

Article

DGDE-CSFNet: Difference-Guided Dual-Branch Encoding and Cross-Scale Fusion for Maritime Scene Segmentation

Yutong Yang ¹, Yaxiong Chen ^{1,2}, Changchun Xie ², Sai Zhong ³ and Guirong Feng ^{4,*}

¹ Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China; yytlyx@whut.edu.cn (Y.Y.); chen yaxiong@whut.edu.cn (Y.C.)

² Hubei Key Laboratory of Transportation Internet of Things, School of Artificial Intelligence, Wuhan University of Technology, Wuhan 430072, China; 271621@whut.edu.cn

³ School of Informatics and Mathematics, Yangtze University, Jingzhou 434000, China; zhongs@yangtzeu.edu.cn

⁴ College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China

* Correspondence: guirfeng@fzu.edu.cn

Highlights

What are the main findings?

- A Difference-Guided Dual-Branch Encoding Cross-Fusion Network (DGDE-CSFNet) is developed for sea–land–ship semantic segmentation, integrating CNN and Transformer branches to capture complementary local details and global contextual information.
- The proposed Hierarchical Difference-Guided Attention (HDGA) and Multi-Level Cross-Scale Fusion (MCSF) modules effectively enhance small ship recognition and boundary refinement by exploiting inter-branch differences and multi-scale feature interactions.

What are the implications of the main findings?

- DGDE-CSFNet provides a reliable solution for maritime scene interpretation, enabling more accurate sea–land–ship segmentation under complex conditions such as background interference and irregular object distributions.
- The proposed difference-guided cross-branch learning strategy provides a promising direction for leveraging heterogeneous feature representations in high-resolution remote sensing image segmentation.

Abstract

Semantic segmentation of maritime imagery is challenging due to the coexistence of large homogeneous regions, such as sea and land, and small, sparsely distributed ship targets, particularly in complex port environments. Further, strong background dominance and intricate boundary structures often result in missed detections and inaccurate ship delineation with conventional segmentation frameworks. To address these challenges, this paper proposes DGDE-CSFNet, a Difference-Guided Dual-Branch Cross-Fusion Network with Cross-Scale Feature Fusion specifically designed for sea–land–ship segmentation. The network employs a dual-branch encoder in which convolutional and Transformer backbones operate in parallel to capture local spatial structures and long-range contextual dependencies. To exploit their complementary behaviors, the encoding stage incorporates a Hierarchical Difference-Guided Attention (HDGA) module. By explicitly characterizing inter-branch inconsistencies, the module selectively enhances discriminative responses that are essential for accurate small-object perception. The decoder further adopts a Multi-Level Cross-Scale Fusion (MCSF) to aggregate multi-resolution features, enabling effective interaction between shallow spatial details and deep semantic representations. This design improves boundary localization and structural consistency while suppressing background



Academic Editor: Haopeng Zhang

Received: 25 August 2026

Revised: 2 October 2026

Accepted: 3 October 2026

Published: 9 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

interference and preserving subtle ship details. Experimental results on the ISDSD and HRSC2016-SL datasets demonstrate that our method achieves consistent performance gains over existing state-of-the-art methods.

Keywords: marine remote sensing; sea-land-ship segmentation; semantic segmentation; difference-guided attention mechanism

1. Introduction

Sea–land–ship segmentation is a crucial task in the processing of ocean remote sensing imagery [1,2], requiring accurate pixel-level classification of sea, land, and ship regions. This task has become increasingly significant in a wide range of applications, including maritime security, port management, and ocean monitoring. As a specific pixel-level semantic segmentation task in remote sensing, sea-land-ship segmentation is closely related to general semantic segmentation techniques.

Semantic segmentation of remote sensing images has advanced considerably with the development of deep neural networks. However, challenges such as complex scene structures, high intraclass variability, interclass similarity, and the need for fine spatial detail and global contextual modeling remain central research issues. To address the limited availability of pixel-level annotations, Muhtar et al. [3] proposed IndexNet, a self-supervised dense representation learning framework that incorporates object spatial position cues and jointly performs image-level and pixel-level contrastive learning, demonstrating improved generalization over conventional ImageNet-pretrained models. Beyond representation learning, attention mechanisms have been widely explored to enhance feature discrimination. Li et al. [4] identified attention bias when spatial and channel affinities are modeled independently, and introduced a synergistic attention perception network that jointly captures channel-wise and spatial contextual relationships to better segment weak and sparsely distributed objects. To further improve contextual reasoning under varying object scales, Chen et al. [5] developed an adaptive attention network that integrates multiscale channel attention, multidimensional spatial attention, and contextual fusion to enhance feature expressiveness while maintaining computational efficiency. Chen et al. [6] developed a semantic information-guided global–local fusion framework for UAV image small-object detection, which enhances small target representation by integrating semantic and spatial features. Chen et al. [7] developed MCSNet, which enhances remote sensing image segmentation by jointly modeling global context and spatial details through multilevel feature fusion. Lightweight design has also gained traction. Yue et al. [8] proposed Lite-Seger, which combines MobileViT with global and local attention refinement, achieving competitive accuracy with significantly reduced parameters and high inference speed. Meanwhile, Li et al. [9] introduced GLVMamba, a global–local visual state-space model that enhances global context reasoning and edge contour refinement through state-space modeling and multiscale pyramid pooling. Incorporating domain knowledge has also proven effective. Zhao et al. [10] constructed a geographical scenario knowledge graph and designed a graph structure attention network to leverage semantic neighborhood relationships, improving differentiation among visually similar ground objects. Furthermore, Gao et al. [11] proposed an adaptive frequency enhancement network that modulates spatial and frequency-domain features to better handle diverse land-cover patterns. Although these approaches have improved remote sensing semantic segmentation from different perspectives, effectively balancing fine-grained spatial details and global semantic understanding remains challenging, especially in complex maritime scenes.

Within the broader context of remote sensing semantic segmentation, maritime scene understanding represents a challenging application scenario due to large homogeneous regions, irregular boundaries, and complex interactions between targets and backgrounds. Unlike conventional land-cover segmentation, maritime scenes require accurate discrimination among different semantic categories, particularly small ship targets embedded in complex coastal environments. Ship detection and segmentation in remote sensing imagery have been widely studied in optical, SAR, and infrared scenarios. For SAR images, Wei et al. [12] introduced LFG-Net to enhance instance-level ship segmentation by incorporating low-level texture cues and high-resolution mask refinement, improving small ship delineation under speckle noise. To enhance fine-grained ship recognition, Ouyang et al. [13] proposed MGANet, which aligns local part features and global object features through multigranularity self-attention, enabling better detection of densely distributed and visually similar ships. Beyond detection accuracy, interpretability has also been explored. Xiong et al. [14] designed an explainable attention network that highlights discriminative structural components of ships, providing visual reasoning without degrading classification performance. For modality gaps across visible and infrared imaging, Zhang et al. [15] presented a two-stage domain adaptation segmentation framework to improve infrared ship segmentation where pixel-level labels are limited. SEEvIF-Net [16] introduces an evidence-theoretic semi-supervised framework that fuses probabilistic and evidential information via Dempster's rule, explicitly quantifying epistemic uncertainty to achieve robust ship segmentation with limited labeled data. In SAR-based ship detection, Chen et al. [17] exploited polarization diversity through dual-polarization feature fusion to enhance robustness in cluttered ocean scenes. Recent methods further combine CNN and Transformer architectures to capture both detailed and global context. For instance, MLANet [18] employs multi-level feature enhancement and attention fusion for robust segmentation in complex coastal environments, while CGA-Det [19] integrates CNN and GNN to model structural relations among densely packed ships. More recently, ShipMS-BSNet [20] tackles multi-scale ship segmentation using a large-scale dataset and a novel nnU-Net-based network that enhances multi-scale context, suppresses background interference via channel attention, and refines target boundaries at the decoder. Similarly, MPFT-UNet [21] enhances UAV-based ship segmentation through a multi-scale dynamic sparse cross-gating mechanism for small targets, a boundary refinement module for edge accuracy, and Transformer-based feature fusion for global context. With the recent emergence of vision foundation models, Yang et al. [22] further extended SAM 2 to UAV-based ship segmentation through a training-free constrained boundary enhancement strategy, where weak contours are reinforced, and candidate boundaries are refined through prompt consensus without additional fine-tuning.

