Cross-Domain Few-Shot Infrared Ship Segmentation With Class-Specific Adapters and SAM Refinement

Ting Zhang, Pengyi Liu, Zhaoying Liu, *Member, IEEE*, Yingchun Chen, Hisham Alfuhaid, *Member, IEEE*, Mohammad Al-Antary, Muhammad Waqas, *Senior Member, IEEE*, and Syed Mudassir Hussain, *Member, IEEE*

***Abstract*—Infrared (IR) ship semantic segmentation is pivotal for all-weather maritime surveillance and national early-warning applications. However, the significant domain gap between IR and visible images, combined with the extreme scarcity of annotated samples in the target IR domain, poses a dual challenge that severely restricts the cross-domain application of segmentation networks. Existing few-shot segmentation methods mostly assume similar data distributions, while domain adaptation methods rely on abundant unlabeled data; consequently, neither approach effectively addresses the combined problem of cross-domain and few-shot learning. To address this issue, a two-stage framework named class-specific adapters and segment anything model (SAM) refinement Network is proposed for cross-domain few-shot IR ship semantic segmentation. The framework first learns domain-invariant features and then leverages a handful of annotated IR samples for class-specific adaptation and mask refinement. Specifically, in the first stage, a dual-branch network integrating wavelet transform and convolution is designed to extract robust cross-domain features, employing a random convolution perturbation strategy to stabilize adversarial training. In the second stage, independent, lightweight class-specific adapters are introduced for each target class and efficiently fine-tuned via self-supervised contrastive learning. Furthermore, the SAM is incorporated during inference through a prototype-driven prompt generation mechanism to refine initial segmentation results and enhance boundary accuracy. Experimental results on the VI-Ship and Agriculture-Vision datasets demonstrate that the proposed method outperforms state-of-the-art approaches across multiple evaluation metrics. Notably, under the extreme one-shot setting on the VI-Ship dataset, the proposed method achieves a mean Intersection over Union of 63.19%, exceeding the second-best method by 4.43%, thereby validating its effectiveness.**

***Index Terms*—Class-specific adapters (CAS), cross-domain few-shot learning, infrared (IR) ship imagery, prototype learning, segment anything model (SAM).**

I. Introduction

PRECISE ship segmentation plays a pivotal role in maritime surveillance and national defense. With the advancement of intelligent vision technology, infrared (IR) imaging offers strong all-weather perception capabilities, especially in low-visibility environments, such as night, rain, and fog, and has become a key modality for ship monitoring and target recognition [1], [2], [3]. However, applying deep learning models directly to IR images faces two inherent limitations: first, the intrinsic imaging characteristics of IR images, which feature low contrast, high noise, and a lack of texture details; second, the extreme difficulty and high cost of acquiring pixel-level annotations, resulting in a severe scarcity of training data [4], [5], [6]. In contrast, visible (VIS) images possess natural advantages in clarity and accessibility, serving as a vital data source for training high-performance deep segmentation models. Consequently, transferring rich knowledge learned from the VIS domain to the IR domain is a viable strategy for enhancing IR image segmentation performance [7], [8], [9]. However, such cross-modal knowledge transfer is nontrivial, as models must confront a dual dilemma constituted by a massive domain discrepancy and extreme sample scarcity. On the one hand, the distinct physical imaging principles of VIS and IR images result in fundamental differences in color, texture, and light response. On the other hand, unlike existing unsupervised domain adaptation (UDA) methods that rely on abundant unlabeled target data, IR image acquisition in many real-world scenarios is inherently difficult, preventing models from bridging the domain gap by aligning large-scale data distributions. When a model is simultaneously trapped between a significant domain gap and sparse target domain supervisory signals, standard transfer strategies are prone to failure. First, limited supervision is insufficient to support robust adaptation to the new domain, leading to overfitting on the few-shot target domain and insufficient generalization [10], [11], [12]. More

Received 16 February 2026; revised 15 May 2026; accepted 13 June 2026. Date of publication 16 June 2026; date of current version 10 July 2026. This work was supported by the National Natural Science Foundation of China under Grant 61906005, in part by the Beijing Municipal Education Commission through the General Project of Science and Technology Plan under Grant KM202110005028, in part by the Beijing University of Technology through the Project of Interdisciplinary Research Institute under Grant 2021020101, in part by the International Research Cooperation Seed Fund of Beijing University of Technology under Grant 2021A01, and in part by the King Khalid University, KSA, through the Deanship of Scientific Research and Graduate Studies, Small Research Group, under Grant RGP.1/39/46. *(Corresponding authors: Zhaoying Liu; Syed Mudassir Hussain.)*

Ting Zhang, Pengyi Liu, and Zhaoying Liu are with the School of Computer Science, Beijing University of Technology, Beijing 100124, China (e-mail: zhangting@bjut.edu.cn; liupengyi@emails.bjut.edu.cn; zhaoying.liu@bjut.edu.cn).

Yingchun Chen is with the State Key Laboratory of Bridge Engineering Safety and Resilience, Beijing University of Technology, Beijing 100124, China (e-mail: ychen08089@163.com).

Hisham Alfuhaid is with the Department of Computer Science, King Khalid University, Abha 61421, Saudi Arabia (e-mail: alasmary@kku.edu.sa).

Mohammad Al-Antary and Muhammad Waqas are with the School of Computing and Mathematical Science, Faculty of Engineering and Science, University of Greenwich, SE10 9LS London, U.K. (e-mail: m.t.m.alantary@greenwich.ac.uk; engr.waqas2079@gmail.com).

Syed Mudassir Hussain is with the School of Physics, Engineering and Computer Science, University of Hertfordshire, AL10 9AB Hatfield, U.K. (e-mail: s.m.hussain2@herts.ac.uk).

Digital Object Identifier 10.1109/JSTARS.2026.3704472

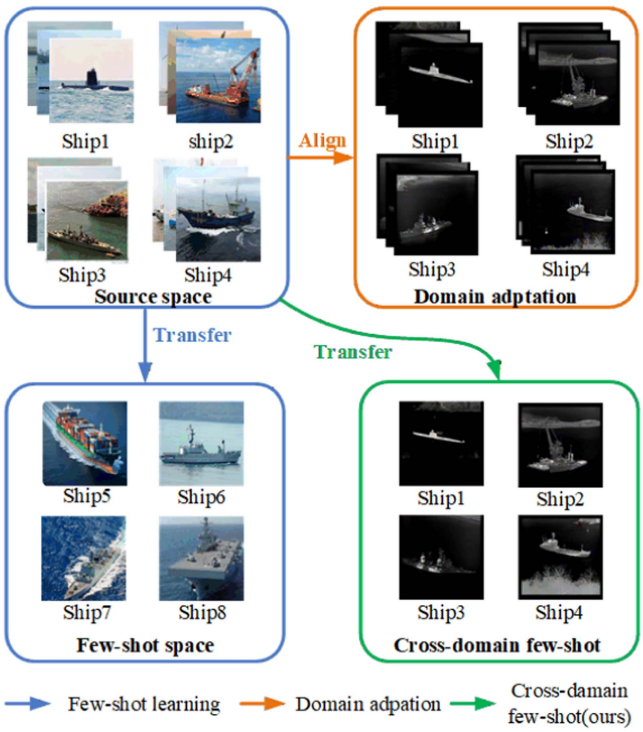

Fig. 1. Schematic comparison of few-shot learning, domain adaptation, and the cross-domain few-shot learning (CDFSS) studied in this article. Few-shot learning aims to address the generalization problem for new categories within the same data domain. Domain adaptation focuses on overcoming distribution discrepancies. CDFSS integrates these challenges.

critically, the massive domain discrepancy can induce negative transfer.

In recent years, significant progress has been made in domain adaptation, few-shot learning, and cross-domain few-shot semantic segmentation (CDFSS), as illustrated in Fig. 1. UDA methods, particularly those based on adversarial training (e.g., ADVENT) or self-training (e.g., CPSL) strategies, perform well in handling certain domain differences but generally rely on large amounts of unlabeled target domain data [16]. To address massive domain gaps, such as visible-to-infrared (VIS-IR), researchers have proposed more refined strategies. For instance, Zhang et al. [19] proposed T-DANet, which systematically addresses the problem through style transfer and a multistage training system. Gan et al. [20] proposed MDA-Net, designing a multidomain attention network to intelligently transfer cross-domain invariant features and avoid negative transfer. However, these advanced UDA methods remain limited by their dependence on large-scale target domain data and cannot handle scenarios with data scarcity. Conversely, few-shot semantic segmentation (FSS) methods (e.g., PANet, HSNet) [21], [22] excel at learning new categories from a few samples, but their fundamental premise is that the data follows an intradomain identical distribution. This assumption is completely broken in cross-domain VIS-IR scenarios, causing sharp performance declines. To address this challenge, CDFSS has emerged. Existing works, such as PATNet and APSeg, enhanced generalization by introducing domain-specific transformations or leveraging large foundation models [24], [25], [26]. Nevertheless, when facing extreme differences, such as VIS-IR, these methods still suffer from insufficiently robust feature alignment, rough modeling of class-specific attributes, and unrefined segmentation boundaries, limiting their application in practical scenarios.

To solve the aforementioned problems, we propose a two-stage method for IR ship segmentation, termed class-specific adapters and segment anything model (SAM) refinement Network (CAS-Net). The framework primarily consists of two phases: cross-domain feature extraction and class-adapter fine-tuning. Specifically, in the first stage, we design a dual-branch network fusing wavelet transform and convolution, introducing a random convolution perturbation (RCP) strategy to stabilize adversarial training, thereby constructing a shared feature space highly robust to significant domain gaps. In the second stage, targeting new classes in the target domain, we design independent lightweight class-specific adapters (CAS) assisted by self-supervised contrastive learning for efficient fine-tuning. This achieves precise class semantic modelling from extremely few annotated samples. Furthermore, the SAM is introduced in the inference stage. We propose a prototype-driven prompt generation mechanism to refine the initial segmentation mask, significantly improving the boundary precision and structural integrity of the final segmentation results. The main contributions of our work are as follows.

1) We propose a two-stage cross-domain few-shot segmentation framework that decouples the task into domain-shared feature learning and few-shot class adaptation, effectively addressing the dual challenges of VIS-IR domain discrepancy.
2) We construct a robust mechanism for cross-domain shared feature learning. By fusing wavelet transform with RCP, we enhance the stability of the model in extracting shared features under significant domain gaps.
3) We design an efficient few-shot class adaptation strategy. Through independent lightweight adapters and self-supervised contrastive learning, we achieve precise and parameter-efficient semantic modeling for new categories.
4) We introduce a prototype-driven SAM mask refinement module, which translates learned class semantic information into geometric prompts to effectively guide SAM in optimizing segmentation boundaries.
5) Experimental results on the VI-Ship and Agriculture-Vision datasets indicate that our method significantly outperforms existing state-of-the-art methods in all few-shot settings.

II. Related Work

A. Cross-Domain Semantic Segmentation

Cross-domain semantic segmentation aims to transfer knowledge from a label-rich source domain to a target domain, with the core challenge lying in bridging the distribution discrepancy between different domains [13], [14], [15]. Existing works primarily focus on two paradigms. The first is global distribution alignment based on adversarial learning. Representative works, such as ADVENT [16], employ adversarial training on the entropy maps of the segmentation network predictions.

This implicitly encourages the model to produce low-entropy, high-confidence predictions on the target domain, thereby indirectly achieving feature distribution alignment. Hoffman et al. [27] proposed cycle-consistent adversarial domain adaptation, which conducts adversarial training in both input and feature spaces. By utilizing a cycle-consistency loss, this method ensures that source and target images retain their original structural content after style translation, effectively aligning both image styles and feature distributions. Tsai et al. [28] introduced the adapt semantic segmentation model, which enhances the adaptation of low-level features by performing domain adaptation at multiple feature layers via a multilevel adversarial network.

The second category involves self-training based on iterative pseudolabel optimization, such as CPSL [17]. This approach utilizes pseudolabels generated by the model in high-confidence regions to supervise target domain learning, employing strategies like class balancing and online clustering to improve pseudolabel quality [18]. Zhang et al. [19] proposed prototypical denoising domain adaptation, which first filters reliable pseudolabels based on the distance between target image features and class prototypes, and then utilizes these labels to retrain the segmentation network. From an image style perspective, Yang and Soatto [29] proposed Fourier domain adaptation. This method replaces the low-frequency components of the source domain with those of the target domain via Fourier transform, narrowing the style difference between domains and indirectly improving pseudolabel quality [29].

