

Article

Multi-Level Cross-Modal Interactive-Network-Based Semi-Supervised Multi-Modal Ship Classification

Xin Song , Zhikui Chen *, Fangming Zhong *, Jing Gao, Jianing Zhang and Peng Li

The School of Software Technology, Dalian University of Technology, Dalian 116621, China; songxin72@mail.dlut.edu.cn (X.S.)

* Correspondence: zkchen@dlut.edu.cn (Z.C.); fmzhong@dlut.edu.cn (F.Z.)

Abstract: Ship image classification identifies the type of ships in an input image, which plays a significant role in the marine field. To enhance the ship classification performance, various research focuses on studying multi-modal ship classification, which aims at combining the advantages of visible images and infrared images to capture complementary information. However, the current methods simply concatenate features of different modalities to learn complementary information, which neglects the multi-level correlation between different modalities. Moreover, the existing methods require a large amount of labeled ship images to train the model. How to capture the multi-level cross-modal correlation between unlabeled and labeled data is still a challenge. In this paper, a novel semi-supervised multi-modal ship classification approach is proposed to solve these issues, which consists of two components, i.e., multi-level cross-modal interactive network and semi-supervised contrastive learning strategy. To learn comprehensive complementary information for classification, the multi-level cross-modal interactive network is designed to build local-level and global-level cross-modal feature correlation. Then, the semi-supervised contrastive learning strategy is employed to drive the optimization of the network with the intra-class consistency constraint based on supervision signals of unlabeled samples and prior label information. Extensive experiments on the public datasets demonstrate that our approach achieves state-of-the-art semi-supervised classification effectiveness.

Keywords: ship classification; deep multi-modal learning; semi-supervised learning



Citation: Song, X.; Chen, Z.; Zhong, F.; Gao, J.; Zhang, J.; Li, P. Multi-Level Cross-Modal Interactive-Network-Based Semi-Supervised Multi-Modal Ship Classification. *Sensors* **2024**, *24*, 7298. <https://doi.org/10.3390/s24227298>

Academic Editor: Jingjing Wang

Received: 4 September 2024

Revised: 6 November 2024

Accepted: 12 November 2024

Published: 15 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Ship classification has aroused increasing attention and played a significant role in the marine field, such as combating illegal fishing and managing maritime traffic. With the development of Earth observation technology, ship images collected by the maritime surveillance system have surged, which makes ship classification more attractive for relevant applications. At present, a great deal of research focuses on ship image classification, which utilizes the characteristics extracted from the input image to recognize the ship category [1–3].

Mainstream ship images for classification contain synthetic aperture radar (SAR) images, visible images, and infrared images [4–6]. SAR images are obtained from radars and are vulnerable to other electromagnetic interference, which can only distinguish large vessels from the background [7–9]. Compared with SAR images, visible images have higher resolution and provide more texture details for classifying fine-grained ship targets and have attracted great attention from researchers [10]. Visible images are generated by reflected light and can capture abundant visual details, while they cannot capture explicit ship target information under poor lighting conditions. Infrared images can ameliorate the problem by providing complementary ship target information for visible images of the same scene. The multi-modal ship image (visible and infrared image) classification combines the advantages of visible images and infrared images to mine the complementary

information for classification, which has demonstrated significant advantages over single-modal images.

Traditional multi-modal methods manually incorporate hand-crafted features of different modalities into common representations to learn complementary information and then feed common representations to a classifier like SVM for classification [11]. It is noted that these methods depend heavily on image quality and show poor classification performance in complex imaging conditions, where illumination or viewing angles change significantly. To solve this problem, deep-learning methods are proposed to automatically extract deep features of different modalities through convolutional neural networks and then concatenate these features to capture complementary information. Then, like traditional methods, deep-learning methods feed concatenated deep common representations into a classifier. For example, Ren et al. [10] concatenates high-level features extracted by CNN and spatial location features extracted by Hu invariant moments [12] as common representations for ship image classification.

Despite the success of multi-modal ship classification, the current network-based methods usually simply concatenate features of different modalities, which ignores the correlation between different modalities. However, capturing the correlation between different modalities from texture and semantic perspectives is critical for learning comprehensive complementary information about ship images. Furthermore, all the above methods have the common drawback that the training progress requires a large amount of labeled ship images [13–15]. The existing methods neglect that the number of labeled ship images is usually limited and cannot utilize unlabeled data effectively, which can pose a great difficulty for model training. For the current semi-supervised multi-modal ship classification methods, how to make full use of the cross-modal correlation implied in unlabeled and labeled data has not been well studied.

In this paper, a novel semi-supervised multi-modal ship classification approach (Semi-MMSC) is proposed (see Figure 1) that utilizes the multi-level cross-modal interactive network to learn comprehensive complementary information and introduces a novel semi-supervised contrastive learning strategy to guide the optimization of the network. The multi-level cross-modal interactive network mines the correlation of different modalities in both local-level and global-level branches. In the local-level branch, to model the relationship between different modalities from the texture perspective, Semi-MMSC first utilizes local feature extractors to obtain texture features of different modalities. Then, the interactive network employs the local-level fusion module to fuse these texture features with the spatial and channel attention mechanism. In the global-level branch, Semi-MMSC extracts global features of different modalities first. After that, to capture the relationship between different modalities from the semantic perspective, the network builds the feature interactive map among global-level features with local-level fused features as a medium via the global-level interactive module.