Recent studies have demonstrated that hybrid CNN-Transformer architectures can effectively combine local spatial representation and global contextual modeling for remote sensing semantic segmentation. CNN-based networks are effective in extracting fine-grained spatial details, whereas Transformers provide stronger capabilities in capturing long-range dependencies. Chen et al. [23] proposed CTFuse, which integrates CNN and Transformer structures to exploit local feature extraction and global semantic modeling capabilities. Similarly, Wu et al. [24] developed CMTFNet by combining CNN-based local information extraction with multiscale Transformer-based global context modeling, achieving improved semantic representation for high-resolution remote sensing images. More recently, dual-branch CNN-Transformer architectures have been explored to further enhance feature complementarity. Xiang et al. [25] introduced a parallel CNN-Transformer framework, where features from both branches are interactively fused to capture spatial details and global contextual information simultaneously. Zhu et al. [26] proposed IC-

TransUNet, which employs CNN and Transformer branches as dual backbones to extract detailed and semantic representations, respectively. Yang et al. [27] further designed a dual-branch encoder consisting of a CNN branch and a Shunted Transformer branch, together with multi-scale feature enhancement mechanisms to improve remote sensing image segmentation. Ji et al. [28] developed DBRSNet, which enhances CNN-Transformer feature interaction through adaptive selection and multi-scale fusion strategies. For maritime scene understanding, Zhang et al. [29] proposed a spatial-semantic alignment fusion network based on parallel Transformer-CNN encoders for sea-land port segmentation. Although these methods have achieved promising performance by exploiting complementary information between CNN and Transformer branches, they mainly focus on feature fusion, alignment, or interaction. The intrinsic discrepancy between heterogeneous CNN and Transformer representations, which may contain valuable discriminative information for complex scenes, remains insufficiently explored. However, whether such discrepancy-aware feature learning can effectively improve sea-land-ship segmentation under complex maritime environments remains insufficiently investigated.

Most existing studies have focused either on binary sea-land segmentation or on ship target extraction alone, while research addressing the joint pixel-level classification of sea, land, and ship targets remains relatively limited. Moreover, current sea-land segmentation algorithms still face difficulties in complex maritime scenes, where the presence of large-scale sea and land areas often leads to suboptimal segmentation of small ship targets. These limitations indicate that joint sea-land-ship segmentation remains challenging in complex maritime scenes. As illustrated in Figure 1a, ship segmentation in sea-land port environments is challenged by complex backgrounds comprising diverse man-made structures such as docks, containers, and buildings. Many of these objects exhibit high visual similarity to ships in terms of color, texture, and shape, which can easily lead to confusion during segmentation and increase the risk of both false positives and missed detections. As shown in Figure 1b, when ships are moored at ports, they are often tightly adjacent to docks, resulting in visual overlaps or occlusions that blur object boundaries and complicate pixel-level differentiation, which further intensifies the segmentation difficulty. Moreover, Figure 1c highlights the inherent scale imbalance in this task where ships typically occupy much smaller areas compared to dominant sea and land regions, and their sizes vary considerably, from small fishing vessels to large cargo ships. This significant scale imbalance makes it difficult to preserve discriminative representations of small ships while maintaining accurate segmentation of dominant background regions.

To address the above challenges, we propose DGDE-CSFNet, a Difference-Guided Dual-Branch Cross-Fusion Network for sea-land-ship segmentation. Unlike existing CNN-Transformer fusion approaches that mainly aggregate complementary features, DGDE-CSFNet explicitly models hierarchical differences between CNN and Transformer branches through the Hierarchical Difference-Guided Attention (HDGA) module, enabling more discriminative feature learning for small ship targets and complex backgrounds. Furthermore, the Multi-Level Cross-Scale Fusion (MCSF) module integrates multi-level semantic and spatial information to improve boundary preservation and scale-aware representation. The main contributions of this work are summarized as follows:

- We propose DGDE-CSFNet, a Difference-Guided dual-branch Cross-Fusion network that integrates hierarchical difference-guided attention with multi-scale feature fusion for accurate sea-land-ship segmentation in complex maritime scenes.
- We introduce an HDGA module to fuse local-global features via cross-branch difference modeling, and a MCSF module to progressively integrate semantics and spatial details for fine-grained ship boundary segmentation.

- Experimental results on the ISDSD and HRSC2016-SL port scene sea-land-ship segmentation datasets demonstrate that our method achieves better segmentation precision than other state-of-the-art approaches.

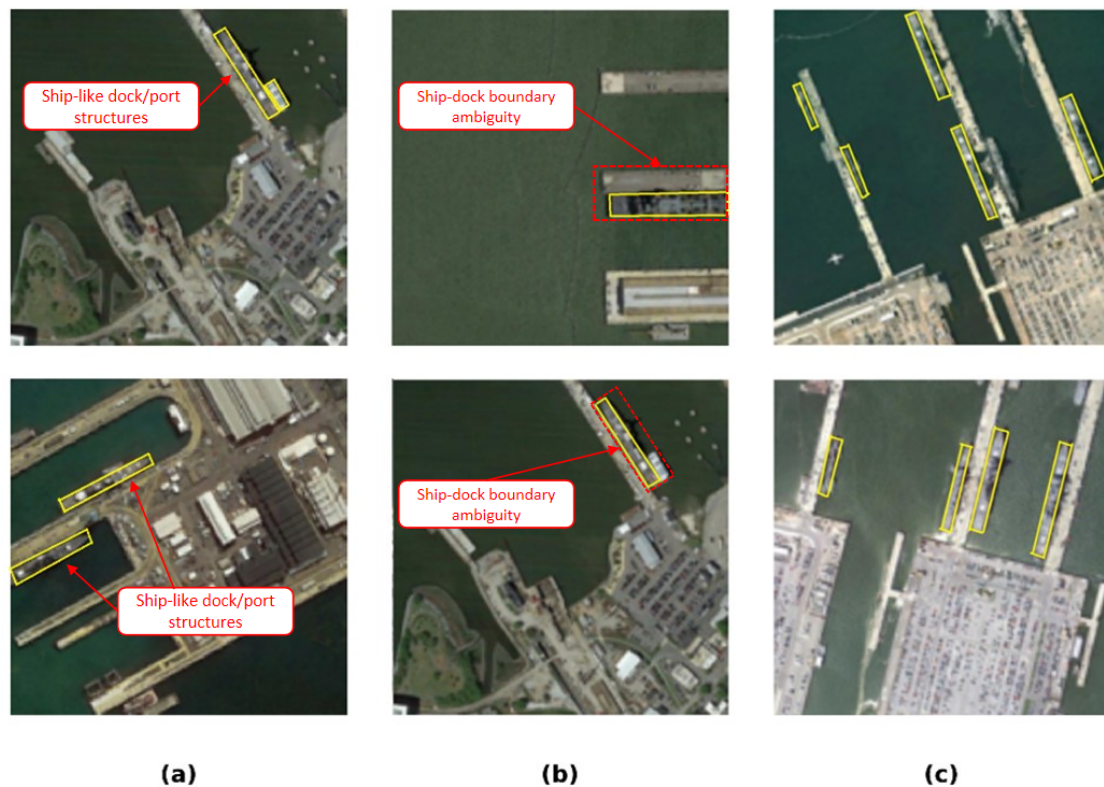


Figure 1. Challenges of joint sea-land-ship segmentation in complex maritime scenes. The yellow boxes highlight representative ship targets that are challenging to segment in complex maritime scenes. (a) Example of a port area, featuring various objects such as docks, containers, and buildings. Some of these objects share high similarity with ships in terms of color, texture, and shape. (b) Image of ships docked at the port, where ships are often closely adjacent to docks, resulting in visual overlap and even occlusion between ship and dock boundaries. (c) Distribution of ship targets. Compared to the vast ocean and land areas, ship targets generally occupy smaller regions and appear at a relatively smaller scale.

2. Materials and Methods

2.1. Overall Architecture

The core of the proposed DGDE-CSFNet is to categorize each pixel in a given image and then correspond it to the predefined objects of sea area, land area, and ship targets in our task. As illustrated in Figure 2, the model follows an encoder-decoder paradigm. The encoder adopts a dual-branch architecture that integrates CNN and Transformer backbones to extract both local and global features, which are subsequently fused through the proposed HDGA module. The decoder then progressively integrates multi-scale features via the MCSF module, followed by layer-wise upsampling and regression operations to generate the final segmentation map.