However, when facing scenarios like VIS-IR, where distinct physical imaging principles lead to massive domain discrepancies, the effectiveness of the aforementioned general strategies declines significantly. To address this, researchers have proposed more refined solutions. For instance, Zhang et al. [19] proposed a two-stage framework (T-DANet) that systematically fuses style transfer in the input space with attention modules in the feature space. Gan et al. [20] proposed MDA-Net, which focuses on avoiding negative transfer by designing a multidomain attention network to adaptively transfer truly cross-domain invariant features rather than forcibly aligning all features. In addition, some orthogonal studies focus on reinforcing feature representation itself. For example, gated spatial CNN (Gated-SCNN) introduced a gated convolution mechanism to strengthen the extraction of object boundary information [30]. ACNet utilizes asymmetric convolution blocks to enhance core feature expression [31], [32]. DANet explicitly models feature dependencies in spatial and channel dimensions via dual attention mechanisms [33]. APM decomposes features into amplitude and phase spectra at the feature level, introducing learnable masks to filter frequency components before reverting to the spatial domain for enhanced features [34].

Despite the significant progress achieved by these methods in UDA tasks, their core designs imply two key assumptions incompatible with the scenario addressed in this article. First, dependence on target data volume: the vast majority of UDA methods assume the existence of abundant unlabeled data in the target domain to allow the model to fully learn its data distribution. This contradicts real-world applications where IR image acquisition is costly and data are scarce. Second, vulnerability under massive domain gaps: in cases where target domain data are severely insufficient, these alignment methods are not only limited in effectiveness but are also more prone to forgetting source domain knowledge or overfitting target domain noise due to sparse supervisory signals, resulting in a sharp decline in performance.

B. Few-Shot Semantic Segmentation

Metric learning-based methods construct a prototype for each new category in an embedding space and perform classification by calculating the distance or similarity between query pixels and class prototypes [35]. For instance, Wang et al. [21] proposed PANet, which introduces prototype alignment regularization to enhance the consistency between support and query prototypes. Tian et al. [36] proposed PFENet, combining feature pyramids with attention mechanisms to endow prototypes with multiscale expressive capabilities. The authors in [22] and [23] presented HSNet, which elevates metric learning from the pixel level to a structured region matching level by constructing multiscale 4-D correlation tensors and introducing 4-D convolution.

Augmentation-based methods aim to mitigate dependence on support set prototypes and handle appearance variations between support and query sets [37], [38], [39]. For example, Fan et al. [40] proposed the self-support prototype (SSP) network, which mines high-confidence regions from the query image itself to generate adaptive prototypes, thereby alleviating reliance on the appearance of the support set. The authors in [41] and [42] proposed PGNet, utilizing graph neural networks to explicitly model semantic relations, guiding the generation of semantically more complete prototypes.

However, most FSS methods operate on a fundamental assumption: the base classes in the meta-training stage and the new classes in the few-shot testing stage share a similar data distribution [43], [44]. This assumption is fundamentally violated in the VIS-IR cross-domain scenario studied in this article. Due to the massive domain discrepancy, directly applying the metric space learned in the VIS domain to the IR domain is ineffective, as prototypes derived from VIS images are semantically incomparable to IR features. Consequently, while we draw upon the concept of prototype matching, we fundamentally reconstruct the process: our class adaptation is not performed in the raw feature space, but on the shared features output by the first stage—features that have already undergone deep domain adaptation. Furthermore, through independent CAS and self-supervised contrastive learning, we force the model to learn the unique attributes of the categories within the target domain itself.

C. Cross-Domain Few-Shot Semantic Segmentation

CDFSS integrates the challenges of both paradigms, aiming to complete segmentation tasks in the target domain using abundant labeled data from the source domain and extremely few labeled samples from the target domain, which aligns with the task in this study [45], [46]. Existing works can be broadly summarized

into two categories: meta-learning-based and fine-tuning-based approaches [47], [48].

Meta-learning-based methods aim to learn a fast-adapting meta-capability. Representative works, such as PMNet, achieve generalization of meta-knowledge by learning domain-specific feature transformations [49], while RestNet explores leveraging priors from large models to enhance cross-domain generalization [50], [51]. However, the limitation of these methods is that when the discrepancy between the source and target domains is excessive, the effectiveness of the learned meta-knowledge declines sharply, severely constraining generalization capabilities.

In contrast, fine-tuning-based methods, have garnered significant attention due to their flexibility and practicality. These methods typically pretrain a strong backbone network on the source domain, then freeze most parameters and fine-tune only specific modules using a few samples from the target domain [52], [53], [54]. For instance, ABCDFSS achieves parameter-efficient adaptation by fine-tuning only lightweight adapter modules [55]. TGCM explicitly models target-guided knowledge transfer for robust cross-modal segmentation [70]. IFA provides stronger supervision signals for the fine-tuning process by constructing bidirectional predictions [47]. Moreover, recent works have further advanced this field. Informative Structure Adaptation (ISA) proposes a source-domain-free, test-time adaptation strategy that automatically identifies and progressively updates the most crucial model structures based on support samples without retraining on the source domain [71]. Another novel method, LLF, highlights that low-level features are highly sensitive to domain shifts. It addresses this by introducing a Low-level Enhancing Module via RCPs and a low-level calibration module to correct cross-domain distortions [72], [73]. Li et al. [74] introduced a dual-agent optimization framework that simultaneously rectifies feature domain sensitivity via frequency-domain mutual aggregation and corrects cross-domain matching errors in the spatial domain. However, significant differences exist in design philosophy and applicable scenarios among different adapters. For example, AdapterHub and AdaptFormer are designed for Vision Transformer architectures, focusing on interaction with self-attention modules [56], [57]. Conversely, Conv-Adapter, though designed for convolutional networks, has mostly been validated in image classification tasks. When these adapters are directly applied to scenarios where the architecture or task objective does not match (e.g., dense prediction in semantic segmentation), their performance may be limited [58], [59].

Although CDFSS methods have made progress, they universally face a cascading bottleneck when addressing extreme scenarios like VIS-to-IR ship segmentation: 1) Insufficient feature foundation: Many methods insufficiently model the massive domain discrepancy during the feature alignment stage. This results in low-quality shared features provided by the pretrained model, burying latent risks for subsequent few-shot learning; 2) Rough class adaptation: In few-shot settings, models struggle to finely model the unique attributes of new target domain categories from sparse supervisory signals, leading to class confusion and loss of detail; 3) Missing boundary refinement: The vast majority of methods focus optimization on core mask prediction while neglecting boundary quality, resulting in final outputs with evident defects in structural integrity and boundary accuracy.

III. Proposed Method

A. Overall Framework

To address the significant domain discrepancy and the scarcity of target domain samples in VIS-IR ship image segmentation, we propose CAS-Net, as illustrated in Fig. 2. It decouples the segmentation task into two sequential stages: cross-domain feature extraction and CAS fine-tuning.

The first stage aims to learn shared feature representations that are robust to both VIS and IR images under conditions of significant domain discrepancy. To this end, we design a dual-branch network fusing frequency domain and convolutional information to extract cross-modal invariant features. Meanwhile, to stabilize adversarial training under data-scarce target domain conditions, we introduce a RCP strategy. This strategy increases the difficulty of discriminator training, thereby compelling the extractor to learn more generalizable shared knowledge.

The second stage aims to achieve efficient adaptation for new categories using extremely few annotated samples from the target domain. To achieve this, we introduce independent, lightweight CAS for each target category and fine-tune them in conjunction with a self-supervised contrastive learning strategy.

Finally, in the inference stage, to address the issue of coarse segmentation boundaries prevalent in existing methods, we introduce the SAM. We propose a prototype-driven prompt generation mechanism that utilizes learned class semantics to guide SAM in the fine-grained refinement of initial prediction masks, ultimately outputting high-quality segmentation masks.

B. Stage I: Cross-Domain Feature Extraction

1) Feature Extraction Network: To simultaneously capture spatial semantic information and structural texture information critical for cross-domain tasks, we construct a parallel dual-branch network architecture comprising a spatial branch and a frequency branch. We term the feature enhancement structure composed of the wavelet branch and fusion convolution as the wavelet-based feature enhancement (WFE) module. For a given input image $\boldsymbol{X}$, the two branches extract features in parallel

$$\boldsymbol{F}_{sp} = E_{sp}(\boldsymbol{X}). \tag{1}$$

The frequency branch first processes the input image $\boldsymbol{X}$ via discrete wavelet transform (DWT). Unlike the Fourier Transform, which discards spatial localization and only provides global frequency statistics, the wavelet transform yields a localized time–frequency representation. This property is particularly crucial for IR target segmentation, where precise discrimination relies heavily on localized contour boundaries and shapes rather than fine-grained surface textures. By retaining vital spatial structural details alongside high-frequency edge responses, DWT effectively enhances cross-domain feature robustness without losing critical geometric semantics. Low-pass

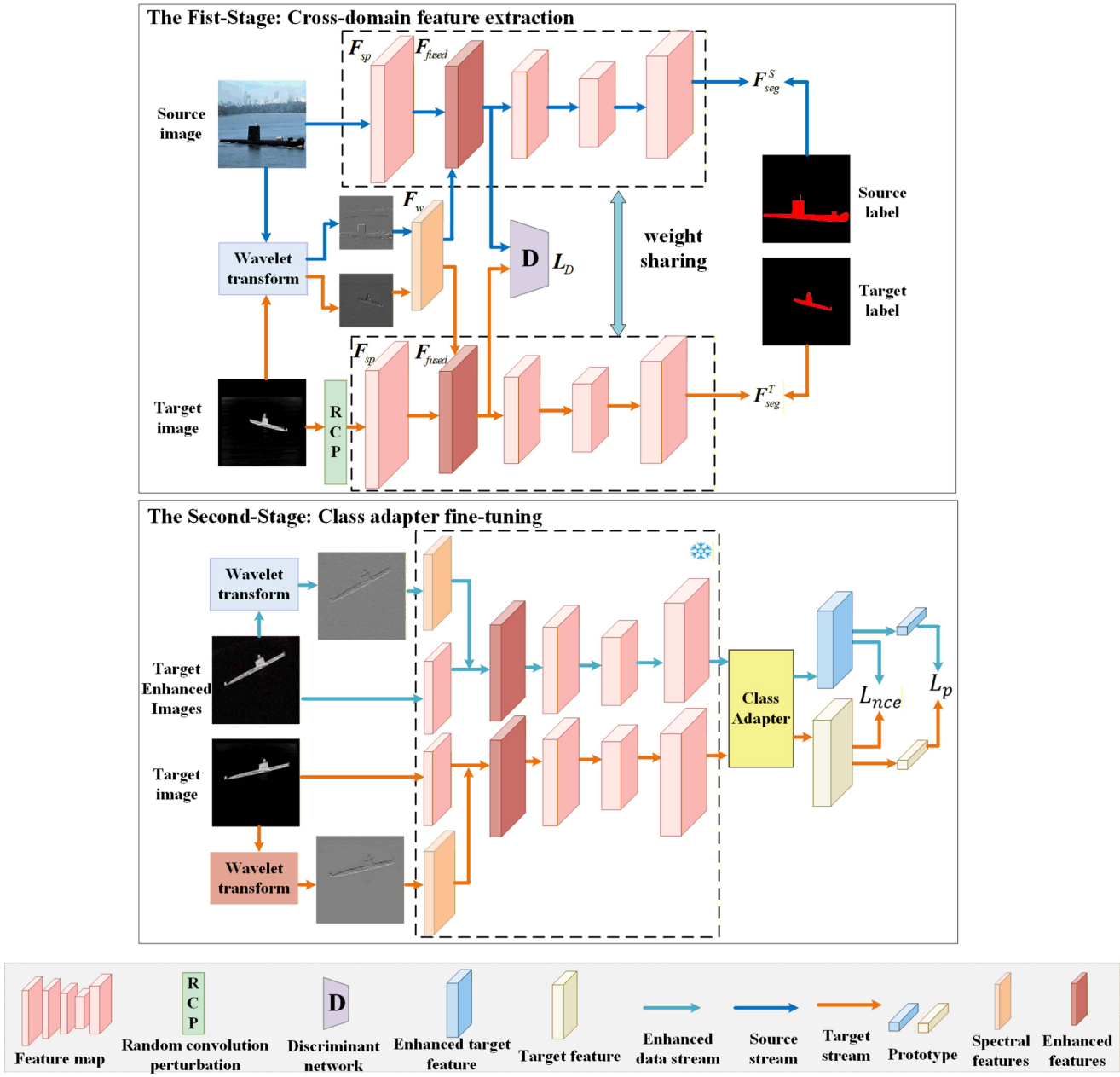

Fig. 2. Overview of the proposed two-stage CDFSS framework.