To alleviate the influence of label scarcity on model training, the semi-supervised contrastive learning strategy incorporates contrastive loss into the semi-supervised learning and optimizes the network with these two loss functions. Specifically, with the unsupervised contrastive loss, the contrastive representations are encoded by maximizing the similarity between correlated features of the same sample, which can learn similarity distribution from data and be used as supervised signals of unlabeled data. To train the network with obtained supervision signals of unlabeled data, the memory mechanism is designed to generate class prototypes based on contrastive representations. Then, the self-supervised loss function is designed to constrain the intra-class consistency between unlabeled and labeled multi-modal data based on obtained supervision signals and prior label information. The experimental results on the public datasets confirm that the proposed approach outperforms the state-of-the-art semi-supervised models in terms of classification performance.

The main contributions of this paper can be summarized as follows:

1. The multi-level cross-modal interactive network is proposed to implement local-level and global-level cross-modal feature interaction, which encodes deep correlated features containing more comprehensive complementary information for classification.
2. The semi-supervised contrastive learning strategy is designed to derive supervised signals from unlabeled samples with contrastive learning and learn the intra-class consistency constraint between unlabeled and labeled data to optimize the model.
3. The performance comparison shows the enhanced classification performance of the proposed approach over the current state-of-the-art on the public datasets.

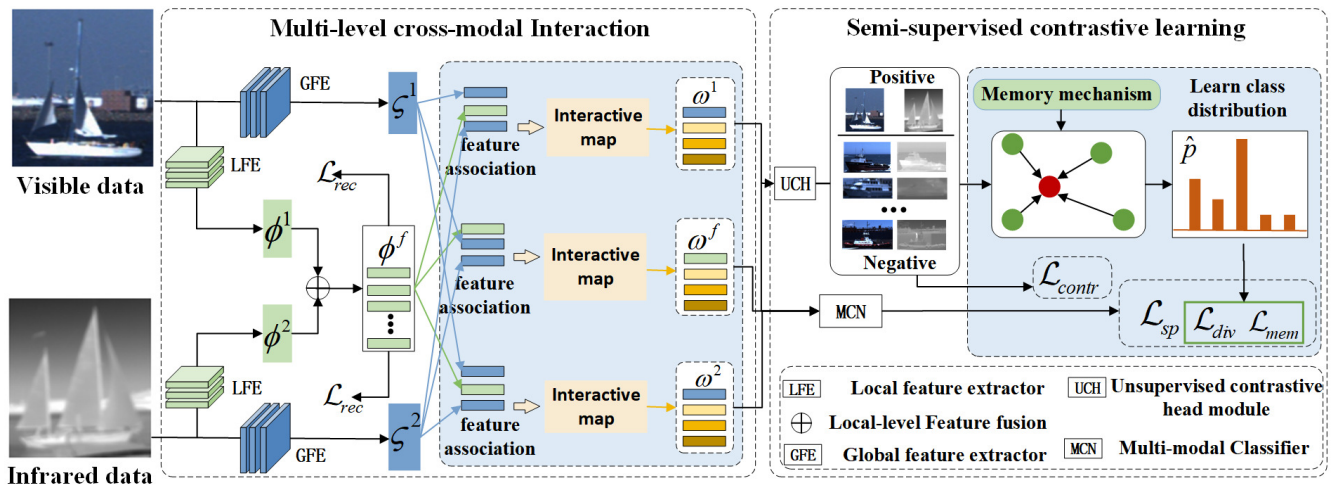


Figure 1. Framework of the proposed approach. Given a batch of multi-modal images, the approach utilizes the multi-level cross-modal interactive network to encode deep correlated features ω^i of multi-modal data through mining the correlation among global-level features ζ^i for each modal data and local-level fused features ϕ^f . ϕ^f are based on the fusion between texture features ϕ^i of each modal datum through the spatial and channel attention mechanism. For the optimization of the network, the self-supervised contrastive learning strategy achieves the intra-class consistency constraint based on class distribution ω^f of unlabeled data generated through the memory mechanism and prior label information, which contains unsupervised contrastive loss $\mathcal{L}_{contr}$ and semi-supervised divergence loss ($\mathcal{L}_{div}, \mathcal{L}_{mem}$). In addition, A reconstructive loss $\mathcal{L}_{rec}$ is used to guide the model pre-training. A supervised loss $\mathcal{L}_{sp}$ is used to learn label information. Contrastive representations ω^i are outputs of the unsupervised contrastive head module, which are the input of the memory mechanism.

2. Related Works

2.1. Ship Classification

Ship classification methods extract visual characteristics from input sensors and then utilize these characteristics to recognize which categories the input belongs to, which play a crucial role in many marine tasks, such as vessel management and navigation safety protection.

Single-modal Ship Classification: The traditional methods attempt to learn visual characteristics only based on the single image source, including SAR data, infrared data, and visible data. SAR is one kind of active microwave imaging sensor that can collect all-day and all-weather data [16,17]. Wang et al. [18] designed a hierarchical ship classifier for COSMO-SkyMed SAR images where the geometric and backscattering characteristics of various ship types were analyzed and used. Wu et al. [19] proposed a ship classification method based on high-resolution SAR images to distinguish bulk carriers, container ships, and oil tankers. The method first designed one three-element feature vector that was composed of kernel density mean features, three structural features, and a backscattering coefficient estimation mean feature, and then, this feature vector was input into one SVM classifier for the final ship classification. Li et al. [2] proposed a clustering-based size-adaptive safer oversampling technique for SAR ship classification that selects minority

samples adaptively and then generates new instances with the safe metric to alleviate the imbalanced dataset problem. Generally, SAR images are used for distinguishing differences between large-size ships due to the limitation of image resolution.