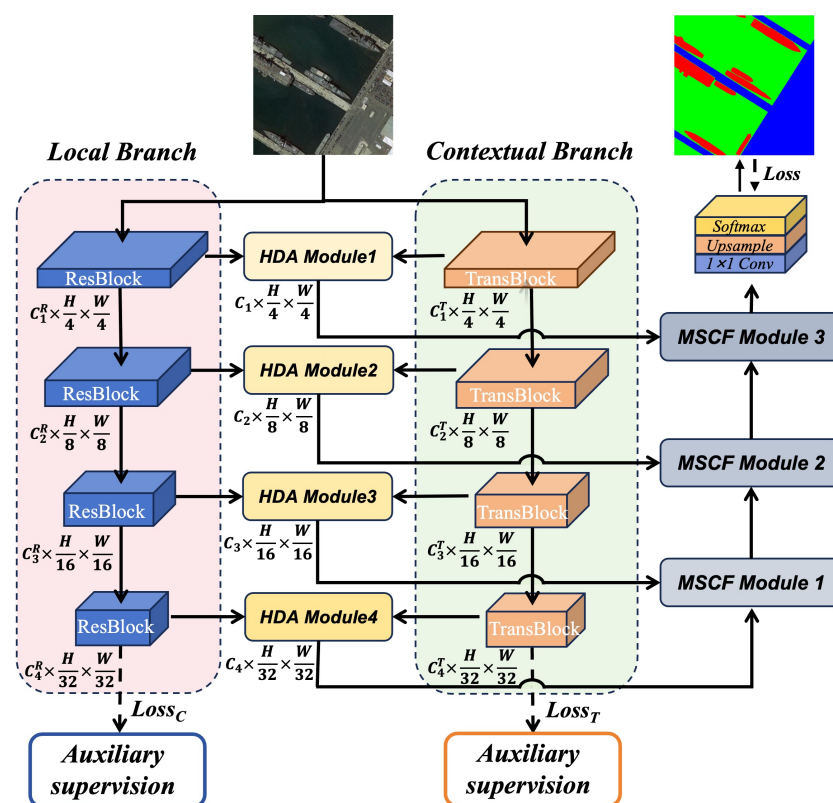


Figure 2. Overview of the proposed DGDE-CSFNet framework. The model follows an encoder–decoder architecture. The encoder adopts a dual-branch design, combining CNN and Transformer backbones to extract local details and global contextual features, respectively. A HDGA module is employed to adaptively integrate multi-level features from both branches. The final representations from each backbone are fed into corresponding FCN and SegFormer decoders to compute auxiliary supervision. In the decoding stage, an MSCF module is utilized to fuse features from different encoder levels, enhancing the model’s ability to segment small-scale ship targets under complex maritime conditions.

2.2. Dual Branch Encoder

The proposed framework employs a dual-branch encoder structure that synergistically combines a ResNet50 CNN backbone [30] and a SegFormer Transformer backbone [31] for complementary feature extraction. Given an input image $F \in \mathbb{R}^{C \times H \times W}$, where C , H , and W represent the channel, height, and width dimensions, respectively, the network processes features through four hierarchical stages:

ResNet Branch stage outputs: $R_i \in \mathbb{R}^{C_i^R \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$, SegFormer Branch stage outputs: $T_i \in \mathbb{R}^{C_i^T \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ for $i \in 1, 2, 3, 4$. C_i^R and C_i^T represent the number of feature channels extracted at each stage of the ResNet and SegFormer backbones, respectively. Following feature extraction, the HDGA module receives multi-stage features from the two backbones above as input and leverages the complementary characteristics between local and global features to enhance the extraction and fusion of discriminative features. The output of the HDGA module at each stage is represented as $F_D \in \mathbb{R}^{C_i^D \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ for $i \in 1, 2, 3, 4$, where C_i^D indicates the channel dimension of the fused features. Finally, the outputs from all four HDGA stages are fed into the decoder network for hierarchical feature refinement and final prediction generation.

Additionally, the final stage features extracted from both backbone networks are independently fed into their corresponding decoder heads: FCN decoder for CNN features and SegFormer decoder for Transformer features, for the computation of auxiliary supervision losses during training. The implementation of dual-branch auxiliary loss supervision

provides three key benefits: (1) it strengthens gradient flow throughout the entire network architecture during backpropagation; (2) through deep supervision, potential information degradation is mitigated; (3) it enhances overall training stability and convergence properties. Importantly, this mechanism ensures that both the CNN and Transformer branches maintain their distinctive feature extraction capabilities while simultaneously optimizing their collaborative performance for the final fused segmentation output.

2.3. Hierarchical Difference-Guided Attention Module

The HDGA module captures inter-feature differences and employs an attention mechanism to enhance the adaptability of the model to complex scenarios, thereby overcoming the limited receptive field of CNNs and the local detail neglect of Transformers. Figure 3 illustrates the overall structure of the proposed HDGA module.

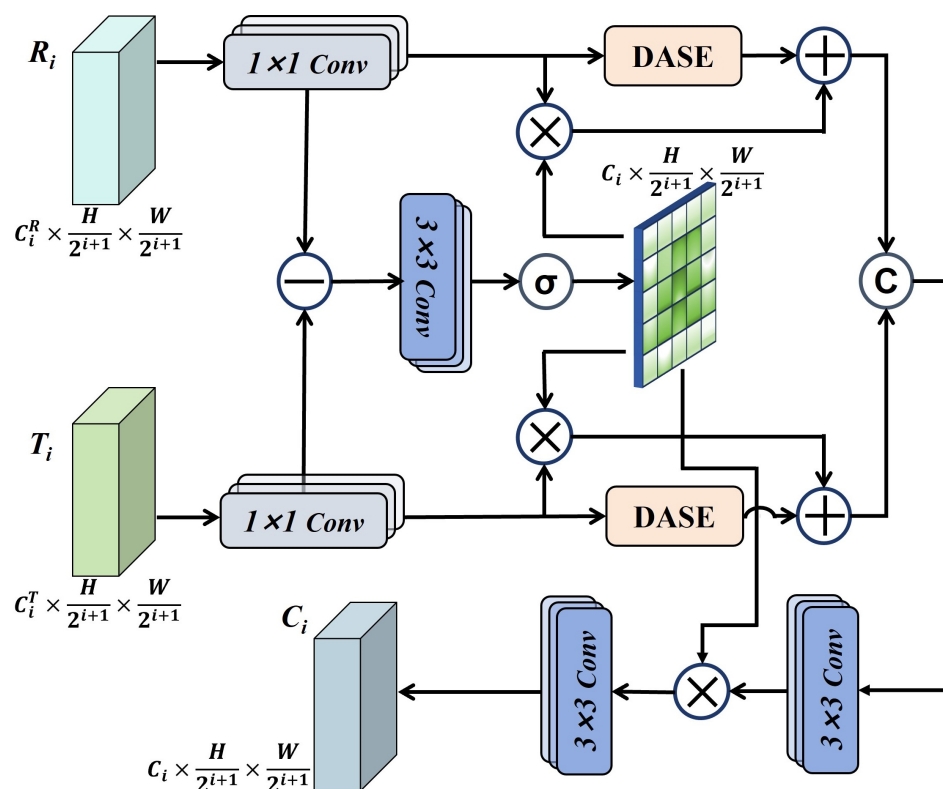


Figure 3. Architecture of the HDGA module. The CNN and Transformer features are first channel-aligned, and their differences are compared to generate an attention map. This map is then applied to the original features via residual connections, which are further refined by a Dynamic Adaptive squeeze-and-excitation mechanism and subsequently concatenated for convolutional integration.

2.3.1. Dynamic Adaptive SE Block

The Dynamic Adaptive SE (DASE) block is designed to dynamically capture feature dependencies between channels. Fundamentally, the Squeeze-and-Excitation (SE) block [32] enhances feature representation by explicitly modeling channel-wise dependencies through an attention mechanism; this process can be formulated as

$$Z_c = \text{Squeeze}(X_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_c(i, j), \quad (1)$$

where X_c represents the input feature map for the c -th channel; $X_c(i, j)$ denotes the pixel value at spatial location (i, j) of the c -th channel; and Z_c corresponds to the c -th element of the global channel descriptor vector Z .

Specifically, an adaptive recalibration process is performed to comprehensively capture channel-wise dependencies, namely the Dynamic Adaptive SE (DASE) block. This excitation operation employs two fully-connected layers: the first reduces channel dimensionality from C to C/r through compression, while the second restores the original channel dimension. The complete process can be formally expressed as

$$s = \text{Excitation}(Z_c) = \sigma(W_2 \delta(W_1 Z)), \quad (2)$$

where W_1 and W_2 represent the weight matrices of the two fully connected layers, respectively; $\delta(\cdot)$ denotes the ReLU activation function; $\sigma(\cdot)$ indicates the Sigmoid activation function; and the output s corresponds to the learned channel attention weights. This formulation enables the network to dynamically emphasize informative features while suppressing less useful ones through channel-wise feature recalibration.