and high-pass filters are applied in the horizontal and vertical directions, followed by $2\times$ downsampling, to decompose the image $\boldsymbol{X}$ into one low-frequency subband $\boldsymbol{X}_{LL}$ and three high-frequency subbands (horizontal $\boldsymbol{X}_{HL}$, vertical $\boldsymbol{X}_{LH}$, and diagonal $\boldsymbol{X}_{HH}$) [60], specifically expressed as:

$$[\boldsymbol{X}_{LL}, \boldsymbol{X}_{LH}, \boldsymbol{X}_{HL}, \boldsymbol{X}_{HH}] = \mathrm{DWT}(\boldsymbol{X}). \tag{2}$$

Considering that the three high-frequency subbands contain more stable edge and texture details, we concatenate these subbands and feed them into an independent frequency encoder E_w to generate deep frequency features $\boldsymbol{F}_w$

$$\boldsymbol{F}_w = E_w(\mathrm{Concat}(\boldsymbol{X}_{LH}, \boldsymbol{X}_{HL}, \boldsymbol{X}_{HH})). \tag{3}$$

Finally, to achieve complementarity between the two modalities, the spatial features $\boldsymbol{F}_{sp}$ and frequency features $\boldsymbol{F}_w$ are concatenated and fused through a fusion convolution layer to form a unified cross-domain feature representation $\boldsymbol{F}_{\text{fused}}$

$$\boldsymbol{F}_{\text{fused}} = \mathrm{Conv}_{\text{fuse}}(\mathrm{Concat}(\boldsymbol{F}_{sp}, \boldsymbol{F}_w)) \tag{4}$$

where $\mathrm{Concat}(\cdot)$ denotes feature concatenation along the channel dimension; $\boldsymbol{F}_w$ and $\boldsymbol{F}_{sp}$ are the feature maps output by the spatial and frequency branches, respectively; $\mathrm{Conv}_{\text{fuse}}(\cdot)$ represents the fusion convolution layer that outputs the unified cross-domain feature representation $\boldsymbol{F}_{\text{fused}}$.

2) RCP for Robust Adversarial Training: To guide the network in learning domain-invariant features, we introduce domain adversarial training. Fundamentally, adversarial training

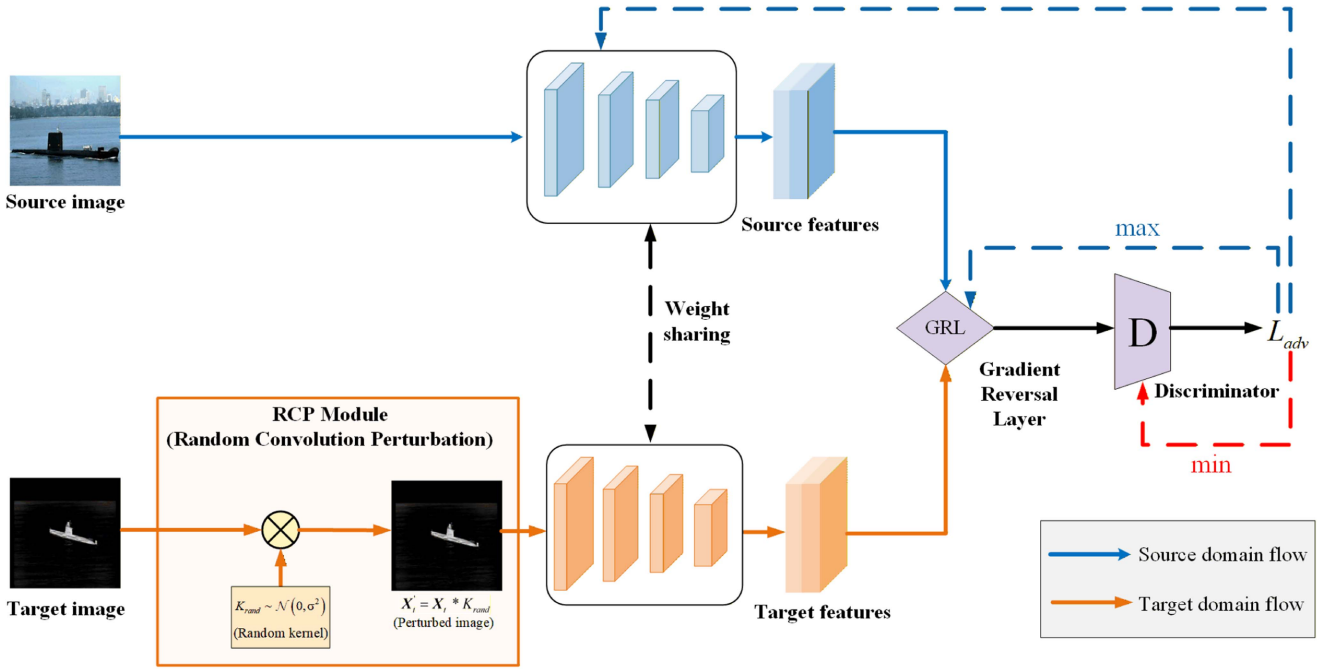

Fig. 3. Illustration of the convolutional perturbation adversarial training strategy.

constitutes a minimax game between the feature extraction network and the discriminator network. Specifically, the discriminator aims to distinguish the domain origin of the input features, whereas the feature extractor endeavors to deceive the discriminator by generating domain-invariant features. This adversarial process can be formulated as follows:

$$\min_E \max_D V(D,E) = \mathbb{E}_{\boldsymbol{X}_s \sim \mathbb{P}_s}\left[\log D(E(\boldsymbol{X}_s))\right] + \mathbb{E}_{\boldsymbol{X}_t \sim \mathbb{P}_t}\left[\log\left(1 - D(E(\boldsymbol{X}_t))\right)\right] \quad (5)$$

where $E(\cdot)$ denotes the feature extractor; $D(\cdot)$ is the domain discriminator, which is employed to predict whether the input features originate from the source domain or the target domain; $\boldsymbol{X}_s$ and $\boldsymbol{X}_t$ represent the source and target domain images, respectively; $\mathbb{P}_s$ and $\mathbb{P}_t$ denote their corresponding data distributions; $V(D,E)$ is the adversarial objective function.

When target domain samples are extremely scarce, the discriminator is susceptible to overfitting on the limited data, thereby backpropagating erroneous or unstable gradient signals to the feature extraction network [61]. To address this, we introduce the RCP strategy, as illustrated in Fig. 3. The core idea of this method is to apply a subtle, RCP to the target domain image before it enters the feature extraction network, thereby generating hard samples in an online manner to increase the difficulty of the discrimination task. Specifically, a random convolution kernel K_{rand}, whose parameters follow a normal distribution, is convolved with the target domain image $\boldsymbol{X}_t$ to generate a perturbed image, denoted as

$$\boldsymbol{X}_{\text{perturbed}} = \boldsymbol{X}_t * K_{\text{rand}}, \quad K_{\text{rand}} \sim \mathcal{N}\left(0,\sigma^2\right) \quad (6)$$

where K_{rand} denotes the randomly generated convolution kernel, whose parameters are sampled from a normal distribution with a mean of 0 and a variance of σ^2. This perturbation augments the diversity of target domain inputs while preserving the core semantic content. By increasing the difficulty of the discrimination task, it effectively mitigates the risk of discriminator overfitting and encourages the feature extraction network to learn shared features that are less sensitive to domain discrepancies.

C. Stage II: Class Adapter Fine-Tuning

In this stage, as illustrated in Fig. 4, we freeze the pre-trained shared feature extraction network and utilize extremely few annotated samples from the target domain to achieve efficient adaptation and refined segmentation. To this end, we design independent CAS for semantic modeling and employ SAM for boundary refinement.

1) Class-Specific Adapter: To achieve rapid adaptation to new categories in the target domain while maximally preserving the pretrained universal cross-domain knowledge and mitigating the risk of overfitting, we design an independent, lightweight adapter A_c for each target category c. In contrast to fine-tuning the entire network or employing a single adapter shared across all categories, this class-specific design allows each adapter to specialize in learning the unique semantic features of its corresponding category. This approach effectively avoids interclass interference, thereby enabling more precise semantic modeling even with extremely scarce samples.

The adapter adopts a lightweight bottleneck architecture, consisting of a downprojection convolution, a nonlinear activation, and an upprojection convolution, supplemented by a residual connection. It is embedded within the high-level feature paths of the frozen backbone network [61]. The operation is

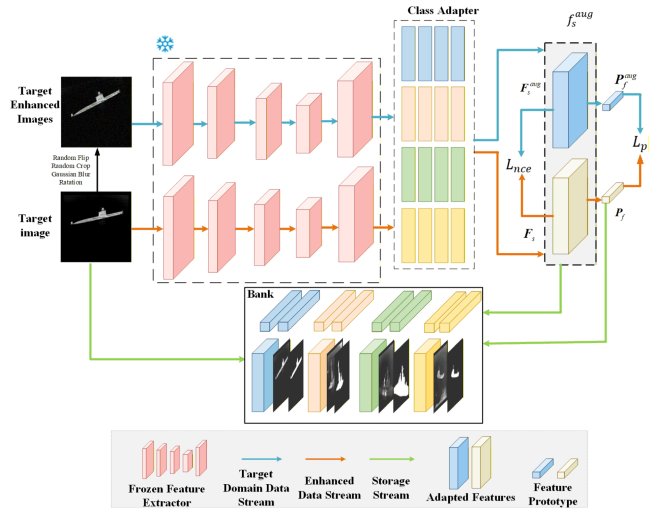

Fig. 4. Illustration of Stage II: Adapter Fine-tuning. In this stage, the shared feature extraction network trained in the first stage is frozen, and parameter fine-tuning is performed exclusively on the CAS introduced for the new categories in the target domain. For target domain images, two data streams—an original view and an augmented view—are constructed via random rotation, cropping, and noise addition. These streams pass through the frozen backbone, where lightweight bottleneck adapters (consisting of down-projection convolution, ReLU, and up-projection convolution with a residual connection) specific to the corresponding category are inserted into the high-level feature paths to learn class-unique semantics. Subsequently, at the pixel level, dense contrastive alignment is performed using spatially corresponding features from the original and augmented views to formulate self-supervised signals. At the class level, foreground and background prototypes are aggregated from masked regions to enforce prototype-level alignment, thereby enhancing global class discriminability. Simultaneously, the adapted feature representations and updated class prototypes are stored in a Memory Bank, serving as the basis for generating initial masks via dense matching and for generating positive/negative SAM prompt points based on prototype similarity.

formulated as

$$A_c(\boldsymbol{F}) = \boldsymbol{F} + \boldsymbol{W}_{c,2} * \sigma\left(\boldsymbol{W}_{c,1} * \boldsymbol{F}\right) \tag{7}$$

where $\boldsymbol{F}$ denotes the features extracted by the backbone network, $\boldsymbol{W}_{c,1}$ and $\boldsymbol{W}_{c,2}$ represent the trainable weights within the adapter, $\sigma(\cdot)$ is the ReLU activation function, and A_c represents the adapter of category c.

2) Adapter Fine-Tuning Via Self-Supervised Contrastive Learning: Under conditions where only K-shot annotated samples are available in the target domain, directly fine-tuning the entire network is prone to overfitting and training instability. To address this, in the second stage, we freeze the shared feature extractor trained in the first stage and perform parameter-efficient fine-tuning exclusively on the CAS corresponding to the new categories in the target domain. Furthermore, we exploit the structural and semantic information within the limited annotated samples by constructing self-supervised signals. This strategy comprises two components: a pixel-level local consistency constraint and a class-level prototype alignment constraint. The former emphasizes feature consistency at spatially corresponding positions across different views, while the latter enhances global semantic separability at the category level.

