. *IEEE Trans. Cybern.* **2022**, *52*, 12661–12674. [CrossRef] [PubMed]
22. Nirgudkar, S.; DeFilippo, M.; Sacarny, M.; Benjamin, M.; Robinette, P. MassMIND: Massachusetts Maritime INfrared Dataset. *Int. J. Robot. Res.* **2023**, *42*, 21–32. [CrossRef]
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
25. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Granada, Spain, 16–20 September 2018; pp. 421–429. [CrossRef]
26. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519. [CrossRef]
27. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5229–5238. [CrossRef]
28. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 801–818. [CrossRef]

29. Liu, Y.; Zhao, G.; Fan, S.; Fei, C.; Liu, J.; Zhang, Z.; Wang, L.; Li, Y.; Zhao, X.; Liu, Z. Tri-band vehicle and vessel dataset for artificial intelligence research. *Sci. Data* **2025**, *12*, 592. [CrossRef] [PubMed]
30. Wang, L.; Dong, Y.; Fei, C.; Liu, J.; Fan, S.; Liu, Y.; Li, Y.; Liu, Z.; Zhao, X. A lightweight CNN for multi-source infrared ship detection from unmanned marine vehicles. *Heliyon* **2024**, *10*, e26229. [CrossRef] [PubMed]
31. Bu, J.; Wu, Z.; Jiang, Y. Wavelet-domain CNN for wave texture enhancement in infrared sea surface images. In Proceedings of the 2026 International Conference on Image, Signal Processing and Pattern Recognition (ISPP), Guilin, China, 10–12 April 2026; pp. 586–594. [CrossRef]
32. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018, Granada, Spain, 16–20 September 2018; pp. 440–448. [CrossRef]
33. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
34. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; pp. 1140–1156. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.