The final step involves a channel-wise rescaling operation, where the learned attention weights are applied to the original feature maps through element-wise multiplication. This reweighting process is formally expressed as

$$\text{Output} = \text{Scale}(X, s) = X \otimes s, \quad (3)$$

where $\otimes$ denotes channel-wise multiplication between the input feature map X and the attention weight matrix s .

2.3.2. Hierarchical Difference-Guided Attention Process

As shown in Figure 3, the hierarchical difference-guided attention process operates on input features $R_i \in \mathbb{R}^{C_i^R \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ from the ResNet backbone and $T_i \in \mathbb{R}^{C_i^T \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ from the Transformer backbone at stage i , and the processes are as follows:

Initially, outputs from two encoder backbones are fed into the module to model the distinctive characteristics between the two complementary information sources. The processes are as follows:

$$D_i = |R'_i - T'_i| = |\text{proj}_{c1}(R_i) - \text{proj}_{c2}(T_i)|, \quad (4)$$

where R_i and T_i represent the input features from the ResNet and Transformer branches, respectively; proj_{c1} and proj_{c2} are two 1×1 convolutions with different parameter weights. R'_i and T'_i are intermediate outputs of the R_i and T_i transformed by proj_{c1} and proj_{c2} operations, respectively. $|\cdot|$ indicates the absolute value operation.

Then, the above difference feature D_i is further processed to get a differential attention weight that measures divergence scores between the local and context information on the overall feature map, which can be formulated as

$$w = \sigma(\text{proj}_c(D_i)), \quad (5)$$

where σ is the sigmoid activation function, and proj_c is a 3×3 convolution operation.

Next, a refinement operation is performed on the transformed features of the two branches to further enhance their feature representations. Specifically, the Dynamic Adaptive SE (DASE) block is introduced to adaptively recalibrate the channel responses of each branch, enabling the network to emphasize informative features while suppressing redundant responses. The refinement process is formulated as

$$R''_i = w \odot R'_i + \text{DASE}(R'_i), \quad (6)$$

$$T''_i = w \odot T'_i + \text{DASE}(T'_i), \quad (7)$$

where R'_i and T'_i denote the transformed feature representations of the CNN and Transformer branches at the i -th stage, respectively. R''_i and T''_i represent the refined feature representations after adaptive enhancement. $\text{DASE}(\cdot)$ denotes the Dynamic Adaptive SE block, which generates adaptive channel-wise attention weights to recalibrate the input features. w is a learnable weighting coefficient that controls the contribution of the original transformed features, and $\odot$ represents the element-wise multiplication operation.

To further enhance the joint representation capability of the two branches, a fusion operation is performed on the two refined features, as formally defined below:

$$C'_i = F_{rt}([R''_i : T''_i]), \quad (8)$$

where $[\cdot]$ is a concatenation operation, and F_{rt} denotes a 3×3 convolution operation.

Finally, to reinforce the overall discriminability of the fused representation, the final output of the HDGA module is obtained as follows

$$C_i = \text{proj}_{c'}(w \odot C'_i), \quad (9)$$

where $\odot$ denotes element-wise multiplication, and $\text{proj}_{c'}$ is a 3×3 convolution operation.

2.4. Multi-Level Cross-Scale Fusion Module

Figure 4 illustrates the architecture of the proposed MCSF module, which employs a scale attention mechanism for feature integration. The module processes two input features from adjacent encoder stages: $C_i \in \mathbb{R}^{C_i^T \times H \times W}$ (lateral connection preserving spatial details) and $C_{i+1} \in \mathbb{R}^{C_{i+1}^T \times H \times W}$ (deeper features with advanced semantics), where C_{i+1} has been upsampled by a factor of two via bilinear interpolation to match spatial dimensions.

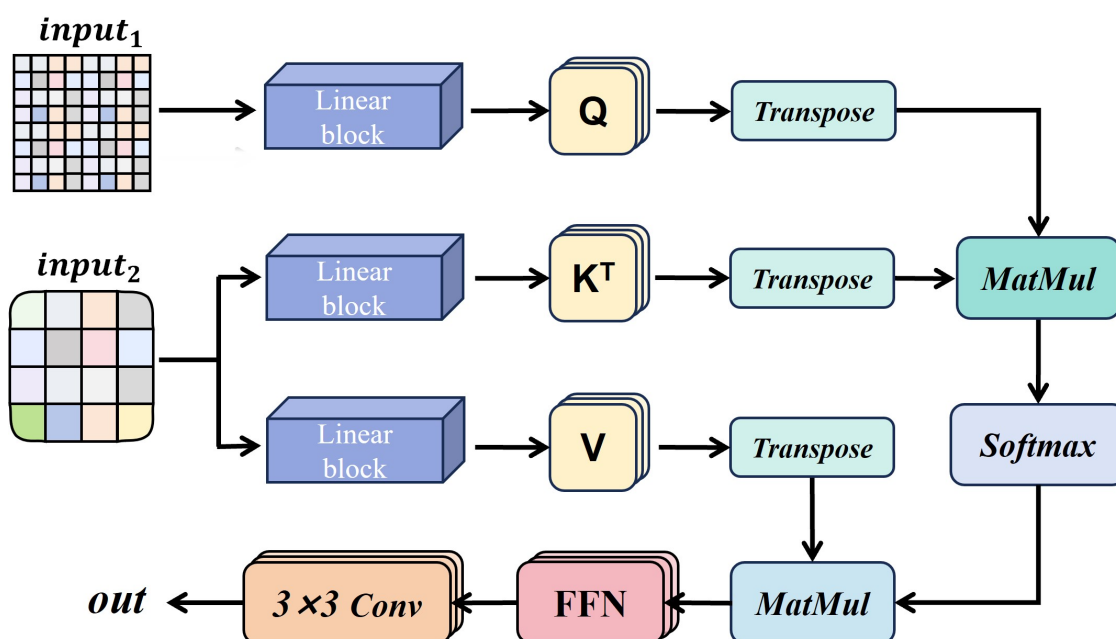


Figure 4. MCSF module. Linear projections and transpose operations are applied to the input features, followed by two matrix multiplications and a feed forward network, with a convolution applied before producing the output.

The fusion process proceeds as follows: First, two input features from adjacent encoder stages are projected through three linear layers to obtain query (Q), key (K), and value (V) matrices. This process can be formulated as follows:

$$Q = W_q C_i, \quad K = W_k C_{i+1}, \quad V = W_v C_{i+1}, \quad (10)$$

where W_q , W_k , and W_v are linear projection parameters.

Subsequently, to obtain the similarity scores between the query and key, a dot-product operation is performed to measure their correspondence. This process can be mathematically represented as

$$\text{attn} = \text{Cross}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \quad (11)$$

where d is the attention head dimension; $(\cdot)^T$ represents the transpose operation; and $\text{Softmax}(\cdot)$ denotes the softmax activation function.

The similarity scores are multiplied with the dimensionally transformed value matrix V to produce the scale-fused feature representation Z . This operation can be formally expressed as

$$Z = \text{attn} \cdot V, \quad (12)$$

At last, the scale-fused feature Z is further processed for local enhancement. The process can be mathematically represented as

$$E_i = \text{Re}(\text{FFN}(Z)), \quad (13)$$

where Re denotes a 3×3 convolution operation, and $\text{FFN}(\cdot)$ represents a feed-forward network (FFN) with ReLU activation. This module effectively combines shallow spatial details with deep semantic information while maintaining dimensional consistency.

2.5. Objective Function

The final segmentation loss $\mathcal{L}$ is formulated as a composite objective consisting of the cross-entropy loss $\mathcal{L}_{ce}$ and the focal loss $\mathcal{L}_{focal}$ [33]. In addition to the final segmentation loss, the proposed sea-land-ship segmentation model employs two auxiliary decoder heads to compute auxiliary loss terms. Auxiliary losses $\mathcal{L}_c$ and $\mathcal{L}_T$ are generated by the FCN and SegFormer decoder heads, respectively, both based on standard cross-entropy loss. The overall loss function consists of two components: the final segmentation loss and auxiliary losses. The final segmentation loss combines both cross-entropy loss and focal Loss, while the auxiliary losses comprise the outputs from two auxiliary decoder heads, both computed using cross-entropy loss. This can be formally expressed as

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{Focal}, \quad (14)$$

$$\mathcal{L}_{aux} = \mathcal{L}_C + \mathcal{L}_T, \quad (15)$$

$$\mathcal{L}_{total} = \mathcal{L} + \theta \cdot \mathcal{L}_{aux}, \quad (16)$$

where β is a weighting coefficient empirically set to 1, and θ is another weighting coefficient balancing the influence between main and auxiliary losses, set to 0.4 in this work. $\mathcal{L}_{CE}$ and $\mathcal{L}_{Focal}$ denote the cross-entropy loss and focal loss computed by the model's final predictions, while $\mathcal{L}_C$ and $\mathcal{L}_T$ represent the auxiliary losses from the FCN decoder head and SegFormer decoder head, respectively.