Given a target domain input image $\boldsymbol{X}_t$ and its corresponding annotation mask $\boldsymbol{Y}_t$, we apply random data augmentations (e.g., random cropping, rotation, blurring, and noise perturbation) to $\boldsymbol{X}_t$ to generate an augmented view

$$\boldsymbol{X}_t^{\text{aug}} = \mathcal{T}(\boldsymbol{X}_t) \tag{8}$$

where $\mathcal{T}(\cdot)$ denotes the data augmentation operation. Let $E(\cdot)$ denote the shared feature extractor obtained from the first stage, and $A^c(\cdot)$ denote the adapter corresponding to target category. Both the original and augmented views are processed by the frozen backbone and the corresponding CAS to obtain the adapted feature representations

$$\boldsymbol{F}_t = A^c(E(\boldsymbol{X}_t)) \tag{9}$$

$$\boldsymbol{F}_t^{\text{aug}} = A^c(E(\boldsymbol{X}_t^{\text{aug}})) \tag{10}$$

where $\boldsymbol{F}_t$ and $\boldsymbol{F}_t^{\text{aug}}$ denote the adapted feature maps. During the training process, only the parameters of the adapter A^c are updated, while the feature extractor $E(\cdot)$ remains frozen. This strategy effectively mitigates the risk of overfitting under few-shot conditions and maintains the stability of the cross-domain shared representations. We establish pixel-level supervisory signals between different views by constructing pixel-wise correspondences within the feature space $\boldsymbol{F}_t$ and $\boldsymbol{F}_t^{\text{aug}}$. These correspondences serve to construct positive pairs for the dense contrastive learning objective, constraining local representations to remain consistent across different views, as formulated in (19).

Sole reliance on local consistency may result in insufficient semantic discriminability at the class level. Consequently, we further construct prototypes to enhance global semantic separability. Utilizing the annotation mask, we partition the features into foreground and background regions and compute the foreground prototype and background prototype via masked average pooling

$$P_f = \frac{1}{|\Omega_f|} \sum_{u \in \Omega_f} f_t(u) \tag{11}$$

$$P_b = \frac{1}{|\Omega_b|} \sum_{u \in \Omega_b} f_t(u) \tag{12}$$

where $\Omega_f = \{u \mid Y_t(u) = 1\}$ and $\Omega_b = \{u \mid Y_t(u) = 0\}$ represent the sets of foreground and background pixels, respectively. Similarly, the prototypes P_f^{aug} and P_b^{aug} for the augmented view are obtained in the same manner. The prototype class alignment loss is formulated in (20).

To stabilize training and accumulate cross-sample semantic priors, we maintain a Memory Bank organized by category to cache class feature and prototype information. During the inference stage, the Memory Bank serves a dual purpose as follows. Dense matching: It facilitates similarity calculation between query image features and the class prototypes stored in the Bank. This determines the most relevant category for the current query, guiding the selection of the corresponding CAS for feature modulation and initial segmentation. Prompt generation: Based on the foreground and background prototypes of the selected category, similarity retrieval is performed on the query feature map to localize positive and negative prompt points, thereby identifying high-confidence geometric prompts

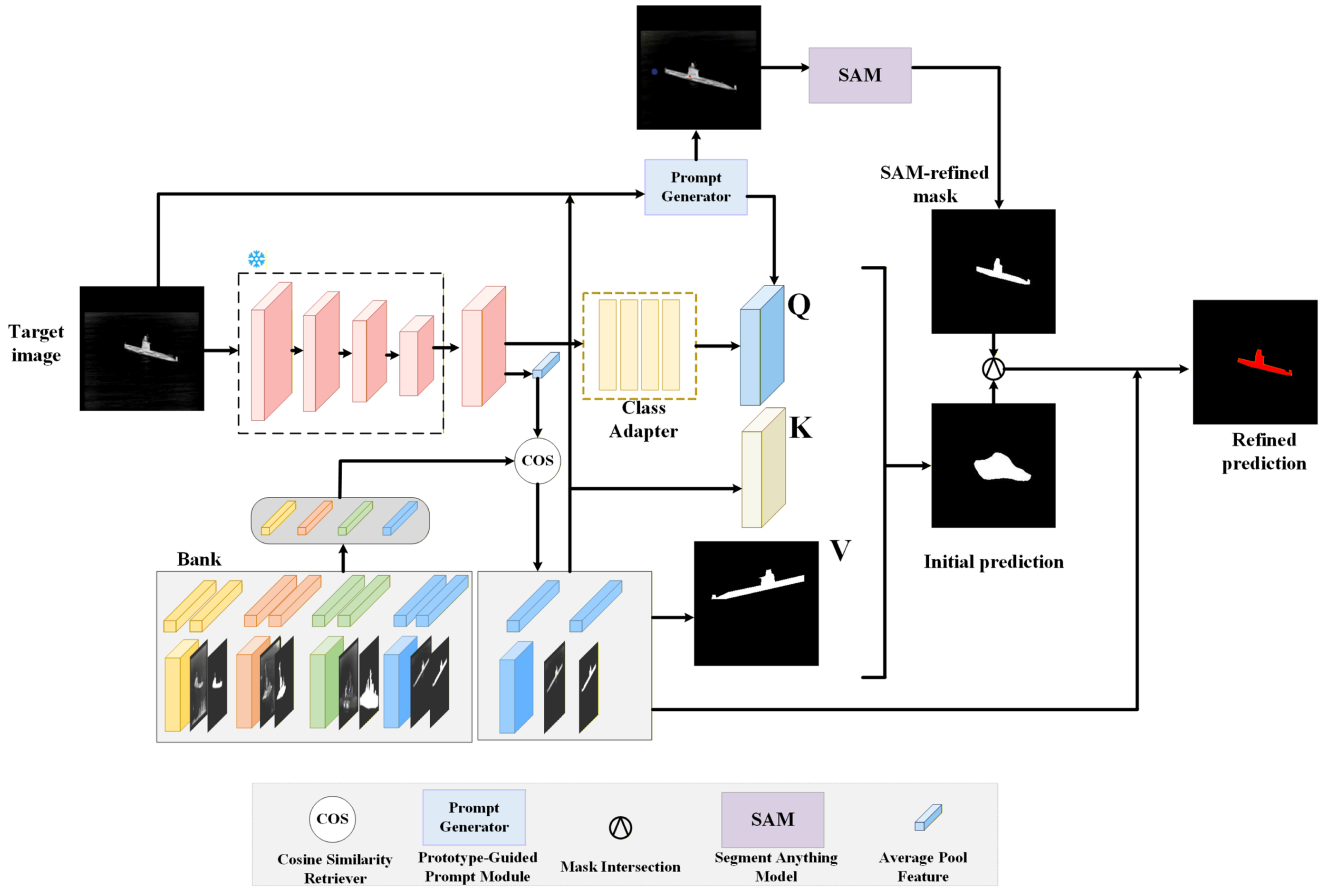

Fig. 5. Schematic of the inference stage.

for the subsequent SAM boundary refinement

$$P_f = \frac{1}{|\Omega_f|} \sum_{u \in \Omega_f} f_t(u). \tag{13}$$

D. Inference Stage

In the inference stage, we design a two-step prediction pipeline to balance semantic accuracy and boundary precision, as illustrated in Fig. 5.

Step 1—Initial semantic mask generation: First, we leverage the stored Memory Bank. Upon inputting a query image, we extract its features and perform dense matching against the class prototypes stored in the Memory Bank. This yields an initial probability map, which is subsequently thresholded to obtain a binary initial mask. While this mask possesses high semantic correctness, its boundaries may exhibit coarseness due to the resolution limits of the feature map.

Step 2—SAM-based boundary refinement: To enhance boundary quality and recover fine-grained structural details, we incorporate the SAM for mask refinement. The core innovation lies in our proposed prototype-driven prompt generation mechanism, which translates the learned class-level semantic knowledge into geometric prompts intelligible to SAM. Specifically, utilizing the foreground prototype p_i^f and background prototype p_i^b of the identified category, we perform a global similarity search on the query feature map f_q. The location exhibiting the highest similarity to the foreground prototype is selected as the Positive Prompt, while the location most similar to the background prototype serves as the negative prompt. This process is formulated as

$$\text{pos} = \arg\max_{(i,j)} \text{sim}\left(f_{\text{test}}^{(i,j)}, p_x^f\right) \tag{14}$$

$$\text{neg} = \arg\max_{(i,j)} \text{sim}\left(f_{\text{test}}^{(i,j)}, p_x^b\right) \tag{15}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity operation.

Subsequently, the query image, together with the generated positive and negative prompt points, is fed into SAM to yield a mask M_{sam} characterized by refined structures and boundaries. Finally, to fuse the semantic reliability of the initial prediction with the structural precision of the SAM refinement, the final segmentation mask M_{final} is derived from the pixelwise intersection of the two

$$M_{\text{final}} = M^q \cap M_{\text{sam}}^q. \tag{16}$$

E. Loss Functions

This section details the loss functions utilized to train the proposed two-stage framework. *Stage I:* The loss function in the first stage comprises a source domain segmentation loss

and a domain adversarial loss. The source domain segmentation loss ensures the supervised learning performance of the segmentation network on the source domain. Meanwhile, the domain adversarial loss encourages the segmentation network to learn more domain-invariant representations by employing a discriminator to distinguish between source and target domain features. The total loss function for the segmentation network is defined as

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N_s} \sum_{(\boldsymbol{X}_s, \boldsymbol{Y}_s) \in \mathcal{D}_s} \boldsymbol{Y}_s \log(G(\boldsymbol{X}_s)) \tag{17}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss on the source domain, $\boldsymbol{X}_s$ represents the source domain samples, $\boldsymbol{Y}_s$ denotes the corresponding ground truth labels, N_s is the number of source domain training samples, and $G(\cdot)$ represents the segmentation network.

The domain adversarial loss is formulated as

$$\begin{aligned} L_{\text{adv}} = {} & \mathbb{E}_{\boldsymbol{X}_s \sim \mathcal{D}_s}[\log D(G(\boldsymbol{X}_s))] \\ & + \mathbb{E}_{\boldsymbol{X}_t \sim \mathcal{D}_t}[\log(1 - D(G(\boldsymbol{X}_t)))] \end{aligned} \tag{18}$$

where $D(\cdot)$ denotes the discriminator network, $\boldsymbol{X}_t$ represents the target domain samples, and $\boldsymbol{X}_s$ represents the source domain samples.

In summary, the total loss for the first stage is defined as

$$L_{\text{first}} = L_{\text{CE}} + \alpha L_{\text{DA}} \tag{19}$$

where α is a weighting coefficient that controls the relative importance (tradeoff) between the cross-entropy segmentation loss and the domain adversarial loss.

Stage II: In the second stage, we employ a self-supervised contrastive learning strategy to fine-tune the CAS. The total loss comprises a dense self-supervised embedding alignment loss and a global class prototype alignment loss.

The dense self-supervised embedding alignment loss is formulated as

$$L_{nce} = -\frac{1}{H_f W_f} \sum_{i=1}^{H_f W_f} \log \frac{\exp\left(\sin\left(\boldsymbol{f}_i, \boldsymbol{f}_i^{\text{aug}}\right)/\tau\right)}{\sum_{j=1}^{H_f W_f} \exp\left(\sin\left(\boldsymbol{f}_i, \boldsymbol{f}_j^{\text{aug}}\right)/\tau\right)} \tag{20}$$

where $\boldsymbol{f}_i$ and $\boldsymbol{f}_i^{\text{aug}}$ denote the feature vector at the ith spatial position in the original view and the feature vector at the corresponding position in the augmented view, respectively; H_f and W_f represent the height and width of the feature map; τ is the temperature parameter.

The global class prototype alignment loss is defined as

$$L_p = -\log \frac{\exp\left(\text{sim}\left(p_f, p_f^{\text{aug}}\right)\right)}{\exp\left(\text{sim}\left(p_f, p_f^{\text{aug}}\right)\right) + \exp\left(\text{sim}\left(p_f, p_b^{\text{aug}}\right)\right)} \tag{21}$$

where p_f and p_b represent the foreground and background prototypes of the original view, respectively; p_f^{aug} and p_b^{aug} denote the corresponding prototypes computed from the augmented view.

The total loss for this stage is expressed as

$$L_{\text{second}} = L_{nce} + \beta L_p \tag{22}$$

Fig. 6. Examples of the VI-Ship dataset images.

where β is a hyperparameter used to balance the weights of the dense embedding alignment loss and the class prototype alignment loss.

IV. Experimental Results and Analysis

To systematically validate the effectiveness of the proposed two-stage CDFSS method, this section presents a series of experiments and analyses. First, we introduce the datasets, evaluation metrics, and specific implementation details adopted in the experiments. Subsequently, we compare the performance of our method against various CDFSS methods and UDA methods. Finally, we conduct ablation studies to analyze the actual contributions of key components within the model. All experiments in this article were implemented using the PyTorch deep learning framework on a hardware platform equipped with an NVIDIA RTX A4000 GPU.