On the contrary, visible images and infrared images can provide richer visual details and explicit texture information for classification that attract a large amount of attention. Visible images are generated by the visible sensor capturing reflected light and contain rich texture information while being sensitive to illumination changes [20–22]. Infrared images are generated by the infrared sensor capturing thermal radiations and contain rich target information by presenting the different distributions of temperature in the scene [23–25]. However, infrared images contain little texture information due to the problem of low resolution and blurred edges. To complement the deficiency of each modal datum, various researchers have attempted to study multi-modal ship classification, which captures complementary information hidden in multi-modal data for ship classification.

Multi-modal Ship Classification: Recently, multi-modal ship classification methods have been proposed that extract generic features from different modalities and combine these features to learn cross-modal complementary information hidden in the data. Traditional multi-modal methods utilize hand-crafted feature extractions to extract features from the data. The common feature extractions include the histogram of oriented gradients (HOG) [26], scale-invariant feature transform (SIFT) [27], local binary patterns (LBP) [28], etc. Then, a classification model like support vector machine (SVM) is used to classify images based on the obtained hand-crafted features. For example, because HOG features have better representation ability than other hand-crafted features, Lin et al. [28] proposed to utilize the manifold learning SAR-HOG features to classify ship objects. However, these hand-crafted features can only learn low-level visual characteristics, and the extraction process consumes a large amount of human effort. More importantly, traditional multi-modal methods have high requirements for image quality, which are impractical for ship classification since the actual data collection environment is complex.

To solve the above problems, many deep-learning methods are proposed to adopt deep convolutional networks as the base architecture to extract deep features automatically, which can learn high-level features from images. These high-level features are more invariant than low-level visual features, which are important for distinguishing the differences in different classes. For example, Zhang et al. [1] proposed to combine HOG features and deep CNN features with the hog feature fusion mechanism to improve the classification performance. To both preserve the internal structure of the category information and select the minimal dimension features, the feature fusion architecture based on spectral regression discriminant analysis is built to combine features from infrared and visible images [29]. Ren et al. [10] proposed a hybrid method that fuses high-level features extracted by CNN and spatial location features extracted by Hu invariant moments for ship image classification. Although the current multi-modal ship classification methods have achieved significant success, the application of deep learning methods is still limited due to the expensive cost of massive labeled data collection. To solve this problem, some works attempt to introduce semi-supervised learning into the maritime field [30–32].

2.2. Semi-Supervised Learning

Semi-supervised learning focuses on utilizing a few labeled data and a large amount of unlabeled data to implement the classification task, which has been a long-standing problem in computer vision. The recent self-supervised methods consist of two branches: pseudo-label-based methods and consistency-regularization-based methods. The former focuses on producing pseudo labels based on a pre-trained model for unlabeled samples and then training the model with these pseudo labels [33,34]. The whole training process is viewed as a self-training process, which suffers the overfitting problem caused by incorrect pseudo labels. Iscen et al. [34] made a detour by the label propagation to spread labeled data distribution on unlabeled data. Consistency-regularization-based methods aim to produce consistent predictions on different image views of the same image and then adopt the

pseudo-label-based consistency learning strategy to train the model [35–37]. The previous methods created different views of an image by strong augmentations to simulate image perturbations of an object. At present, pseudo-label-based consistency training, which uses a weak augmentation for creating pseudo labels and a strong augmentation for consistency training, has dominated and achieved excellent performance gain.

In the ship classification task, some methods try to transfer learned prior class knowledge into the training of unlabeled ship images to achieve better classification results. For example, Yang et al. [30] transfers automatic identification system (AIS) knowledge to train the network on the image domain, which extracts static ship information from AIS data to assist in learning ship category information. However, there is no literature for semi-supervised multi-modal ship classification, which aims at dealing with the classification problem of unlabeled multi-modal ship data. Hence, we propose a novel approach, which can make use of cross-modal complementary information from unlabeled data and limited labeled data for multi-modal ship classification.

3. The Proposed Approach

In this section, details of the proposed approach are introduced, including the preliminaries, the network architecture, and loss functions.

3.1. Preliminaries

The multi-modal ship dataset is defined as $D = \{X^1, \dots, X^M\}$, $M = 2$, where M denotes the number of the modalities. $X^m = \{x_i^m \in R^{C \times H \times W}\}_{i=1}^N$ denotes m -th modal data. N denotes the number of datasets, C denotes the image channel, H and W , respectively, denote the image height and width. Two ResNets are adopted as the global feature extractor to extract deep global-level features $\zeta^m \in R^{N_g}$ of different modalities, $\zeta^m = F^m(X^m; W^m)$, $m \in 1, \dots, M$, where F^m represents the m -th convolutional network and W^m denotes the corresponding network weights, respectively. N_g denotes the global-level feature dimension.

3.2. Multi-Level Cross-Modal Interactive Network

Concatenating deep global-level features of these two modalities linearly cannot learn comprehensive complementary information hidden in unlabeled data. To alleviate this problem, the multi-level cross-modal interactive network is proposed to mine the feature correlation of different modalities in the local-level and global-level branches.

3.2.1. Local-Level Fusion Module

In the local branch, to build the correlation of texture features, the network first employs the local feature extractor to extract texture features from images of different modal data and then combines these features with the spatial and channel attention mechanism. The local-level fusion module consists of two stages, extraction and fusion. To learn the texture information hidden in data with the local-level fusion module, the reconstruct loss is designed to train the module, which is demonstrated in the following sections.