2.6. Datasets and Evaluation Metrics

Extensive experiments were conducted on two datasets used for sea-land-ship segmentation: the ISDSD dataset [34] and the HRSC2016-SL dataset derived from HRSC2016 [35].

The ISDSD dataset contains 200 high-resolution remote sensing images collected from multiple port areas worldwide from Google Earth imagery. The images have a spatial resolution of approximately 1 m, and their spatial dimensions range from approximately

1000 × 1000 to 3000 × 3000 pixels. Since the original ISDSD data do not directly provide the pixel-level sea-land-ship semantic segmentation masks used in this study, we manually annotated the original images to construct the corresponding three-class semantic segmentation dataset, where the semantic categories are sea, land, and ship. To avoid information leakage, the 200 original images were first randomly divided at the image level into 140 training images and 60 validation images. No independent test set was used in our experiments. After the image-level split, patch extraction and data augmentation, including random cropping, rotation, and flipping, were independently performed on the training and validation subsets. This process produced 947 training samples and 252 validation samples with an image size of 512 × 512 pixels. Since the data partition was performed before augmentation, augmented samples derived from the same original image do not appear simultaneously in the training and validation subsets.

HRSC2016 is a high-resolution optical remote sensing dataset originally developed for ship recognition and detection [35]. The images were collected from Google Earth and cover several well-known port and maritime areas under different acquisition conditions. The original HRSC2016 benchmark does not provide pixel-level sea-land-ship semantic segmentation annotations. In this study, we construct a sea-land-ship semantic segmentation dataset, termed HRSC2016-SL, based on 443 selected HRSC2016 images. The pixel-level semantic segmentation masks were manually annotated by us, with three semantic categories, namely sea, land, and ship. For overlapping regions between ship and land, the ship category is assigned priority to preserve the complete structure of ship targets and maintain consistent semantic labeling. The 443 images were first divided into training and validation subsets at the original image level, and no independent test set was used. Patch extraction and data augmentation, including random cropping, rotation, and flipping, were subsequently applied to the two subsets independently. After preprocessing, 1912 training samples and 292 validation samples with an image size of 512 × 512 pixels were obtained.

To prevent image-level and spatial data leakage, the 444 original images used for HRSC2016-SL were repartitioned into training, validation, and test subsets before patch generation. Image patches of 512 × 512 pixels were then generated independently within each subset. All patches originating from the same original image were assigned exclusively to one subset. Data augmentation was applied only to the training subset, whereas the validation and test subsets were kept unchanged.

To evaluate segmentation performance, four widely used metrics were adopted: F1-score, overall accuracy (OA), per-class Intersection over Union (IoU), and mean Intersection over Union (mIoU). A comprehensive analysis of these metrics enables an objective and thorough evaluation of the effectiveness of the proposed DGDE-CSFNet Network.

2.7. Experiment Details

The proposed DGDE-CSFNet network was implemented using the PyTorch 1.12.0 framework and trained on a Linux server equipped with an NVIDIA Quadro RTX 5000 GPU. Extensive comparative experiments were conducted on two publicly available port scene segmentation datasets, with a batch size of 4 for both training and validation, and all models trained for 100 epochs. During training, various data augmentation techniques such as random rotation, random flipping, and photometric distortion were applied to enhance generalization. The AdamW optimizer was employed for parameter optimization, with an initial learning rate of 0.0001, momentum parameters set to (0.9, 0.999), and a weight decay of 0.001. A PolyLR learning rate schedule was adopted, with the power parameter controlling the decay curve set to 0.9. To ensure the reliability of the reported results, all experiments were repeated three times under identical experimental settings,

and the values presented in the tables were the averaged results over these repeated runs rather than those from a single run.

3. Results

3.1. Methods Comparison

Extensive experiments are conducted on the two aforementioned sea–land–ship segmentation datasets to validate the effectiveness of the proposed DGDE-CSFNet method, along with comparisons against eight representative baseline approaches:

- FCN [36]: Fully Convolutional Network (FCN) is a classical semantic segmentation model that replaces fully connected layers with convolutional layers to enable end-to-end dense prediction.
- DeepLabV3+ [37]: DeepLabV3+ enhances DeepLabV3 by adding a decoder module to refine segmentation boundaries and applies depthwise separable convolutions for faster and more effective encoding and decoding.
- CCNet [38]: CCNet introduces a criss-cross attention module to aggregate contextual information along horizontal and vertical paths for each pixel. A recurrent mechanism further enhances global feature capture while maintaining low memory and computation costs.
- SegFormer [31]: SegFormer integrates a hierarchical Transformer encoder with a lightweight MLP decoder, capturing multiscale features without positional encoding. Its simple yet effective design balances local and global attention, enabling efficient and accurate semantic segmentation.
- ST-UNet [39]: ST-UNet combines a Swin transformer and CNN in parallel, using modules to enhance spatial detail and fuse global and local features for improved segmentation.
- CTFuse [23]: CTFuse is a hybrid CNN-Transformer model that combines multi-scale convolutional attention with global context modeling. It enhances both local and global feature extraction through spatial and channel attention modules in both the CNN and Transformer components.
- CTCFNet [40]: CTCFNet combines CNN and PVT blocks to extract complementary local and global features. A feature aggregation module and bi-directional decoder are used to enhance multiscale fusion and boundary precision.
- AENet [41]: AENet adopts a dual-branch architecture with SCAM and SAM modules to enhance fine-detail and edge feature extraction. Edge features are fused with segmentation features, and edge loss is used to refine boundary segmentation.

Table 1 presents the comparative results on the ISDSD and HRSC2016-SL datasets. On the ISDSD dataset, the proposed DGDE-CSFNet model consistently surpasses all baseline methods across all evaluation metrics. Specifically, for the ship category, DGDE-CSFNet achieves an IoU of 86.64%, outperforming SegFormer by 1.66% and exceeding ST-UNet, CTFuse, and CTCFNet by 3.35%, 4.18%, and 2.82%, respectively. Compared to AENet, the improvement reaches 1.86%. Regarding overall segmentation performance, DGDE-CSFNet attains 97.09% in F1 score, 94.51% in mIoU, and 99.05% in OA, demonstrating consistent superiority over all other baselines. On the HRSC2016-SL dataset, DGDE-CSFNet again delivers superior performance, outperforming all baselines in all metrics except for the IoU of the sea class. For ship segmentation, DGDE-CSFNet achieves 87.36% IoU, with gains of 1.11%, 1.59%, 1.76%, 1.04%, and 1.82% over SegFormer, ST-UNet, CTFuse, CTCFNet, and AENet, respectively. In terms of overall metrics, DGDE-CSFNet obtains 96.56% F1, 93.46% mIoU, and 98.39% OA, showing consistent improvements across all baselines.

Table 1. The experimental comparison results between DGDE-CSFNet and compared state-of-the-art methods on the ISDSD and HRSC2016-SL datasets.

Method	ISDSD Dataset						HRSC2016-SL Dataset					
	Class IoU(%)			Overall Metrics			Class IoU(%)			Overall Metrics		
	Sea	Land	Ship	F1 (%)	mIoU (%)	OA (%)	Sea	Land	Ship	F1 (%)	mIoU (%)	OA (%)
FCN [36]	97.09	95.99	74.41	93.49	89.17	97.94	98.29	91.53	82.27	95.0	90.7	97.69
DeepLabv3+ [37]	98.36	97.79	80.58	95.77	92.24	98.78	98.58	92.29	83.35	95.4	91.41	97.94
CCNet [38]	98.3	97.88	82.29	96.12	92.82	98.81	98.41	92.4	83.75	95.47	91.52	97.91
SegFormer [31]	98.42	98.09	84.98	96.71	93.83	98.93	98.7	93.74	86.25	96.25	92.9	98.27
ST-UNet [39]	98.29	97.89	83.29	96.32	93.16	98.83	98.73	93.65	85.77	96.14	92.71	98.25
CTFuse [23]	98.49	98.09	82.46	96.22	93.01	98.91	98.74	93.61	85.6	96.1	92.65	98.24
CTCFNet [40]	98.54	98.2	83.82	96.52	93.52	98.97	98.74	93.93	86.32	96.3	93.0	98.31
AENet [41]	98.43	98.1	84.78	96.67	93.77	98.93	98.72	93.45	85.54	96.06	92.57	98.22
DGDE-CSFNet	98.57	98.33	86.64	97.09	94.51	99.05	98.73	94.28	87.36	96.56	93.46	98.39

Bold values indicate the maximum value of each indicator.