A. Dataset

As illustrated in Fig. 6, the experiments utilize the dataset employed by Zhang et al. [19], with the source and target domains defined as follows: The source domain comprises 2961 VIS ship images and their corresponding pixel-level semantic labels, including a background class and four distinct ship categories. The target domain consists of 614 IR ship images; however, under the few-shot setting, only an extremely limited number of samples (K-shot) possess pixel-level labels. Fig. 6 displays representative samples from the dataset. Specifically, the first and second rows show VIS images from the source domain and their corresponding labels, exhibiting clear texture details; the third and fourth rows present IR images from the target domain and their corresponding labels. Specifically, the dataset encompasses four fine-grained ship categories. Furthermore, to ensure robustness across various thermal conditions, the IR images were collected at diverse imaging times, including morning, daytime, evening, and night scenarios, covering comprehensive background interference and temperature distributions.

B. Evaluation Metrics

To comprehensively evaluate the segmentation performance of the model on the target domain, we adopt two metrics commonly used in the field: mean Intersection over Union (mIoU) and pixel accuracy (PA) [5], [62], [63].

mIoU is a standard metric for semantic segmentation tasks. It measures the degree of overlap between the predicted segmentation region and the ground truth region by calculating the ratio of their intersection to their union [64], [65]. Assuming there are a total of K categories, for the kth category, the intersection over union (IoU) is defined as

$$\mathrm{IoU}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{23}$$

where TP_k, FP_k and FN_k denote the number of true positive, false positive, and false negative pixels for the kth class, respectively.

mIoU is obtained by calculating the average of the IoU values across all K categories

$$\mathrm{mIoU} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{IoU}_k. \tag{24}$$

This metric provides a comprehensive assessment of segmentation performance across all categories. It exhibits robustness against class imbalance issues and effectively reflects the overall quality of the model's segmentation.

PA measures the proportion of correctly classified pixels relative to the total number of pixels and is expressed as

$$\mathrm{PA} = \frac{\sum_{k=1}^{K} \mathrm{TP}_k}{\sum_{k=1}^{K} (\mathrm{TP}_k + \mathrm{FP}_k)}. \tag{25}$$

Based on PA, the mean Pixel Accuracy (mPA) is calculated by averaging the PA across all K categories. This provides a balanced evaluation metric that equally weights the accuracy of each class, making it robust against category imbalance in the test set.

C. Implementation Details

(a) Network architecture: CAS-Net primarily comprises three components: a feature extraction network, a discriminator network, and CAS. The feature extraction network utilizes ResNet-50 as its backbone. The CAS adopt a lightweight bottleneck architecture. In the adversarial training of the first stage, the discriminator network is employed to distinguish whether features originate from the source domain or the target domain. This network adopts a five-layer convolutional structure. All convolution kernels are set to a size of 4×4 with a stride of 2, and the channel numbers for the layers are 64, 128, 256, 512, and 1, respectively. The last layer uses a single output channel because the discriminator performs domain classification and predicts a domain probability. Finally, the discrimination result is output through an activation function.

(b) Parameter settings: The training process of the proposed model is divided into two stages, with specific parameter settings as follows: In the first stage, both the feature extraction network and the discriminator network are optimized using the Adam optimizer. The initial learning rate η for the feature extraction network is set to 2.5×10^{-4}, with a weight decay of 1×10^{-5}. The initial learning rate for the discriminator network is set to 1×10^{-4}. In the second stage, the backbone parameters of the feature extraction network are frozen, and only the adapters are fine-tuned. The adapters are optimized using the Adam optimizer with an initial learning rate of 5×10^{-5}. This stage aims to adjust the high-level feature distribution via the adapters to adapt to the semantic space of the IR domain.

TABLE I
SEGMENTATION RESULTS OF DIFFERENT α

α	0.5	0.6	0.7	0.8	0.9	1
mIoU(%)	61.44	61.87	62.92	**63.19**	63.05	62.59

TABLE II
SEGMENTATION RESULTS OF DIFFERENT β

β	0.5	0.6	0.7	0.8	0.9	1
mIoU(%)	61.37	62.09	62.43	**63.19**	62.52	61.20

(c) Hyperparameter sensitivity analysis: To investigate the influence of the balancing factors α and β on model performance and determine their optimal values, we designed the following experiments.

Sensitivity analysis for α: In the first-stage loss function, α governs the tradeoff between the backbone cross-entropy loss and the domain adversarial loss. We selected the candidate set $\{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and determined the optimal value by comparing the mIoU results on the validation set. The results are presented in Table I.

As observed in Table I, the highest mIoU and optimal segmentation performance are achieved when $\alpha = 0.8$. Consequently, $\alpha = 0.8$ is adopted for all subsequent experiments.

Sensitivity analysis for β: In the second-stage loss function, β balances the weights of the self-supervised contrastive loss and the class prototype alignment loss. We utilized the candidate set $\{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and compared the mIoU results on the validation set to select the value. The results are shown in Table II.

As observed in Table II, when β is relatively small (e.g., 0.5), the model tends to over-rely on pixel-level self-supervised contrastive learning. While this enhances local consistency, the lack of strong constraints from global class prototypes results in loose semantic modeling for novel categories, thereby increasing susceptibility to class confusion. Conversely, when β is relatively large (e.g., 1.0), the model places excessive emphasis on global alignment with the few-shot prototypes. Due to the extreme scarcity of samples in the support set, the computed prototypes may inevitably exhibit bias. Excessive constraints in this scenario cause the adapter to overfit specific support samples, neglecting the rich internal structural details of the images, which consequently degrades the mIoU to 61.20%.

D. Comparison With State-of-The-Art Methods

In this section, we conduct a comparative analysis of CAS-Net against existing FSS methods and CDFSS methods on the VI-Ship dataset. Experiments are conducted under one-shot & two-shot & three-shot & five-shot & ten-shot settings. The quantitative results are presented in Table III, where bold numbers indicate the best results. Partial visualization results are shown in Fig. 7.

TABLE III
QUANTITATIVE PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE VI-SHIP DATASET

Methods	mIoU					mPA				
	One-shot	Two-shot	Three-shot	Five-shot	Ten-shot	One-shot	Two-shot	Three-shot	Five-shot	Ten-shot
Few-shot Segmentation Methods										
PGNet	34.85	34.89	35.05	34.91	34.98	40.89	42.24	43.56	46.89	47.93
PANet	35.96	35.99	36.23	37.23	37.41	32.96	39.45	42.91	45.76	49.68
PFENet	37.58	37.88	38.25	39.56	39.77	45.89	47.56	50.59	51.86	55.74
HSNet	38.54	39.02	39.24	41.08	42.42	35.96	37.86	39.45	42.46	48.96
SSP	39.41	39.71	40.52	42.87	42.96	39.54	41.89	43.12	48.92	50.23
Cross-domain Few-shot Segmentation Methods										
PATNet	47.63	47.85	48.75	52.24	53.96	45.21	47.96	48.23	50.52	51.89
RestNet	48.10	48.51	48.88	51.69	53.12	45.25	49.65	52.69	55.69	60.58
APSeg	58.76	58.93	59.08	59.54	60.77	51.25	64.89	65.96	69.10	68.42
PMNet	53.89	54.23	54.42	55.95	56.07	62.59	63.45	65.91	66.87	68.24
ABCDFSS	54.07	54.74	55.13	55.24	56.48	64.28	65.82	67.20	68.71	69.05
TGCM	55.12	55.92	56.42	58.89	59.52	57.93	60.59	65.91	66.51	69.62
IFA	56.36	56.89	57.62	58.74	59.10	54.26	55.96	60.25	61.85	63.96
ISA	53.21	53.88	54.75	55.34	56.03	63.89	64.70	66.97	67.01	68.23
LLF	60.04	60.89	61.21	62.17	62.88	62.18	66.07	68.38	69.09	70.05
DATO	58.59	58.85	59.02	60.54	61.08	62.84	64.73	67.22	68.43	69.29
CAS-Net	**63.19**	**63.26**	**63.75**	**64.28**	**64.86**	**65.76**	**70.56**	**71.03**	**72.61**	**73.96**

From Table III and Fig. 7, it can be drawn that the following holds.

1) Under the one-shot setting, CAS-Net achieves an mIoU of 63.19%, consistently outperforming the FSS methods. Specifically, methods, such as PGNet, PANet, and SSP, generally yield mIoU scores below 40% in the one-shot setting. Even as the number of support set samples increases to 10, the performance improvement of these methods remains marginal, failing to fundamentally overcome the bottleneck of domain discrepancy. This demonstrates that in the absence of effective domain adaptation mechanisms, simply increasing the number of support samples cannot bridge the substantial semantic gap between VIS and IR images. In contrast, our method introduces cross-domain feature extraction in the first stage, effectively alleviating domain discrepancies.
2) Compared to existing CDFSS methods, CAS-Net achieves mIoU scores of 63.19%, 63.26%, 63.75%, 64.28%, and 64.86% under one-shot & two-shot & three-shot & five-shot & ten-shot settings, respectively. For example, in the most challenging 1-shot setting, our method (63.19%) significantly outperforms the second-best method, which is now the recently proposed LLF (60.04%), by 3.15%. In the 10-shot setting, our mPA (73.96%) is 3.91% higher than that of LLF (70.05%). This indicates that the proposed two-stage framework possesses stronger generalization capabilities under extremely few-shot conditions, even when compared to the most cutting-edge models.
3) In conclusion, CAS-Net achieves the highest mIoU and PA across all shot settings, with the advantage being particularly pronounced in the extreme one-shot scenario. Furthermore, visualization results demonstrate that the segmentation maps generated by our method exhibit clearer and more complete boundaries, thereby validating the effectiveness of the proposed two-stage cross-domain few-shot IR ship semantic segmentation framework.

E. Analysis of Model Parameters and Computational Efficiency

In addition to segmentation accuracy, computational complexity (measured in FLOPs) and parameter count (Parameters) are crucial metrics for assessing a model's potential for real-world deployment [66], [67], [68]. This section summarizes the resource overhead of various methods during the training phase and provides a comprehensive comparison in conjunction with the one-shot mIoU performance. The results are presented in Table IV and Fig. 8. As illustrated in Fig. 8, CAS-Net resides in the top-left region of the chart. This indicates that the model achieves a balance between performance and efficiency, maintaining reasonable computational overhead while upholding high segmentation accuracy.

It can be observed from Table IV that compared to computationally intensive methods such as APSeg (40.96 M parameters, 57.51 GFLOPs) and PFENet (39.62 M parameters, 60.24 GFLOPs), our method (33.62 M parameters, 47.88 GFLOPs)