Extraction: For the input image of each model, the local-level fusion module adopts the local feature extractor to extract texture features of i -th modal data $\phi_l^i \in R^{C_l \times H_l \times W_l}$. C_l denotes the number of texture feature channels at the l -th layer. H_l and W_l denote the number of the texture feature height and width at the l -th layer. The local feature extractor consists of four convolutional blocks. Each convolutional block contains max-pool and stride convolution. These operations are used as downsample operations to maintain texture information in the feature extraction process. The whole network architecture is shown in Table 1.

Fusion: To obtain local-level fused features based on texture features of different modalities, the local-level fusion module performs the spatial and channel attention mechanism on texture features to generate the spatial and channel attention maps, respectively, and then fuses two attention maps. Specifically, the spatial attention mechanism calculates

the spatial attention map by the weighted sum of the features of different modes along the spatial dimension $\phi_l^p \in R^{C_l \times H_l \times W_l}$

$$\phi_l^p = \sum_{j=1}^M \frac{\|\phi_l^j(a, b)\|_1}{\sum_{i=1}^M \|\phi_l^i(a, b)\|_1} \times \phi_l^j(a, b) \quad (1)$$

where $\|\cdot\|_1$ denotes l_1 -norm, M denotes the number of the modalities. $\phi_l^i(a, b)$ indicates features of the i -th modal at the pixel position (a, b) . Similarly, the local-level fusion module utilizes the channel attention mechanism to calculate the channel attention map by the weighted sum of features of different modalities along the channel dimension $\phi_l^c \in R^{C_l \times H_l \times W_l}$

$$\phi_l^c = \sum_{j=1}^M \frac{P(\phi_l^j(c))}{\sum_{i=1}^M P(\phi_l^i(c))} \times \phi_l^j(c) \quad (2)$$

where P denotes the global pooling operator, $\phi_l^i(c)$ indicates features of the i -th modal at the channel c . Local-level fused features are obtained by combining the spatial attention map and the channel attention map $\phi_l^f = (\phi_l^p + \phi_l^c) \times \frac{1}{2}$.

Table 1. The architecture of the local-level fusion module.

Name	Type	Output Size
ConvEn_1	Conv + Relu	(16, 224, 224)
ConvEn_2	Conv + Relu	(8, 224, 224)
	Conv + Relu	(64, 224, 224)
	MaxPooling	(64, 112, 112)
convEn_3	Conv + Relu	(32, 112, 112)
	Conv + Relu	(112, 112, 112)
	MaxPooling	(112, 56, 56)
convEn_4	Conv + Relu	(56, 56, 56)
	Conv + Relu	(160, 56, 56)
	MaxPooling	(80, 28, 28)
convEn_5	Conv + Relu	(40, 28, 28)
	Conv + Relu	(208, 28, 28)

3.2.2. Global-Level Interactive Module

As shown in Figure 2, in the global-level branch, the global-level interactive module aims at building the feature interactive map between global-level features of visible and infrared data with local-level fused features as a medium, which encodes correlated features that contain deep complementary information. First, we define the input feature list of the global-level interactive module as

$$\begin{aligned} \text{InputFeatureList} &= [\zeta^1, \dots, \zeta^M, \zeta^{M+1}], \\ \text{where } \zeta^{M+1} &= F^{m+1}(\phi_1^f; W^{M+1}) \in R^{N_g} \end{aligned} \quad (3)$$

Although there are four layers of local-level fused features obtained by the local-level fusion module, this paper only adopts features of the first layer as the input feature because features ϕ_1^f can contain more comprehensive texture information than other layers. For each pair of features, the pairwise interactive matrix is first computed by performing matrix multiplication on two feature vectors

$$\gamma(i, \bar{i}) = \zeta^i \zeta^{\bar{i}T}, \text{ where } i \in [1, M+1] \quad (4)$$

where $\bar{i} = \{1, \dots, M+1\} \setminus i$. Then, the feature interactive map of the i -th input with respect to other feature vectors is defined as follows:

$$C^i = \sum_{(i,\bar{i}) \in S^i} \gamma(i, \bar{i}) \quad (5)$$

where C^i denotes the vectorization of the matrix, and $S^i = \{(\zeta^i, \zeta^{\bar{i}})\}_{\bar{i}=\{1,\dots,M+1\}\setminus i}$ denotes the sets of feature pairs. The feature interactive map builds the correlations among global-level features of different modalities and local-level fused features, and models the inter-modal relationship from the semantic perspective.

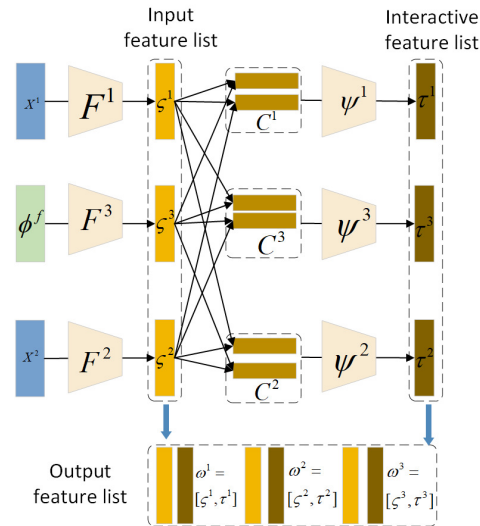


Figure 2. Framework of the proposed global-level interactive module.

The above feature interactive map needs to calculate the feature vectors of all pairs, which may cause feature redundancy. The global-level interactive module further reduces the feature dimension through a multi-layer fully connected network ψ^i

$$\tau^i = \psi^i(C^i; W^\psi) \quad (6)$$

where $W^\psi \in R^{N_g \times N_I}$ denotes the network weights. Then, the deep correlated features are generated by combining interactive features and input features of the global-level interactive module $\omega^i = [\zeta^i, \tau^i] \in R^{N_g + N_I}$.