3.2. Qualitative Analysis

As illustrated in Figure 5, visual comparisons are conducted using several representative baseline models on the ISDSD dataset and the HRSC2016-SL dataset, respectively, with black boxes highlighting the key regions of interest. Specifically, (a) shows the original image, (b) the ground truth annotation, (c) the output of FCN, (d) DeepLabv3+, (e) CCNet, (f) SegFormer, (g) CTCFNet, and (h) the proposed DGDE-CSFNet model.

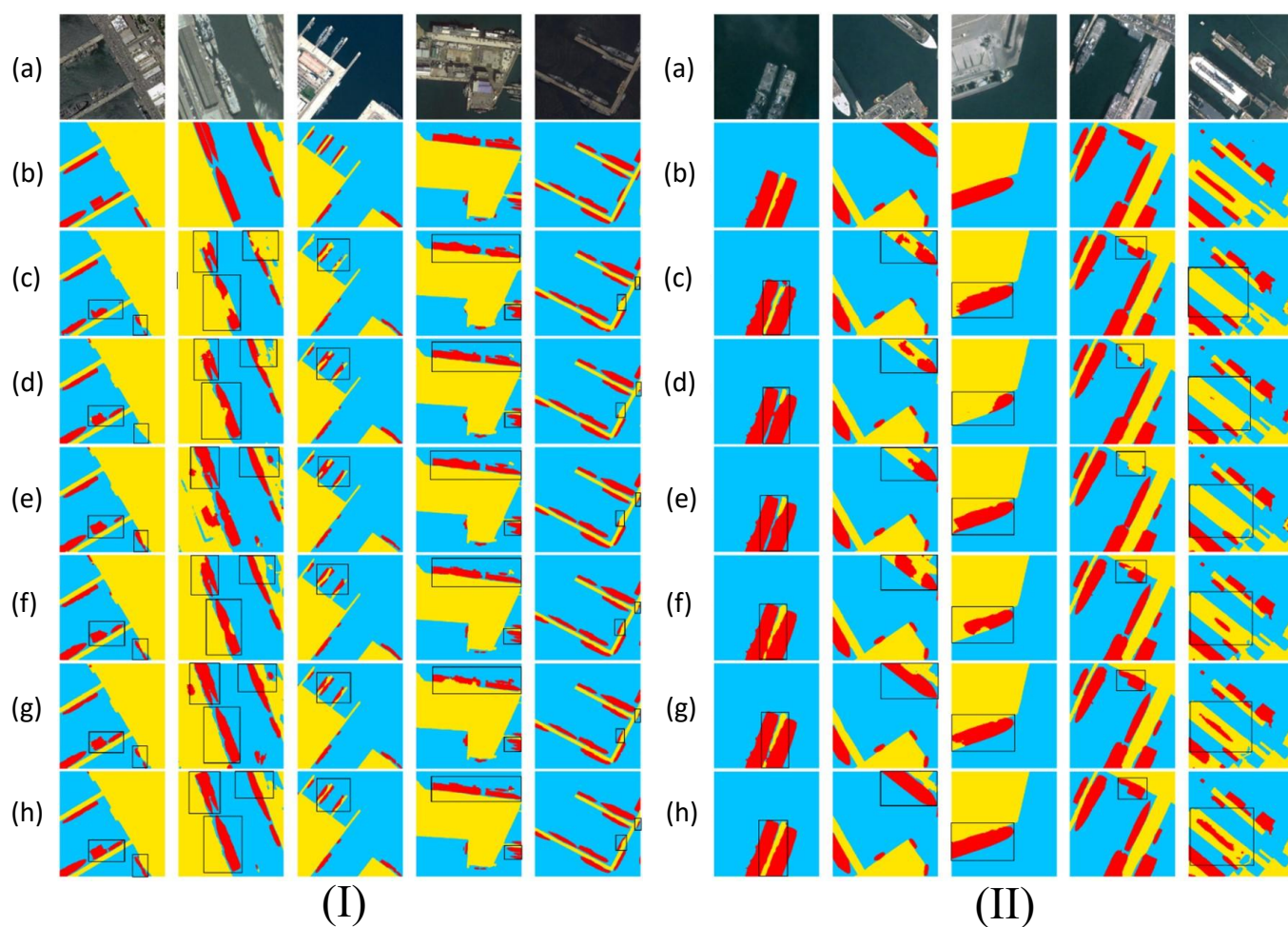


Figure 5. (I) Visualization results on the ISDSD dataset. (II) Visualization results on HRSC2016-SL dataset. Specifically, (a) denotes the original image, (b) the ground truth annotation, (c) the result of FCN, (d) DeepLabv3+, (e) CCNet, (f) SegFormer, (g) CTCFNet, and (h) the proposed DGDE-CSFNet model.

In Figure 5I, it is obvious that the proposed DGDE-CSFNet model significantly reduces background interference and mitigates issues caused by edge overlap and occlusion, thereby preventing misclassification between ships and other land-based structures such as docks. As shown in Figure 5II, some CNN-based methods tend to misclassify ship targets in complex environments, particularly in regions with blurred boundaries. Although models such as SegFormer and CTCFNet have shown improvements in overall segmentation performance, they still suffer from detail loss and fail to fully capture the complete structure of ship targets. In contrast, the proposed DGDE-CSFNet effectively mitigates misclassification issues, thereby achieving more accurate and complete segmentation performance. Overall, the visual evidence clearly highlights the robustness and effectiveness of the proposed DGDE-CSFNet model in handling challenging sea-land-ship segmentation tasks.

3.3. Ablation Experiment

To validate the effectiveness of the proposed modules in this work, ablation experiments are conducted on the ISDSD dataset. The ablation study primarily focuses on the following aspects:

3.3.1. Effectiveness of the Dual Branch Encoder and HDGA Module

To evaluate the effectiveness of the Dual Branch Encoder and the proposed HDGA module, five comparative variants are designed for the ablation study: (a) Using only the CNN encoder without any fusion mechanism; (b) Using only the Transformer encoder for feature extraction; (c) Employing a CNN and Transformer dual-branch encoder, with their features fused via a simple absolute difference operation; (d) Employing a CNN and Transformer dual-branch encoder, with their features concatenated along the channel dimension followed by a 1×1 convolution for conventional feature fusion; (e) Incorporating the proposed HDGA module to perform adaptive fusion of the output features of the dual-branch encoder. By comparing the performance under different configurations, the contribution of the HDGA module to feature representation and segmentation accuracy can be thoroughly evaluated.

As illustrated in Table 2, the proposed HDGA module achieves the best performance among the different encoder and feature fusion configurations. Compared with the CNN-only encoder, HDGA improves the F1-score, OA, and mIoU by 1.73%, 0.50%, and 2.98%, respectively. Compared with the Transformer-only encoder, the corresponding improvements are 0.49%, 0.06%, and 0.85%. In addition, simply fusing features through absolute element-wise subtraction results in lower performance, indicating that difference information alone cannot preserve sufficient local and global semantic representations. Furthermore, compared with conventional feature concatenation followed by a 1×1 convolution, the proposed HDGA module achieves improvements of 0.37%, 0.05%, and 0.59% in F1-score, OA, and mIoU, respectively. This demonstrates that explicitly modeling the hierarchical differences between CNN and Transformer representations is more effective than directly aggregating two branch features. By adaptively exploiting complementary information and discriminative differences between the two branches, HDGA enhances key region representation while suppressing background interference, leading to more accurate maritime-land-ship segmentation.

Table 2. Ablation experiment results of the HDGA module.

Methods	HDGA Module	F1 (%)	OA (%)	mIoU (%)
(a)	–	95.36	98.55	91.53
(b)	–	96.60	98.99	93.66
(c)	–	96.30	98.85	93.12
(d)	–	96.72	99.00	93.92
(e)	✓	97.09	99.05	94.51

✓ indicates that the HDGA module is used.