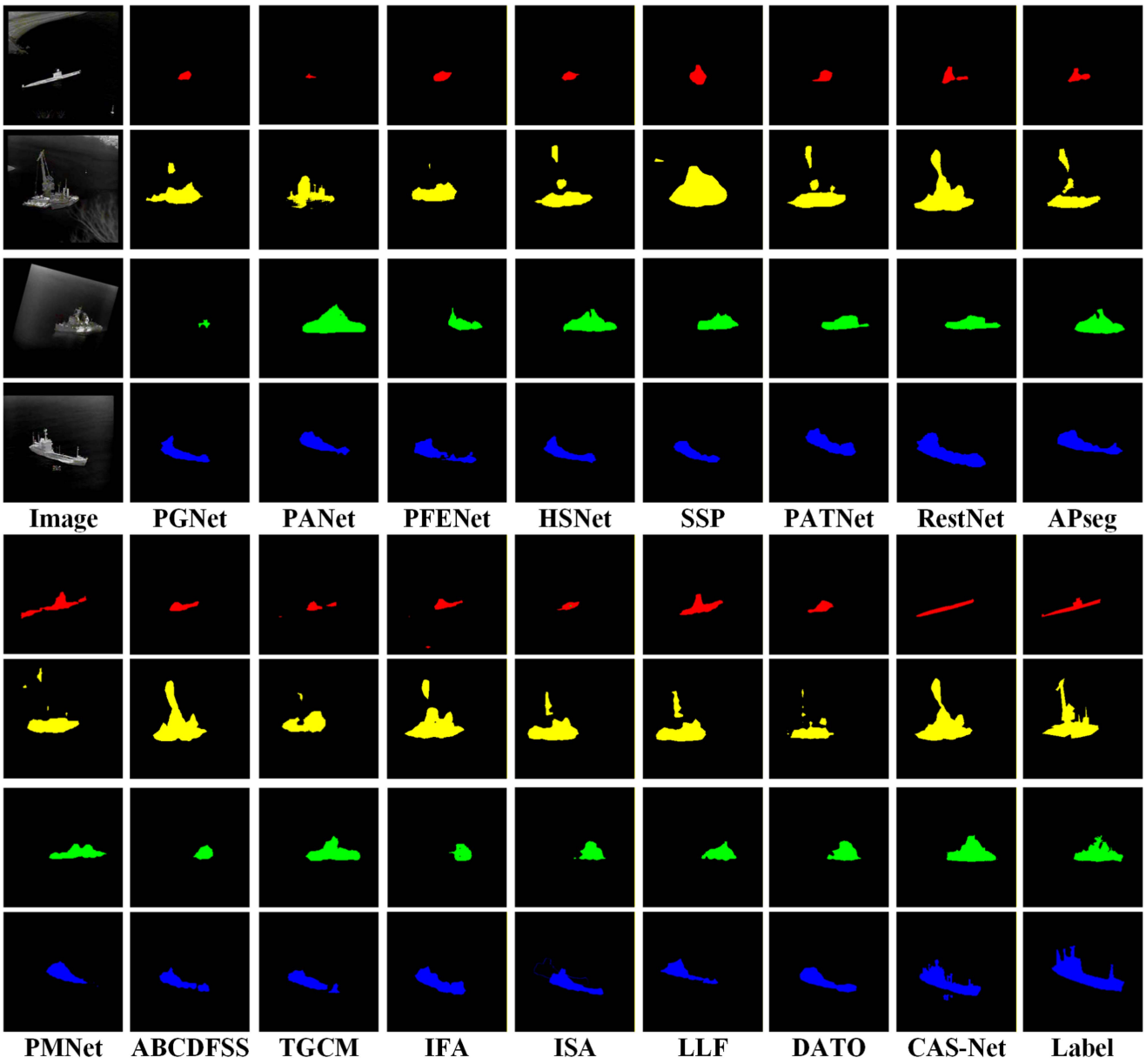

Fig. 7. Qualitative visualization of segmentation results on the VI-Ship dataset.

TABLE IV
ONE-SHOT MIOU, COMPUTATIONAL LOAD, AND PARAMETERS OF COMPARED METHODS

Methods	One-shot mIoU	FLOPs (GFLOPs)	Parameters (M)
PGNet	34.85	45.06	**22.40**
PANet	35.96	47.23	24.50
PFENet	37.58	60.24	39.62
HSNet	38.54	**32.51**	29.10
SSP	39.41	51.21	30.60
PATNet	47.63	53.77	33.20
RestNet	48.10	43.98	35.80
APSeg	58.76	57.51	40.96
PMNet	53.89	53.84	34.60
ABCDFSS	54.07	53.01	35.10
TGCM	55.12	58.62	34.90
IFA	56.36	48.69	32.80
CAS-Net	**61.44**	47.88	33.62

exhibits a significant reduction in both parameter count and computational cost. Conversely, the segmentation accuracy (mIoU) is improved by 2.68% and 23.86%, respectively. This suggests that merely increasing model parameters does not necessarily yield improvements in cross-domain performance; rather, rational architectural design is paramount.

During the second training stage, the parameters of the massive backbone network are frozen, and only the lightweight CAS are updated. This design not only circumvents the substantial computational burden associated with full-parameter fine-tuning of the entire network but also effectively prevents overfitting to target domain data under few-shot conditions. Consequently, significant performance gains are achieved at a lower parameter cost.

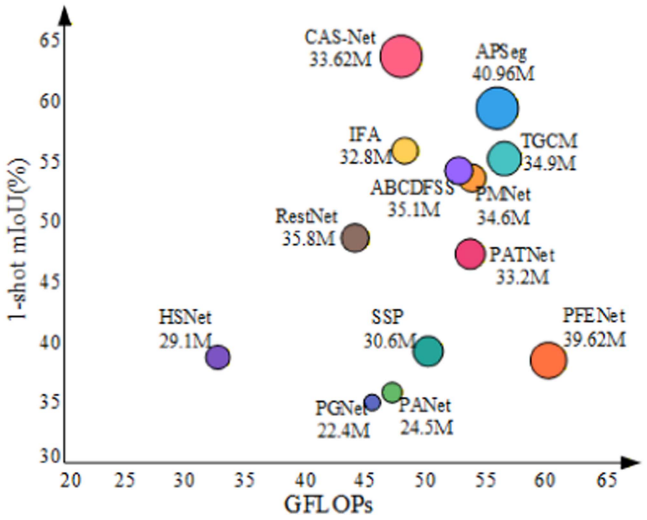

Fig. 8. Comparison of computational efficiency and model size among different methods. The horizontal axis represents computational cost (GFLOPs), and the size of each circle corresponds to the number of parameters (Params). Different colors indicate different methods.

TABLE V
PERFORMANCE COMPARISON WITH UDA METHODS

Methods	mIoU(%)	mPA(%)	FLOPs (GFLOPs)	Parameters (M)
MDA-Net	58.56	50.63	49.05	43.36
ADVENT	54.59	55.63	**47.80**	40.87
CPSL	58.32	57.96	48.23	40.78
Sepico	61.39	62.95	48.25	42.95
T-DANet	60.09	60.89	49.77	44.69
CAS-Net	**63.19**	**65.76**	47.88	**33.62**

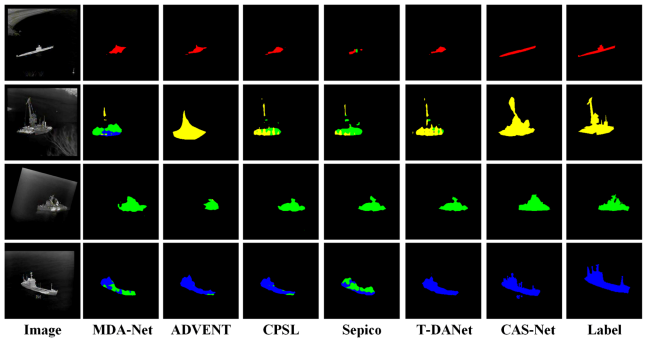

Fig. 9. Qualitative comparison of segmentation results against various UDA methods.

F. Comparison With UDA Methods

To validate the effectiveness of our method under extremely data-scarce conditions, this section compares its performance against several classic UDA methods. Quantitative comparison results are presented in Table V, and visualization results are shown in Fig. 9. It is worth noting that UDA methods utilize all unlabeled data from the target domain for training, whereas our method employs only one-shot labeled samples.

As observed in Table V and Fig. 9 the following holds.

1) In terms of segmentation accuracy, CAS-Net achieves an mIoU of 63.19% using only one-shot annotation, surpassing all compared UDA methods. Specifically, compared to the best-performing UDA method, Sepico, our method improves mIoU by 1.80% and mPA by 2.81%. Compared to ADVENT, mIoU and mPA are improved by 8.6% and 10.13%, respectively. This indicates that in VIS-IR scenario, a small amount of precise supervisory signals is more effective than adaptive alignment based on large-scale unlabeled data.
2) From the visualization results, as shown in the second row of Fig. 9, CAS-Net yields the most complete segmentation, whereas other UDA methods exhibit issues, such as missing hull regions or fragmented boundaries. UDA methods typically require a substantial number of target domain images to estimate statistical distribution features, whereas our method achieves superior performance with only one sample per class. This demonstrates the high practical value of the proposed framework, particularly for tasks, such as IR ship segmentation, where data acquisition costs are high and annotation is difficult.
3) In summary, the proposed method not only outperforms UDA methods that rely on large-scale data in terms of performance but also possesses an absolute advantage in terms of data cost. Experimental results verify that through rational cross-domain feature decoupling and few-shot adaptation strategies, high-precision semantic segmentation can be achieved in the IR target domain where data are extremely scarce.

G. Impact of Different Components

To quantify the contribution of each key component in CAS-Net, we conduct an ablation study by selectively enabling the wavelet branch, class adapters, and the inference-stage SAM Refinement module while keeping the remaining settings unchanged. Table VI reports the resulting mIoU and mPA on the target domain under different component configurations.

1) Class adapters are pivotal for performance enhancement. Upon introducing class adapters to the Baseline (Exp. ii), mIoU surged from 30.75% to 53.12%, representing a substantial increase of 22.37%. This demonstrates that in cross-domain few-shot scenarios, relying solely on the backbone for feature extraction is insufficient; targeted adaptation modules are indispensable for rectifying semantic deviations.
2) The wavelet branch effectively bolsters the robustness of cross-domain features. Comparing Exp. (i) with the Baseline, and Exp. (iv) with Exp. (ii), the incorporation of the Wavelet Branch yielded mIoU improvements of 1.31% and 0.95%, respectively. This indicates that frequency domain information serves as an effective complement to spatial features, thereby enhancing the model's fundamental discriminative capability.

TABLE VI
ABLATION STUDY OF KEY MODEL COMPONENTS

Methods	Wavelet Branch	Class Adapters	SAM Refinement	mIoU(%)	mPA(%)
Baseline				30.75	31.63
(i)	✓			32.06	33.48
(ii)		✓		53.12	54.32
(iii)			✓	30.85	29.26
(iv)	✓	✓		54.07	55.64
(v)	✓		✓	32.45	34.72
(vi)		✓	✓	62.03	62.51
(vii)	✓	✓	✓	**63.19**	**65.76**

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT FEATURE ENHANCEMENT MODULES

Modules	w/o Wavelet	Gated-SCNN	ACNet	DANet	APM	WFE
One-shot	62.03	62.74	59.81	62.54	62.72	**63.19**
Five-shot	63.15	63.52	60.76	63.21	63.99	**64.28**
Ten-shot	63.51	63.64	61.29	63.99	64.02	**64.86**

3) The efficacy of the SAM refinement module is highly contingent upon adapter guidance, exhibiting significant synergy. When SAM was employed in isolation during the inference stage (Exp. iii), performance showed negligible improvement over the Baseline. This is fundamentally because directly applying SAM to the original IR images without targeted semantic guidance leads to severe oversegmentation. SAM behaves as a class-agnostic foundation model; in complex IR scenes, it tends to indiscriminately segment irrelevant salient regions (e.g., sea waves, vessel wakes, or thermal background artifacts) due to the complete lack of target-domain class priors. However, when combined with Class Adapters (Exp. vi), mIoU further increased significantly by 8.91% (reaching 62.03%) over the adapter-only configuration. This underscores that SAM requires high-quality and class-specific semantic initial masks as prompts to function effectively.
4) The full model achieves optimal performance through multilevel synergy. Upon simultaneously integrating all three modules (Exp. vii), the model attained an mIoU of 63.19%, reflecting a total improvement of 32.44% compared to the Baseline. This confirms that the three proposed components complement one another at distinct levels, collectively constituting a comprehensive closed-loop framework ranging from feature extraction to semantic adaptation and detailed refinement.

H. Impact Analysis of Feature Enhancement Modules

To validate the effectiveness of our WFE module in cross-domain few-shot segmentation, we compare it with representative enhancement strategies (e.g., Gated-SCNN, ACNet, DANet, and APM). For a fair comparison, all methods are implemented with the same backbone and training protocol, and only the feature enhancement branch is replaced. Table VII presents the mIoU results under one-shot, five-shot, and ten-shot settings.

1) The proposed WFE achieves the best performance across all settings. Under one-shot, five-shot, and ten-shot settings, our method achieves mIoU scores of 63.19%, 64.28%, and 64.86%, respectively. Compared to the baseline (without any enhancement structure), our method yields improvements of 1.16%, 1.13%, and 1.35%, respectively. This indicates that explicitly introducing frequency domain information can effectively compensate for the deficiencies of purely spatial convolution when handling data with significant domain discrepancies.
2) Compared to spatial or channel attention mechanisms (e.g., DANet), the Wavelet method is better suited for cross-domain scenarios. Although DANet achieves an mIoU of 63.99% under the ten-shot setting, it remains inferior to our method. This suggests that traditional attention mechanisms exhibit limitations in cross-domain generalization. Furthermore, ACNet performs poorly, with an mIoU of only 59.81% under the one-shot setting. This demonstrates that merely altering the receptive field or enhancing the backbone via asymmetric convolution fails to bridge the semantic gap induced by domain shifts.
3) Compared to frequency-based methods, such as APM, our wavelet branch is more fine-grained. Although APM also utilizes frequency domain information and outperforms the Baseline (one-shot mIoU: 62.72%), it falls short of our method. This is attributed to the fact that APM is typically based on the Fourier transform or simple frequency domain masks, which are prone to losing spatial positional information.
4) In summary, the designed wavelet branch successfully exploits domain invariance within the frequency domain, providing a more robust feature basis for the subsequent adapter fine-tuning.