3.3. Semi-Supervised Contrastive Learning Strategy

After obtaining deep correlated features of the unlabeled data that can provide rich complementary information, the approach needs to further alleviate the influence of label scarcity. Toward this goal, the semi-supervised contrastive correlation learning strategy is proposed, which comprises two major components: unsupervised contrastive loss and class-wise memory loss.

3.3.1. Unsupervised Contrastive Loss

Contrastive learning utilizes different correlated features from the same image as positive samples and other samples from the training batch as negative samples, which enforces the similarity of correlated features between different modalities and produces contrastive representations to learn complementary information from unlabeled data. Specifically, for correlated features of the i -th sample belonging to two modalities $(\omega_i^1, \omega_i^2, \omega_i^3)$, the contrastive head module first multiplies the correlated features by element to produce contrastive representations, which learn consensus information of the correlated features between different modalities. The contrastive representations of a sample contain three groups, $\omega_i^1 = \|\omega_i^2\| \|\omega_i^3\|$, $\omega_i^2 = \|\omega_i^1\| \|\omega_i^3\|$, $\omega_i^3 = \|\omega_i^1\| \|\omega_i^2\|$. Then, the contrastive

head module measures the similarity of positive samples, $\text{sim}(\omega_i^1) = \frac{(\omega_i^2)^T \omega_i^3}{\omega_i^1}$, $\text{sim}(\omega_i^2) = \frac{(\omega_i^1)^T \omega_i^3}{\omega_i^2}$, $\text{sim}(\omega_i^3) = \frac{(\omega_i^1)^T \omega_i^2}{\omega_i^3}$.

To encourage contrastive representations can learn model consensus information from correlated features, the corresponding unsupervised contrastive loss for all positive pairs of the i -th sample is formulated as

$$\text{Contr}_i = -\frac{1}{M+1} \sum_{m=1}^{M+1} \log \frac{\exp(\text{sim}(\omega_i^m)) / \nu}{\sum_{k=1}^N (\exp(\text{sim_neg}(\omega_i^m))) / \nu} \quad (7)$$

where ν denotes the temperature parameter. The negative samples are denoted as $\text{sim_neg}(\omega_i^1) = \frac{(\omega_i^2)^T \omega_k^3}{\|\omega_i^2\| \|\omega_k^3\|}$, $\text{sim_neg}(\omega_i^2) = \frac{(\omega_i^1)^T \omega_k^3}{\|\omega_i^1\| \|\omega_k^3\|}$, $\text{sim_neg}(\omega_i^3) = \frac{(\omega_i^1)^T \omega_k^2}{\|\omega_i^1\| \|\omega_k^2\|}$. By minimizing $\mathcal{L}_{\text{contr}} = \sum_{i=1}^N \text{Contr}_i$, correlated features of the same sample become more similar, and correlated features of the different samples become more dissimilar. In consequence, contrastive representations of similar samples have large values, and contrastive representations of dissimilar samples have small values.

For reducing computational complexity, this paper only adopts $\omega^f = \omega_i^2$ instead of all pairs of contrastive representations to generate class prototypes of the i -th sample. Specifically, the memory mechanism takes the key-value structure as the backbone, which learns class-wise prototypical features from contrastive representations of the whole dataset and predicts class probability for unlabeled data. Key embedding stores the prototypical feature of each class and the value embedding records the probabilistic prediction of the same class.

First, the memory mechanism utilizes contrastive representations of labeled data to update the key and value embedding.

$$\begin{aligned} k_j &= k_j - \eta \frac{\sum_{i=1}^{N_j} (k_j - \omega_i^f)}{1 + N_j}, \\ v_j &= \frac{v_j - \eta \Delta_j}{\sum_{i=1}^K v_{j,i} - \eta \Delta_{j,i}}, \text{ where } \Delta_j = \frac{\sum_{i=1}^{N_j} v_j - p_i}{1 + N_j} \end{aligned} \quad (8)$$

where η denotes the learning rate, K denotes the number of classes, and N_j denotes the number of samples from the j -th class ($j \in \{1, \dots, K\}$). p_i indicates the network prediction of the sample i . Along with the training progress, the key and value embeddings become more reliable to reflect the class distribution. At the beginning, the key and value are initialized to 0 and $\frac{1}{K}$, respectively.

With these key and value embeddings, class-wise prototypical features of the labeled data are preserved and the class-wise memory prediction of a sample x_i is calculated $\hat{p}_i = \sum_{i=1}^K w(k_i, v_i | I) \omega_i^f$, where $w(k_i, v_i | x_i)$ denotes the class probability. Specifically, for labeled samples, the class probability is based on the label of the sample. For unlabeled samples, the class probability is based on the pairwise similarity between the contrastive representations ω^f and the key embedding k_i

$$w(k_i, v_i | x_i) = \frac{e^{-\text{dist}(\omega^f, k_i)}}{\sum_{j=1}^K e^{-\text{dist}(\omega^f, k_j)}} \quad (9)$$

3.3.2. Class-Wise Memory Loss

Based on class-wise memory predictions of the unsupervised data, the semi-supervised loss function is formulated for the optimization of training, which ensures the class-wise consistency between unlabeled and labeled data. The loss function contains two parts, the class divergence loss and the memory entropy loss. The class divergence loss encourages the consistency of class information between the memory module and the network while

the memory entropy loss is to encourage the memory module to learn discriminative class information.