3.3.2. Effectiveness of the MCSF Module

To validate the effectiveness of the proposed MCSF module, three experimental setups were designed (a) concatenating the two feature maps along the channel dimension followed by a 1×1 convolution to adjust the channel number, replacing the MCSF module; (b) using only the UNet decoder module instead of the MCSF module; and (c) incorporating the proposed MCSF module. The results are shown in Table 3.

Table 3. Ablation experiment results of the Multi-Level Cross-Scale Fusion module.

Methods	MCSF Module	F1 (%)	OA (%)	mIoU (%)
(a)	–	96.69	98.98	93.8
(b)	–	96.84	99.04	94.07
(c)	✓	97.09	99.05	94.51

✓ indicates that the MCSF module is used.

When incorporating the Multi-Level Cross-Scale Fusion (MCSF) module, compared to experiment (a), the F1, OA, and mIoU metrics improved by 0.4%, 0.07%, and 0.71%, respectively. Compared to experiment (b), improvements of 0.25%, 0.01%, and 0.44% were observed on these three metrics. These results demonstrate that the MCSF module effectively integrates multi-scale features from the encoder at each layer, leveraging the global contextual information of deep features alongside the detailed information of shallow features. This reduces information loss during decoding and significantly enhances segmentation performance for ship targets.

3.3.3. Effectiveness of the Auxiliary Decoder Supervision

As shown in Table 4, four experimental setups were designed to validate the effectiveness of the loss supervision components in the proposed land-sea-ship segmentation model. Specifically, (a) represents the absence of both FCN and SegFormer auxiliary decoders; (b) employs only the FCN auxiliary decoder in the CNN encoder branch; (c) applies only the SegFormer auxiliary decoder in the Transformer branch; and (d) utilizes both FCN and SegFormer auxiliary decoders simultaneously.

Table 4. Ablation experiment results of the auxiliary decoder supervision.

Methods	FCN	SegFormer	F1 (%)	OA (%)	mIoU (%)
(a)	–	–	96.99	98.98	94.32
(b)	✓	–	97.02	99.01	94.39
(c)	–	✓	97.08	99.04	94.47
(d)	✓	✓	97.09	99.05	94.51

✓ indicates the inclusion of the corresponding auxiliary decoder in the network.

When only the FCN auxiliary decoder is incorporated, compared to the baseline experiment (a) without loss supervision, the model achieves improvements of 0.03%,

0.03%, and 0.07% in the F1, OA, and mIoU metrics, respectively. With only the SegFormer auxiliary decoder added, the corresponding improvements over experiment (a) are 0.09%, 0.06%, and 0.15%. When both auxiliary decoders are integrated simultaneously with their respective auxiliary losses, the model attains the highest accuracy, with F1, OA, and mIoU scores reaching 97.09%, 99.05%, and 94.51%, respectively. These results indicate that the FCN and SegFormer auxiliary decoders provide additional supervisory signals during training and contribute to model optimization. Although the resulting improvements are modest, multi-level auxiliary supervision yields small numerical gains in the mean F1, OA, and mIoU scores.

3.3.4. Computational Complexity Analysis

To further evaluate the computational efficiency of the proposed method, we compare the model complexity of DGCFNet with several representative segmentation models in terms of the number of parameters and FLOPs. All models are evaluated with an input resolution of 512×512 under a unified counting protocol. As shown in Table 5, DGCFNet contains 107.95 M parameters and requires 106.50 G FLOPs. Compared with SegFormer-B4, DGCFNet introduces additional computational cost due to its dual-branch feature extraction and cross-branch feature interaction. However, its computational cost remains substantially lower than that of CTFuse and CTCFNet.

Table 5. Computational complexity comparison of different segmentation models.

Model	Input Size	Params (M)	FLOPs (G)
SegFormer-B4	512×512	61.37	58.58
ST-UNet	512×512	168.81	95.12
CTFuse	512×512	147.06	309.31
CTCFNet	512×512	80.73	475.25
DGCFNet	512×512	107.95	106.5

4. Discussion

The superior performance of the proposed DGDE-CSFNet algorithm can be attributed to its ability to capture multi-scale semantic information and enhance feature discrimination, particularly for small-scale ship targets in cluttered port environments. Such performance benefits mainly result from the dual-branch encoding strategy for capturing both local details and global context, along with the differential attention fusion and cross-stage feature integration mechanisms that jointly enhance multi-scale feature representation and suppress background interference.

To further understand the effectiveness of DGDE-CSFNet, we analyze the performance differences among different semantic categories and datasets. The performance variations among different semantic categories are mainly related to their distinct visual characteristics and spatial distributions. Sea and land regions usually occupy large continuous areas with relatively homogeneous appearances, making them easier to distinguish using contextual information. In contrast, ship targets are more challenging due to their small spatial occupancy, diverse scales, complex surrounding backgrounds, and ambiguous boundaries. The consistent improvement in ship segmentation demonstrates the effectiveness of DGDE-CSFNet in enhancing discriminative feature representation through dual-branch feature learning and multi-scale feature fusion.

The performance differences between the ISDSD and HRSC2016-SL datasets can be attributed to their different scene characteristics. The ISDSD dataset contains complex port environments with diverse artificial structures and strong interactions between ships and surrounding facilities, resulting in more challenging segmentation conditions. In com-

parison, HRSC2016-SL provides maritime scenes with different target distributions and background compositions. The stable improvements achieved by DGDE-CSFNet on both datasets demonstrate its robustness in handling variations in scene complexity, object scale, and target-background relationships.

5. Conclusions

In this work, we address the challenges in sea-land-ship segmentation posed by complex port backgrounds, overlapping and occluded ship edges, and the difficulty of segmenting small ship targets. Specifically, a novel land-sea ship segmentation model based on hierarchical difference-guided attention and multi-level cross-scale fusion is proposed. The HDGA module integrates local features extracted by CNNs and global features obtained from Transformers, effectively leveraging their complementary characteristics to minimize feature redundancy and improve the overall understanding of complex scenes. The MCSF module effectively fuses multi-scale encoder features, accurately preserving fine details and amplifying feature responses in small target regions, thereby improving segmentation performance on small ships. Two auxiliary decoders are employed with corresponding auxiliary losses to provide multi-level supervisory signals during training, facilitating model optimization. Extensive experiments on two publicly available sea-land-ship segmentation datasets demonstrate the effectiveness of the proposed model through both quantitative comparisons and qualitative visualizations. Specifically, the proposed method improves mIoU by 0.68 and 0.46 percentage points over the strongest competing methods on the ISDSD and HRSC2016-SL datasets, respectively. For the ship category, which is particularly challenging because of its small spatial extent and complex surrounding backgrounds, the corresponding IoU improvements reach 1.66 and 1.04 percentage points.

Despite these improvements, several limitations remain. First, the dual-backbone architecture introduces additional computational overhead compared with single-backbone segmentation models. Second, the current evaluation is restricted to optical remote sensing imagery. Third, cross-dataset testing has not yet been conducted; therefore, the model's generalization capability across different acquisition conditions, geographic regions, and dataset distributions requires further investigation. Future work will focus on lightweight model design through pruning, quantization, and knowledge distillation, while also extending the evaluation to cross-dataset and multi-source remote sensing scenarios.

Author Contributions: Conceptualization, Y.C. and Y.Y.; methodology, C.X.; validation, S.Z., G.F., Y.C. and Y.Y.; formal analysis, G.F. and C.X.; investigation, resources and data curation, S.Z.; writing—original draft preparation, Y.C. and Y.Y.; writing—review and editing, C.X. and S.Z.; visualization, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Sanya Science and Technology Special Fund under Grant 2024KJFX034, Key Research and Development and Achievement Transformation Program of Inner Mongolia Autonomous Region under Grant 2025YFHH0078, the Project of Sanya Yazhou Bay Science and Technology City under Grant SKJC-JYRC-2024-63, the Hainan Provincial Natural Science Foundation of China under Grant 626QN0720, the PhD Scientific Research and Innovation Foundation of Sanya Yazhou Bay Science and Technology City under Grant No. HSPHDSRF-2024-03-009, the Hainan Province “Nanhai New Star” Technology Innovation Talent Platform Project under Grant NHXXRCXM202361, the General Program the Hubei Natural Science Foundation (No.2025AFB615), the National Natural Science Foundation of China under Grant 62571381, the Project of Sanya Science and Education Innovation Park of Wuhan University of Technology under Grant 2025ZCX005, and the MindSpore, which is a New Deep Learning Computing Framework (<https://www.mindspore.cn/>). The authors acknowledge Beijing PARATERA Technology Co., LTD for providing high-performance and AI computing resources for contributing to the research results reported in this paper. URL: <http://cloud.paratera.com>.