I. Impact Analysis of Different Adapter Architecturess

To verify the effectiveness of the proposed CAS during the few-shot fine-tuning stage, we designed ablation studies focused on adapter structures in this section. We compared our method against mainstream lightweight adapter approaches, including representative methods such as Single-Adapter, AdapterHub, AdaptFormer, and Conv-Adapter. To ensure fairness, all comparative experiments were conducted using the identical backbone network and training pipeline, with only the class adaptation module being replaced. As observed in Table VIII the following holds.

TABLE VIII
PERFORMANCE COMPARISON OF DIFFERENT ADAPTER ARCHITECTURES

Methods	Single-Adapter	AdapterHub	AdaptFormer	Conv-Adapter	CAS
1-shot	55.74	61.54	60.12	55.89	**63.19**
5-shot	56.21	62.87	61.92	57.56	**64.28**
10-shot	56.83	63.51	62.44	58.17	**64.86**

1) The proposed CAS achieves the highest accuracy across all settings. Under one-shot, five-shot, and ten-shot settings, its mIoU reached 63.19%, 64.28%, and 64.86%, respectively. Compared to the baseline method using a single shared adapter (Single-Adapter), our method improved mIoU by 7.45%, 8.07%, and 8.03%, respectively. This indicates that under extremely few-shot conditions, decoupling parameter sharing between categories is crucial for improving segmentation performance.
2) Compared to generic lightweight adapters (e.g., AdapterHub and AdaptFormer), our method demonstrates superior efficacy. The one-shot mIoU scores for AdapterHub and AdaptFormer are 61.54% and 60.12%, respectively. Although they outperform the Single-Adapter, they remain inferior to our method. Specifically, in the one-shot scenario, they are lower than our method by 1.65% and 3.07%, respectively; in the five-shot scenario, the gaps are 1.41% and 2.36%; and in the ten-shot scenario, the gaps are 1.35% and 2.42%. This suggests that discrepancies arise when adapters originally designed for Transformer architectures are directly transferred to CNN architectures.
3) Compared to classification-oriented adapters (e.g., Conv-Adapter), our method performs significantly better. Conv-Adapter exhibited the poorest performance among all methods, with a one-shot mIoU of only 55.89%. This indicates that simple classification adapters are ill-suited for handling complex dense prediction tasks.
4) In summary, the proposed CAS successfully addresses the limitations of generic adapters in cross-domain few-shot segmentation tasks. It not only avoids interclass interference but also maximally preserves spatial structural information conducive to segmentation.

J. Performance Comparison on the Public Agriculture-Vision Dataset

To further validate the generalization capability of the proposed method, we conducted an evaluation on Agriculture-Vision [69], a public agricultural remote sensing dataset depicting a scenario distinct from the ship detection task. This dataset comprises a vast collection of high-resolution aerial farmland imagery characterized by diverse surface crop types and complex textural features, thereby posing a rigorous challenge to the model's cross-domain and few-shot adaptation capabilities. Comparative experiments were conducted against various FSS and CDFSS methods. The quantitative results are presented in Table IX, and visualization results are shown in Fig. 10.

From Table IX and Fig. 10, the following can be obtained.

1) We selected representative methods such as PGNet, PANet, PFENet, HSNet, and SSP for comparison. The experimental results indicate that although the performance of these methods improves as the sample size increases from one-shot to ten-shot, their overall performance encounters a bottleneck. Even SSP, the top-performing method among them, achieves an mIoU of only 26.78% under the one-shot setting and reaches only 31.25% under the ten-shot setting.
2) Under the one-shot setting, CAS-Net attains an mIoU of 42.12%, substantially surpassing the comparative methods. As the sample size increases to ten-shot, our mIoU rises to 45.82%, with an mPA of 69.95%. It is worth noting that the newly introduced methods achieve strong performance among the baselines. However, compared to LLF, which achieves the second-best one-shot performance, our method still improves mIoU by 2.13%. In the ten-shot scenario, our method outperforms the second-best method (ISA) by 1.65% in mIoU, and surpasses LLF by 2.58% in mPA. This demonstrates that, compared to the latest state-of-the-art cross-domain approaches, CAS-Net more effectively overcomes the domain shift from natural scenes to agricultural remote sensing scenes.
3) Combining the quantitative numerical improvements with the qualitative visualization maps (as shown in Fig. 10), the practical efficacy of our method is clearly observable. Existing few-shot methods (e.g., PANet) and cross-domain few-shot methods (e.g., RestNet) often generate fragmented masks when processing farmland regions and struggle to distinguish between different crop areas with similar textures. In contrast, our method produces more accurate segmentation results with sharper contours.
4) In conclusion, the proposed method exhibits excellent cross-scenario generalization capabilities. Whether in IR ship segmentation tasks that emphasize object contours or in agricultural remote sensing segmentation tasks that emphasize surface textures, our method maintains robust high performance. This demonstrates the practical value of the framework in addressing complex cross-domain semantic segmentation tasks.

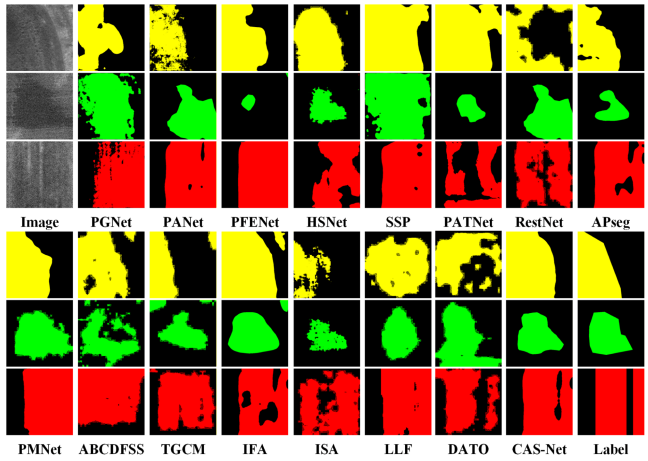

Fig. 10. Visual comparison of segmentation performance on the Agriculture-Vision dataset.

TABLE IX
RESULTS OF DIFFERENT METHODS ON THE AGRICULTURE-VISION DATASET

Methods	mIoU					mPA				
	1-shot	2-shot	3-shot	5-shot	10-shot	1-shot	2-shot	3-shot	5-shot	10-shot
Few-shot Segmentation Methods										
PGNet	22.12	23.56	24.81	26.91	27.15	50.26	51.89	52.63	55.69	60.69
PANet	23.45	24.27	25.62	27.26	28.62	51.19	52.61	53.42	55.21	58.23
PFENet	23.95	24.61	24.84	28.36	29.74	45.29	47.69	47.92	48.65	52.96
HSNet	24.29	25.41	26.08	28.07	30.54	42.91	43.13	43.59	47.96	50.57
SSP	26.78	27.08	27.59	29.63	31.25	43.51	44.66	45.71	51.98	57.62
Cross-domain Few-shot Segmentation Methods										
PATNet	33.97	34.25	35.92	36.59	37.41	48.67	49.56	53.62	55.74	58.62
RestNet	34.52	35.62	37.05	38.24	40.22	47.12	48.62	49.05	52.97	53.83
APSeg	35.19	36.93	38.36	40.32	41.88	48.62	48.99	49.25	51.94	57.92
PMNet	36.94	37.52	38.41	40.35	41.36	50.77	51.85	52.09	54.44	56.28
ABCDFSS	39.52	40.66	40.92	42.63	43.57	55.26	59.62	60.23	63.51	64.85
TGCM	38.95	39.36	40.23	41.89	42.12	42.94	43.65	44.25	46.32	47.65
IFA	37.95	38.43	39.03	41.07	43.09	52.69	53.91	55.39	57.27	59.23
ISA	34.92	38.74	40.52	42.66	44.17	48.63	54.09	56.13	60.31	63.75
LLF	39.99	41.36	42.53	43.78	44.12	50.74	53.54	59.95	59.84	67.37
DATO	38.75	40.37	41.03	42.39	43.21	51.79	52.67	57.44	61.58	64.39
CAS-Net	**42.12**	**43.95**	**44.29**	**44.93**	**45.82**	**58.69**	**60.52**	**63.85**	**65.78**	**69.95**

V. DISCUSSION

While the proposed CAS-Net demonstrates superior performance in cross-domain few-shot IR ship segmentation, several aspects warrant further discussion. First, regarding the modality gap, our framework largely alleviates VIS-to-IR discrepancies through wavelet-based frequency enhancement and adapter tuning. However, extreme maritime weather conditions (e.g., heavy sea fog or rainstorms) can still degrade IR contour visibility, potentially affecting the prototype matching reliability. Second, regarding computational efficiency, although the adapter fine-tuning process is highly parameter-efficient, the introduction of the SAM during the inference phase inherently increases the overall computational burden and memory footprint. This presents a tradeoff between segmentation precision especially boundary refinement and real-time deployment constraints on edge devices. Future research will explore lightweight variants of SAM or knowledge distillation techniques to compress the boundary refinement module, striving to maintain the proposed model's high accuracy while satisfying the stringent latency requirements of real-world maritime surveillance systems.

VI. CONCLUSION

In this article, we propose a two-stage CDFSS framework. In the first stage, we design a dual-branch network integrating wavelet transform and convolution. In the second stage, we design independent CAS and perform efficient fine-tuning on extremely limited samples via a self-supervised contrastive learning strategy. Furthermore, we incorporate the SAM as a postprocessing module. By utilizing a prototype-driven prompt generation mechanism, we refine the initial segmentation masks, significantly enhancing boundary accuracy and detail representation. Experimental results on the VI-Ship and Agriculture-Vision datasets demonstrate that our method outperforms comparative approaches across all metrics, with particularly notable improvements in the extreme one-shot setting, thereby validating the effectiveness and generalization capability of the proposed framework. Despite the significant performance breakthroughs achieved, the introduction of the large-scale foundation model SAM inevitably increases the system's computational complexity and memory overhead, limiting its real-time inference capabilities on resource-constrained edge devices. In the future, we will conduct research on model lightweighting and compression to reduce inference costs while maintaining performance.

REFERENCES

[1] L. Li, L. Liu, F. Cheng, Y. He, and Z. Zhong, "CN-UNet: ConvNeXt UNet with slicing-aided hyper segmentation for infrared small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 84–98, 2026, doi: 10.1109/JSTARS.2025.3631384.

[2] Y. Luo, M. Zhu, T. Chen, and Z. Zheng, "Remaining useful life prediction for stratospheric airships based on a channel and temporal attention network," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 143, Art. no. 108634, 2025, doi: 10.1016/j.cnsns.2025.108634.

[3] F. Chen et al., "MCLL-Diff: Multiconditional low-light image enhancement based on diffusion probabilistic models," *IEEE Sensors J.*, vol. 25, no. 6, pp. 9912–9924, Jun. 2025, doi: 10.1109/JSEN.2025.3534566.

[4] J. Jia et al., "The effect of artificial intelligence evolving on hyperspectral imagery with different signal-to-noise ratio, spectral and spatial resolutions," *Remote Sens. Environ.*, vol. 311, 2024, Art. no. 114291, doi: 10.1016/j.rse.2024.114291.

[5] J. He, Y. Wang, L. Deng, Y. Li, and Z. Li, "Phase congruency image mosaicking approach for aerial mid-wave infrared low-overlap array scanning images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 227, pp. 185–204, 2025, doi: 10.1016/j.isprsjprs.2025.06.007.

[6] D. Zhang, Y. Hu, S. Zhang, and Y. Zhang, "Distributed hierarchical reinforcement learning for dynamic maintenance scheduling of large-scale airline fleets," *Rel. Eng. Syst. Saf.*, vol. 271, 2026, Art. no. 112249, doi: 10.1016/j.ress.2026.112249.

[7] C. HaiChao, H. Wenlong, L. Zuyuan, F. Baiwei, and Z. Qiang, "Enhancing surrogate model accuracy in ship design optimization through intelligent constraint-aware sample selection," *Eng. Appl. Artif. Intell.*, vol. 163, 2026, Art. no. 112716, doi: 10.1016/j.engappai.2025.112716.

[8] S. Chen, X. Long, J. Fan, and G. Jin, "A causal inference-based root cause analysis framework using multi-modal data in large-complex system," *Rel. Eng. System Saf.*, vol. 265, 2026, Art. no. 111520, doi: 10.1016/j.ress.2025.111520.