The class divergence loss aims to expand class information hidden in the memory module into the training of the unlabeled data.

$$\mathcal{L}_{div} = \sum_{j=1}^K p(j) \log \frac{p(j)}{\hat{p}(j)} \quad (10)$$

The class divergence loss penalizes the discrepancy between two distributions, $p(j)$ and $\hat{p}(j)$. When $\mathcal{L}_{div} \rightarrow 0$, it indicates the network predictions based on correlated features match well with memory predictions.

To penalize inconsistent predictions, the memory entropy loss is designed, which is motivated by the entropy minimization principle,

$$\mathcal{L}_{mem} = - \sum_{j=1}^K \hat{p}(j) \log \hat{p}(j) \quad (11)$$

The memory entropy loss reflects the overall memory inference uncertainty and penalizes uncertain memory predictions. A high value of the loss indicates that the memory cannot easily distinguish classes of ship images, which results from assigning inconsistent probabilistic predictions to correlated features within the same class.

3.4. Model Training

In this subsection, the whole progress of the network training is introduced. Before training the whole network, the local-level fusion module needs to be pre-trained, which enforces that the texture information hidden in the data can be captured by local-level fused features. Specially, the pre-trained reconstruct loss is defined as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{pix} + \mathcal{L}_{str} \quad (12)$$

where $\mathcal{L}_{pix}$ and $\mathcal{L}_{str}$ indicate the pixel loss and structure loss, respectively. The pixel loss calculates the distance between input images and the reconstructed images X^{rec} , which makes sure the reconstructed images are more similar to the input images

$$\mathcal{L}_{pix} = \frac{1}{2} \times (\|X^{rec} - X^1\|_F^2 + \|X^{rec} - X^2\|_F^2) \quad (13)$$

where F denotes the Frobenius norm. The structure loss calculates the structural similarity between input images and the reconstructed images

$$\mathcal{L}_{str} = \frac{1}{2} \times (1 - \text{SSIM}(X^{rec}, X^1) + 1 - \text{SSIM}(X^{rec}, X^2)) \quad (14)$$

where SSIM denotes the structural similarity measure. With the value becoming larger, the input images and the reconstructed images have more similarity in structure.

For the labeled data, the supervised training loss is designed to calculate the cross entropy between predictions and labels, which enforces the network to learn class information from the labeled data. The detailed loss function is as follows:

$$\mathcal{L}_{sp} = - \frac{1}{M+1} \sum_{i=1}^{(M+1)} \sum_{j=1}^K \mathbb{1}[y = k] \log(p(y_j | H(\omega_i))) \quad (15)$$

where H represents the multi-modal classifier with multiple full-connection layers. For the unlabeled data, the training loss is based on the semi-supervised contrastive learning strategy, which is demonstrated in the above section. Finally, the overall loss is as follows:

$$\mathcal{L} = \mathcal{L}_{sp} + \alpha \mathcal{L}_{contr} + \beta (\mathcal{L}_{div} + \mathcal{L}_{mem}) \quad (16)$$

where α and β denote loss weights of unsupervised contrastive learning loss and class-wise memory loss, respectively.

4. Experiments

4.1. Datasets

The proposed approach is evaluated on the VAIS dataset [38], which is the first publicly multi-modal maritime ship dataset. For now, as we know, it is also the only existing public multi-modal ship image dataset. The dataset comprises 2865 images (1623 visible ship images and 1242 infrared ship images) and includes six ship categories: cargo ships, middle-other ships, passenger ships, sailing ships, small boats, and tugboats. These images are collected at various times in one day, including dusk and dawn. Some images in the dataset are high-resolution while some images are blurred and hard to recognize. According to the paper [38], there are 1088 pairs of matched visible–infrared unregistered images. Other images are unpaired and low-resolution, including 154 night-time infrared images.

To evaluate the generalization ability of our approach, we use the Sentinel-1 dataset to test the model. The Sentinel-1 dataset [39] is collected from Sentinel-1 satellite single-looked complex (SLC) SAR images of imaging mode Stripmap. The Sentinel-1 SAR mission was launched by the European Space Agency (ESA), operating in C-Band with horizontal and vertical polarizations, providing SAR data in four different modes.

4.2. Comparison Methods

The comparison methods include the supervised learning baseline, a supervised ship classification method P2Net [40], pseudo-label-based baseline, and several state-of-the-art semi-supervised learning methods, including FixMatch [41], CoMatch [42], and Semcls [43]. Fixmatch [41] is a self-training method based on pseudo label generation technology that selects unsupervised samples with high-confidence pseudo labels and uses them as training targets. CoMatch [42] is a graph-based semi-supervised learning method based on the self-supervised learning framework, that generates pseudo labels based on image embeddings and learns better representations by graph-based contrastive learning. Semcls [43] is a semi-supervised contrastive learning method, that designs the class-aware contrastive module to explore reliable in-distribution data clustering and noisy out-of-distribution data contrasting. All the compared methods are trained under single-modal and multi-modal settings. SNL-DNLL [44] is a mutual learning framework based on pseudo-negative labels. In the single-modal setting, all methods take visible and infrared images as input, respectively. In the multi-modal setting, all methods utilize visible and infrared fused images as the input. For producing fused images, an image fusion model, nest-connection-based image fusion framework (NestFuse) [45], is used.

To verify the effectiveness of Semi-MMSC on the Sentinel-1 dataset, three methods (CNN, TL-CNN, and DSN), used in the paper [46], are selected as comparison methods. DSN can be viewed as the multi-modal classification method and combines spatial features and frequency features to classify targets.

4.3. Evaluation Indicators

For the evaluation of the proposed approach, four indicators are used: accuracy, precision, recall, and f1-score.