Data Availability Statement: The datasets used in this paper are publicly available in the following GitHub repository: <https://github.com/TYY92/2026-DGDE-CSFNet>, accessed on 2 October 2026.

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

References

1. Amani, M.; Ghorbanian, A.; Asgarimehr, M.; Yekkehkhany, B.; Moghimi, A.; Jin, S.; Naboureh, A.; Mohseni, F.; Mahdavi, S.; Layegh, N.F. Remote sensing systems for ocean: A review (Part 1: Passive systems). *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *15*, 210–234. [\[CrossRef\]](#)
2. Wang, H.; Li, X. DeepBlue: Advanced convolutional neural network applications for ocean remote sensing. *IEEE Geosci. Remote Sens. Mag.* **2023**, *12*, 138–161. [\[CrossRef\]](#)
3. Muhtar, D.; Zhang, X.; Xiao, P. Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [\[CrossRef\]](#)
4. Li, X.; Xu, F.; Liu, F.; Lyu, X.; Tong, Y.; Xu, Z.; Zhou, J. A synergistical attention model for semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [\[CrossRef\]](#)
5. Chen, Y.; Zhang, Q.; Wang, X.; Dong, Q.; Kang, M.; Jiang, W.; Wang, M.; Xu, L.; Zhang, C. Aanet: Adaptive attention networks for semantic segmentation of high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 14640–14655. [\[CrossRef\]](#)
6. Chen, Y.; Ye, Z.; Sun, H.; Gong, T.; Xiong, S.; Lu, X. Global–Local Fusion with Semantic Information Guidance for Accurate Small Object Detection in UAV Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [\[CrossRef\]](#)
7. Chen, Y.; Wang, Y.; Xiong, S.; Lu, X.; Zhu, X.X.; Mou, L. Integrating Detailed Features and Global Contexts for Semantic Segmentation in Ultrahigh-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [\[CrossRef\]](#)
8. Yue, J.; Wang, Y.; Pan, J.; Liang, H.; Wang, S.; Liu, Q. Less Is More: A Lightweight Deep Learning Network for Remote Sensing Imagery Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–13. [\[CrossRef\]](#)
9. Li, H.; Pan, H.; Liu, X.; Ren, J.; Du, Z.; Cao, J. GLVMamba: A Global-Local Visual State Space Model for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [\[CrossRef\]](#)
10. Zhao, H.; Luo, P.; Cui, W.; Xia, C.; Xu, X.; Feng, Z.; Chen, J.; Wang, J.; Xun, W.; Chen, C. Geographical Scenario Knowledge Informed Graph Structure Attention for Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 1–16. [\[CrossRef\]](#)
11. Gao, F.; Fu, M.; Cao, J.; Dong, J.; Du, Q. Adaptive Frequency Enhancement Network for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [\[CrossRef\]](#)
12. Wei, S.; Zeng, X.; Zhang, H.; Zhou, Z.; Shi, J.; Zhang, X. LFG-Net: Low-level feature guided network for precise ship instance segmentation in SAR images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [\[CrossRef\]](#)
13. Ouyang, L.; Fang, L.; Ji, X. Multigranularity self-attention network for fine-grained ship detection in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9722–9732. [\[CrossRef\]](#)
14. Xiong, W.; Xiong, Z.; Cui, Y. An explainable attention network for fine-grained ship classification using remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [\[CrossRef\]](#)
15. Zhang, T.; Shen, H.; Rehman, S.U.; Liu, Z.; Li, Y.; Rehman, O.U. Two-stage domain adaptation for infrared ship target segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [\[CrossRef\]](#)
16. Dong, Y.; Pan, X.; Yang, L.; Qu, S.; Cao, L. Semi-supervised evidential fusion for robust ship segmentation. *Int. J. Approx. Reason.* **2026**, *197*, 109755. [\[CrossRef\]](#)
17. Chen, J.; Wu, Y.; Gao, X.; Dai, W.; Zeng, X.; Diao, W.; Sun, X. Dpff-net: Dual-polarization image feature fusion network for sar ship detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [\[CrossRef\]](#)
18. Li, Z.; Dong, H.; Zhang, L.; Sun, X.; Kong, D. Mlanet: A robust ship segmentation network based on multi-level multi-attention feature fusion for complex maritime background environments. *IEEE Sens. J.* **2024**, *24*, 42404–42416. [\[CrossRef\]](#)
19. Zhao, C.; Fu, X.; Dong, J. CGA-Det: A CNN-GNN-Based Oriented SAR Ship Detector For Complex Scenes. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. [\[CrossRef\]](#)
20. Liu, D.; Hua, L.; Wang, Z.; Wang, L.; Chu, B.; Zeng, H.; Chen, Z.; Long, Z.; Zhang, Y.; Zhang, H. ShipMS-BSNet: A Multi-Scale Semantic Segmentation Method for Remote Sensing Ships in Complex Marine Environments. *Remote Sens.* **2026**, *18*, 1789. [\[CrossRef\]](#)
21. Shi, M.; Qiu, X.; Li, A.; Yang, Y.; Ke, Y.; Chen, Y. MPFT-UNet: A Boundary-Refined and Multi-Scale Dynamic Fusion Network for UAV-Based Port Ship Segmentation. *J. Mar. Sci. Eng.* **2026**, *14*, 945. [\[CrossRef\]](#)
22. Yang, C.; Yang, S.; Wang, P.; Wang, W.; Huang, Z. Constrained Boundary Enhancement for SAM 2-Based Ship Segmentation in UAV Berthing and Unberthing Videos. *Appl. Sci.* **2026**, *16*, 7730. [\[CrossRef\]](#)
23. Chen, X.; Li, D.; Liu, M.; Jia, J. CNN and transformer fusion for remote sensing image semantic segmentation. *Remote Sens.* **2023**, *15*, 4455. [\[CrossRef\]](#)

24. Wu, H.; Huang, P.; Zhang, M.; Tang, W.; Yu, X. CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. [\[CrossRef\]](#)
25. Xiang, X.; Gong, W.; Li, S.; Chen, J.; Ren, T. TCNet: Multiscale Fusion of Transformer and CNN for Semantic Segmentation of Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 3123–3136. [\[CrossRef\]](#)
26. Zhu, X.; Peng, T.; Guo, J.; Wang, H.; Cao, T. CNN-transformer dual branch collaborative model for semantic segmentation of high-resolution remote sensing images. *Photogramm. Rec.* **2025**, *40*, e12524. [\[CrossRef\]](#)
27. Yang, N.; Tian, C.; Gu, X.; Zhang, Y.; Li, X.; Zhang, F. RST-Net: A Semantic Segmentation Network for Remote Sensing Images Based on a Dual-Branch Encoder Structure. *Sensors* **2025**, *25*, 5531. [\[CrossRef\]](#)
28. Ji, Y.; Shi, W.; Lei, J.; Ding, J. DBRSNet: A dual-branch remote sensing image segmentation model based on feature interaction and multi-scale feature fusion. *Sci. Rep.* **2025**, *15*, 27786. [\[CrossRef\]](#)
29. Zhang, B.; Chen, Y.; Dang, W.; Xiong, S.; Lu, X. A Spatial and Semantic Alignment Fusion Network for Sea-Land Port Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 7420–7435. [\[CrossRef\]](#)
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [\[CrossRef\]](#)
33. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [\[CrossRef\]](#)
34. Tian, T.; Pan, Z.; Tan, X.; Chu, Z. Arbitrary-oriented inshore ship detection based on multi-scale feature fusion and contextual pooling on rotation region proposals. *Remote Sens.* **2020**, *12*, 339. [\[CrossRef\]](#)
35. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017; Volume 2, pp. 324–331. [\[CrossRef\]](#)
36. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [\[CrossRef\]](#)
37. 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), Munich, Germany, 8–14 September 2018; pp. 801–818. [\[CrossRef\]](#)
38. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612. [\[CrossRef\]](#)
39. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [\[CrossRef\]](#)
40. Lu, C.; Zhang, X.; Du, K.; Xu, H.; Liu, G. CTCFNet: CNN-Transformer Complementary and Fusion Network for High-Resolution Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [\[CrossRef\]](#)
41. Xie, C.; Chen, Y.; Zhu, J.; Xiong, S. Attention Module and Edge Enhancement for Sea-Land Segmentation. In Proceedings of the 2024 8th Asian Conference on Artificial Intelligence Technology (ACAIT), Fuzhou, China, 8–10 November 2024; pp. 456–463. [\[CrossRef\]](#)

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.