[9] T. Sun et al., "RID-LIO: Robust and accurate intensity-assisted LiDAR-based SLAM for degenerated environments," *Meas. Sci. Technol.*, vol. 36, no. 3, 2025, Art. no. 036313, doi: 10.1088/1361-6501/adb769.

[10] P. Wu et al., "Fine-grained object recognition algorithm based on self-supervised representation learning of multi-view infrared images," *IEEE Trans. Instrum. Meas.*, vol. 75, 2026, Art. no. 9515418, doi: 10.1109/TIM.2026.3673752.

[11] T. Luo et al., "DiffW: Multi-encoder based on conditional diffusion model for robust image watermarking," *IEEE Trans. Multimedia*, vol. 28, pp. 837–852, 2026, doi: 10.1109/TMM.2025.3632631.

[12] R. Yin, J. Peng, Y. Cai, C. Wu, B. Champagne, and N. Al-Dhahir, "Intelligent 3D trajectory and resource control for multi-UAV 6G networks via GNN and deep unfolding," *IEEE Trans. Commun.*, p. 1, 2026, doi: 10.1109/TCOMM.2026.3655771.

[13] L. Zhou, J. Gao, and D. Li, "An engineering method for simulating dynamic interaction of moored ship with first-year ice ridge," *Ocean Eng.*, vol. 171, pp. 417–428, 2019, doi: 10.1016/j.oceaneng.2018.11.027.

[14] L. Cheng et al., "Fast prediction of underwater vehicle motion under internal solitary waves," *Phys. Fluids*, vol. 38, no. 2, 2026, Art. no. 027115, doi: 10.1063/5.0315969.

[15] C. Wang et al., "Sentinel-1 wave mode SAR monitoring of icebergs around antarctica," *Remote Sens. Environ.*, vol. 332, 2026, Art. no. 115094, doi: 10.1016/j.rse.2025.115094.

[16] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2517–2526.

[17] A. Kumar, T. Ma, and P. Liang, "Understanding self-training for gradual domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5468–5479.

[18] J. Xu et al., "How to characterize imprecision in multi-view clustering," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 9, no. 6, pp. 3821–3832, Dec. 2025, 2025, doi: 10.1109/TETCI.2025.3572135.

[19] T. Zhang, H. Shen, S. U. Rehman, Z. Liu, Y. Li, and O. U. Rehman, "Two-stage domain adaptation for infrared ship target segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4208315.

[20] L. Gan, C. Lee, and S.-J. Chung, "Unsupervised RGB-to-thermal domain adaptation via multi-domain attention network," in *Proc. IEEE Int. Conf. Roboti. Autom.*, London, U.K., 2023, pp. 6014–6020, doi: 10.1109/ICRA48891.2023.10160872.

[21] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9197–9206.

[22] J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6941–6952.

[23] M. Sun et al., "Swin-UNETR: A transformer-based model for 3D pore network segmentation in low-permeability sedimentary rocks," *Int. J. Coal Geol.*, vol. 315, 2026, Art. no. 104951, doi: 10.1016/j.coal.2026.104951.

[24] F. Wang et al., "DABF-Net: A dual-branch attention-guided and bi-directional feature enhancement network for infrared small-target detection with air-to-ground benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5005014, doi: 10.1109/TGRS.2025.3586362.

[25] L. Li et al., "KECS-Net: Knowledge-embedded CSwin-UNet with slicing-aided hypersegmentation for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 23, 2026, Art. no. 7000105, doi: 10.1109/LGRS.2025.3632827.

[26] Y. Shou, T. Meng, W. Ai, H. Liu, and K. Li, "Graph information bottleneck for remote sensing segmentation," *Neurocomputing*, vol. 658, Art. no. 131662, 2025, doi: 10.1016/j.neucom.2025.131662.

[27] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[28] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.

[29] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4084–4094.

[30] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5228–5237.

[31] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1911–1920.

[32] Z. Yang et al., "Band-kernel stochastic learning for unsupervised blind hyperspectral image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 7, 2026, doi: 10.1109/TPAMI.2026.3681688.

[33] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 314–3154.

[34] J. Tong, Y. Zou, Y. Li, and R. Li, "Lightweight frequency masker for cross-domain few-shot semantic segmentation," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 96728–96749, 2024.

[35] C. Ye, Y. Yao, W. Lin, X. Yang, and J. Qiu, "CSGrasp: Category-level semantically-aware grasping method," *IEEE Trans. Ind. Electron.*, vol. 73, no. 2, pp. 2610–2619, Feb. 2026, doi: 10.1109/TIE.2025.3605439.

[36] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1050–1065, Feb. 2020.

[37] B. Xue et al., "ISAR weak feature enhancement with perturbation defense using hybrid clustering oversegmentation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, no. 5, pp. 6256–6274, May 2024, doi: 10.1109/TAES.2024.3408139.

[38] M. Zhu, Y. Gong, D. Gu, and C. Tian, "Improving 3D object detection in neural radiance fields with channel attention," *CAAI Trans. Intell. Technol.*, vol. 10, no. 5, pp. 1446–1458, 2025, doi: 10.1049/cit2.70045.

[39] Y. Lin, Z. Zhu, Y. Guo, Z. Ou, S. Yao, and M. Song, "Exploring inter-domain Wasserstein metric for adaptive object detection," *IEEE Trans. Multimedia*, early access, Feb. 27, 2026, doi: 10.1109/TMM.2026.3668648.

[40] Q. Fan, W. Pei, Y.-W. Tai, and C.-K. Tang, "Self-support few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 701–719.

[41] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9587–9595.

[42] J. Lu, "Path-based knowledge graph link prediction method with graph context," *IEEE Trans. Computat. Social Syst.*, vol. 13, no. 3, pp. 3133–3146, Jun. 2026, doi: 10.1109/TCSS.2026.3669923.

[43] Y. Yin et al., "DAMNet: RGB-LiDAR fusion for traffic and water-surface detection using dynamic direction multi-scale scanning priors and deformable window cross-modal," *IEEE Sensors J.*, vol. 26, no. 11, pp. 17062–17073, Jun. 2026, doi: 10.1109/JSEN.2026.3684845.

[44] C. Wang et al., "Coupling damage characteristics of hull girder with variable cross-section subjected to underwater explosion," *Thin-Walled Structures*, vol. 217, 2025, Art. no. 113910, doi: 10.1016/j.tws.2025.113910.

[45] W. Gong, Y. Zhang, M. Zhu, T. Chen, and Z. Zheng, "Dynamic control of multiple stratospheric airships in time-varying wind fields for communication coverage missions," *Aerosp. Sci. Technol.*, vol. 166, 2025, Art. no. 110514, doi: 10.1016/j.ast.2025.110514.

[46] Z. Li, G. Li, X. Song, and X. Wang, "An efficient and dynamic framework for multi-scale target detection of underwater organisms," *J. Ocean Univ. China*, vol. 25, pp. 150–160, 2026, doi: 10.1007/s11802-026-6062-9.

[47] J. Nie et al., "Cross-domain few-shot segmentation via iterative support-query correspondence mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 3380–3390.

[48] Z. Wang et al., "Building façade delamination quantification framework based on infrared instance segmentation and dual-modality vision calibration," *Adv. Eng. Informat.*, vol. 71, 2026, Art. no. 104341, doi: 10.1016/j.aei.2026.104341.

[49] H. Chen, Y. Dong, Z. Lu, Y. Yu, and J. Han, "Pixel matching network for cross-domain few-shot segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 978–987.
[50] X. Huang, C. Zhu, and W. Chen, "RESTNet: Boosting cross-domain few-shot segmentation with residual transformation network," 2023, *arXiv:2308.13469*.
[51] M. Zhu, Z. Xu, Q. Zhang, Y. Liu, D. Gu, and S.-D. Xu, "GC-STormer: Gated swin transformer with channel weights for image denoising," *Expert Syst. With Appl.*, vol. 284, 2025, Art. no. 127924, doi: 10.1016/j.eswa.2025.127924.
[52] J. Zhuang, W. Chen, B. Guo, and Y. Yan, "Infrared weak target detection in dual images and dual areas," *Remote Sens.*, vol. 16, no. 19, 2024, Art. no. 3608, doi: 10.3390/rs16193608.
[53] X. Yan et al., "PKNet: Infrared small target detection via parallel interactive Kolmogorov-Arnold network," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5010114, doi: 10.1109/TGRS.2025.3635130.
[54] H. Zhang, Q. Zhong, J. Liu, R. Chen, and Z. Wang, "SAFARI-net: Scale-adaptive frequency-aware refinement infrastructure for infrared small target detection," *Opt. Laser Technol.*, vol. 199, 2026, Art. no. 115049, doi: 10.1016/j.optlastec.2026.115049.
[55] J. Herzog, "Adapt before comparison: A new perspective on cross-domain few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 23605–23615.
[56] J. Pfeiffer et al., "AdapterHub: A framework for adapting transformers," 2020, *arXiv:2007.07779*.
[57] S. Chen et al., "AdaptFormer: Adapting vision transformers for scalable visual recognition," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 16664–16678, 2022.
[58] N. A. Almujally et al., "Multi-modal remote perception learning for object sensory data," *Front. Neurorobot.*, vol. 18, 2024, Art. no. 1427786, doi: 10.3389/fnbot.2024.1427786.
[59] Y. Yang, Y. W. Ni, P. F. Chen, and X. S. Feng, "Predicting solar flares using a convolutional neural network with extreme-ultraviolet images," *Astrophysical J.*, vol. 985, no. 1, Art. no. 104, 2025, doi: 10.3847/1538-4357/adc88e.
[60] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-ViT: Unifying wavelet and transformers for visual representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 328–345.
[61] W. Dong et al., "Low-rank rescaled vision transformer fine-tuning: A residual design approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16101–16110.
[62] Z. Zhong, Z. Tang, T. He, H. Fang, and C. Yuan, "Convolution meets LoRA: Parameter efficient finetuning for segment anything model," 2024, *arXiv:2401.17868*.
[63] K. Gu et al., "Infrared image quality estimation with node-to-graph regression," *IEEE Trans. Multimedia*, vol. 28, pp. 2238–2252, 2026, doi: 10.1109/TMM.2026.3651047.
[64] H. Cao et al., "MFINet: A multi-scale feature interaction network for point cloud registration," *Vis. Comput.*, vol. 41, no. 6, pp. 4067–4079, 2025, doi: 10.1007/s00371-024-03646-2.
[65] Q. Su et al., "LeF-MTP: Prioritizing GNN test cases by fusing model uncertainty and feature-space confusability," *Comput. Intell.*, vol. 42, no. 1, 2026, Art. no. e70199, doi: 10.1111/coin.70199.
[66] P. Wan, H. Li, Z. Zhang, C. Duan, and J. Wang, "Motion parameter estimation of near-field drone based on DNSP under sensor position error," *IEEE Wireless Commun. Lett.*, vol. 15, pp. 1613–1617, 2026, doi: 10.1109/LWC.2026.3660298.
[67] W. Zhang, W. Fan, W. Su, and N. Bouguila, "HKANLP: Link prediction with hyperspherical embeddings and Kolmogorov–Arnold networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 37, no. 2, pp. 966–980, Feb. 2026, doi: 10.1109/TNNLS.2025.3614341.
[68] X. Liao, Q. Zhang, B. Sun, and X. Zheng, "Weather forecast-enabled tropospheric scattering path loss modeling for over-the-horizon maritime communications," *IEEE Trans. Antennas Propag.*, vol. 74, no. 2, pp. 1916–1928, Feb. 2026, doi: 10.1109/TAP.2025.3637508.
[69] M. T. Chiu et al., "Agriculture-vision: A large aerial image database for agricultural pattern analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2828–2838.
[70] H. T. Wei et al., "TGCM: Cross-domain few-shot semantic segmentation via one-shot target guided CutMix," in *Proc. Asian Conf. Comput. Vis.*, 2024, pp. 1065–1081.
[71] Q. Fan et al., "Adapting in-domain few-shot segmentation to new domains without source domain retraining," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 21429–21439.
[72] A. Wei et al., "FSINet: A robust feature separation and integration network for multiscale SAR object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 8224–8238, 2026, doi: 10.1109/JSTARS.2026.3665718.
[73] H. Sheng et al., "Discriminative feature learning with co-occurrence attention network for vehicle ReID," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3510–3522, May 2024, doi: 10.1109/TCSVT.2023.3326375.
[74] Z. Li et al., "Dual-agent optimization framework for cross-domain few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 9849–9859.