4.4. Implement Details

The network consists of two components, the local-level fusion module and the global-level interactive module. The local-level fusion module is based on the encoder–decoder architecture, which includes three parts: the local feature extractor, the local feature decoder, and the fusion layer. The local feature extractor is a convolutional network with one

convolutional layer and four convolutional blocks. Each block contains two convolutional layers and one max-pooling operator. The convolutional layers all adopt 3×3 kernel size and are equipped with the Relu activation function. The local feature decoder includes one convolutional layer and six convolutional blocks, which are connected by the nest connection architecture. The convolutional blocks of the decoder are made up of three convolution layers. During training, with the modal-specific encoder and modal-specific decoder with pre-trained loss, deep salient features of the two modalities are extracted, respectively. The fusion layer is built based on spatial and channel attention mechanisms, which are used to fuse these salient features of different modalities.

For mining the global-level correlation further, the global-level interactive module is built. The module consists of three parts: the global feature extractor, feature interactive module, and multi-modal classifier. The global feature extractor is based on ResNet architecture, which is used to extract global-level features. Then, these features are fed into the feature interactive module, which is equipped with three fully connected layers. Three fully connected layers include 400, 200, and 600 units, respectively, followed by a Batchnorm regularization and a Relu activation function. The input of the third fully connected layer is the interactive calculation results of the output of the second fully connected layer. Furthermore, the multi-modal classifier is built to predict the class probability of data points, which is equipped with two fully connected layers.

We adopt the Adam optimizer for optimization, and the learning rate is set to 0.0001. We train Semi-MMSC for 100 epochs to demonstrate its efficiency on the classification task and the batch size is set to 8. The time taken to complete one epoch of training is 110 s, and the duration for a single ship image inference is 0.6 s.

4.5. Dataset Processing

Because the sample number of each class on the VAIS dataset is imbalanced in practice, randomly sampled images with labels like standard semi-supervised learning methods are not class-balanced. For example, the tug boat class only has 50 samples. Hence, for conducting a thorough analysis of the proposed approach, the experiment under two kinds of label fraction settings is conducted in a class-balanced way, 10 samples of each class are used in the training, and a class-imbalanced way, 10% samples of each class are used. The experiment contains two parts, single-modal classification and multi-modal classification. Specifically, in single-modal learning, the experiment takes images of two modalities separately as inputs, which are represented by ‘vis’ and ‘inf’, respectively. The classification results of all compared methods on the VAIS dataset are shown in Table 2.

Table 2. Accuracy for the VAIS dataset on two different label fraction settings.

Methods		Top-1 Accuracy Label Fraction (%)		Top-5 Accuracy Label Fraction (%)	
		10 Samples per Class	10% Samples per Class	10 Samples per Class	10% Samples per Class
single-modal classification	supervised baseline (vis)	32.4449	26.1029	95.4044	89.2463
	supervised baseline (inf)	40.5331	56.6176	88.0515	97.0588
	pseudo label baseline (vis)	32.4449	32.4449	73.8971	95.4044
	pseudo label baseline (inf)	32.4449	32.4449	86.5809	86.5809
	Fixmatch [41] (vis)	29.3199	26.1029	67.5551	95.4044
	Fixmatch [41] (inf)	32.7206	34.5588	95.4044	70.6801
	Comatch [42] (vis)	46.2316	49.5404	98.6213	94.2096
	Comatch [42] (inf)	48.0699	35.2941	95.4044	89.2463
	Semcls [43] (vis)	32.4449	57.0772	73.8971	99.6324
	Semcls [43] (inf)	32.4449	32.4449	89.2463	87.3162
	SNL-DNLL [44] (vis)	12.6838	26.1029	67.5551	86.5809

Table 2. Cont.

	Methods	Top-1 Accuracy Label Fraction (%)		Top-5 Accuracy Label Fraction (%)	
		10 Samples per Class	10% Samples per Class	10 Samples per Class	10% Samples per Class
single-modal classification	SNL-DNLL [44] (inf)	10.7537	32.3420	73.8971	86.5809
	P2Net [40] (vis)	32.4449	26.1029	95.4044	89.2463
	P2Net [40] (inf)	40.5331	56.6176	88.0515	97.0588
multi-modal classification	supervised fusion baseline	34.3750	54.5037	89.9816	93.6581
	pseudo label fusion baseline	39.6140	44.3934	77.7574	94.7610
	Fixmatch [41]	34.8346	32.4449	95.4044	87.3162
	Comatch [42]	32.4449	33.1801	95.4044	95.4044
	Semcls [43]	32.4449	32.4449	95.4044	73.8971
	Semi-MMSC	63.6949	72.0588	100.00	100.00

The experimental visible and infrared images are preprocessed with the unified data augmentation technology, including image scaling, image color jitter, and the direction adjustment of the ship image via horizontal transform. The experimental SAR images are preprocessed using the method in [40]. All experiments are performed on Intel(R) Xeon(R) Gold 5115 CPU @ 2.40GHz (Intel, Santa Clara, CA, USA) and NVIDIA Tesla V100 (NVIDIA, Santa Clara, CA, USA) and are implemented in Python 3.7 using the PyTorch framework.

4.6. Experiment Results

It is observed that deep semi-supervised learning methods based on pseudo labels do not show better performance than the supervised learning baseline, which proves that simply generating pseudo labels for the unlabeled data is not effective. Specifically, many methods achieve the same identical accuracy value as supervised learning, like 32.44% on the first label fraction setting. The reason for this problem should be that a large number of unlabeled samples can confuse the model and cause model overfitting so the model can only classify a few supervised samples. Semi-supervised learning methods show better performance than the supervised baseline and pseudo-label baseline on the VAIS dataset. It indicates that making use of unlabeled data and limited data appropriately is effective in improving the semi-supervised classification performance.

By comparing results under the class-balanced (10 samples of each class) and class-unbalanced (10% samples of each class) conditions, it is noted that a balanced sample size for each category is more conducive to the model learning class information, since most methods show better performance under the class-balanced condition.

Furthermore, some multi-model classification methods show worse performance than single-model classification methods, which indicates that simply concatenating features of different modalities can not learn complementary ship hull and superstructure information effectively, and redundant category independent information from different modalities causes the model performance degradation.

Comparing existing methods, the proposed approach (Semi-MMSC) obtains a top-1 accuracy of 63.69% with 10 samples per class, which achieves a 15.6% improvement. On the condition of 10% samples per class, the performance of Semi-MMSC can further improve to 72.05%. In addition, Semi-MMSC obtains the top-5 accuracy of 100% on two semi-supervised settings. Overall, Semi-MMSC produces ship class predictions that are more confident and accurate in the semi-supervised multi-modal setting.

As shown in Table 3, for evaluating the generalization ability, the proposed approach (Semi-MMSC) conducts classification experiments on the Sentinel-1 dataset. Semi-MMSC adopts the same architecture of DSN [46] to extract frequency features as texture features and then builds the global-level interactive module based on these features to learn cross-modal correlation. Furthermore, the semi-supervised contrastive learning strategy is

imported. Semi-MMSC achieves better performance than other methods, which obtains an average accuracy of 93.90% and an average f1-score of 0.938.

Table 3. Generalization experiments on the Sentinel-1 dataset.

Classes	Accuracy				F1-Score			
	Semi-MMSC	DSN [46]	TL-CNN [46]	CNN [46]	Semi-MMSC	DSN [46]	TL-CNN [46]	CNN [46]
Forest	93.75 $\pm$ 1.76	98 $\pm$ 2.74	99.50 $\pm$ 1.12	96.50 $\pm$ 1.37	0.9677 $\pm$ 0.009	0.958 $\pm$ 0.011	0.959 $\pm$ 0.011	0.949 $\pm$ 0.033
Water	100 $\pm$ 0	100 $\pm$ 0	100 $\pm$ 0	100 $\pm$ 0	1 $\pm$ 0	1 $\pm$ 0	1 $\pm$ 0	1 $\pm$ 0
Agriculture	100 $\pm$ 0	93.50 $\pm$ 2.24	94.50 $\pm$ 2.09	95.00 $\pm$ 3.06	0.969 $\pm$ 0.008	0.959 $\pm$ 0.014	0.96 $\pm$ 0.018	0.955 $\pm$ 0.023
Industrialbuilding	95 $\pm$ 0	90 $\pm$ 5.59	81 $\pm$ 8.77	79.50 $\pm$ 4.47	0.95 $\pm$ 0	0.907 $\pm$ 0.046	0.832 $\pm$ 0.065	0.804 $\pm$ 0.052
Residential	100 $\pm$ 0	100 $\pm$ 0	97.50 $\pm$ 3.06	94.00 $\pm$ 5.18	0.987 $\pm$ 0	0.990 $\pm$ 0.010	0.958 $\pm$ 0.014	0.926 $\pm$ 0.037
Skyscraper	100 $\pm$ 0	92.00 $\pm$ 3.26	91.50 $\pm$ 7.42	89.00 $\pm$ 5.76	1 $\pm$ 0	0.887 $\pm$ 0.029	0.844 $\pm$ 0.025	0.813 $\pm$ 0.045
Storagetank	85 $\pm$ 0	85.50 $\pm$ 6.71	82.00 $\pm$ 4.11	72.50 $\pm$ 6.85	0.877 $\pm$ 0.007	0.858 $\pm$ 0.035	0.817 $\pm$ 0.036	0.741 $\pm$ 0.053
Container	85.25 $\pm$ 3.88	84.50 $\pm$ 4.11	73.50 $\pm$ 10.98	69.50 $\pm$ 9.25	0.88 $\pm$ 0	0.873 $\pm$ 0.023	0.811 $\pm$ 0.075	0.762 $\pm$ 0.0457
Average	93.90 $\pm$ 0.21	92.94 $\pm$ 1.05	89.94 $\pm$ 1.75	87.00 $\pm$ 2.01	0.938 $\pm$ 0.002	0.929 $\pm$ 0.011	0.898 $\pm$ 0.019	0.869 $\pm$ 0.022

The precision–recall curves of our proposed approach and other competitors respectively on two label fractions settings are demonstrated in Figure 3. We compare Semi-MMSC with some state-of-the-art semi-supervised learning methods (i.e., Comatch and Semicls) and some fusion methods. From the results, it is observed that the proposed Semi-MMSC significantly outperforms other methods in four different individual classes (the class ‘small boats’ with the most samples, the class ‘tug boats’ with the fewest samples, the class ‘medium passenger ship’ and the class ‘medium other ship’) at the first label fraction setting. It is noted that the class ‘medium passenger ship’ and the class ‘medium other ship’ are difficult to distinguish. On the second label fraction setting, Semi-MMSC is inferior to Semicls and Fixmatch on the class ‘small boat’, which can be attributed to the fact that these methods may only focus on the difference of the ship shape, like small and medium. On the class ‘tug boat’, the break-even point of Semi-MMSC and Semicls is roughly equal. In general, Semicls and Fixmatch have poor performance on other classes, especially when few samples are obtained. Overall, Semi-MMSC demonstrates the robustness of the semi-supervised multi-modal ship classification task.

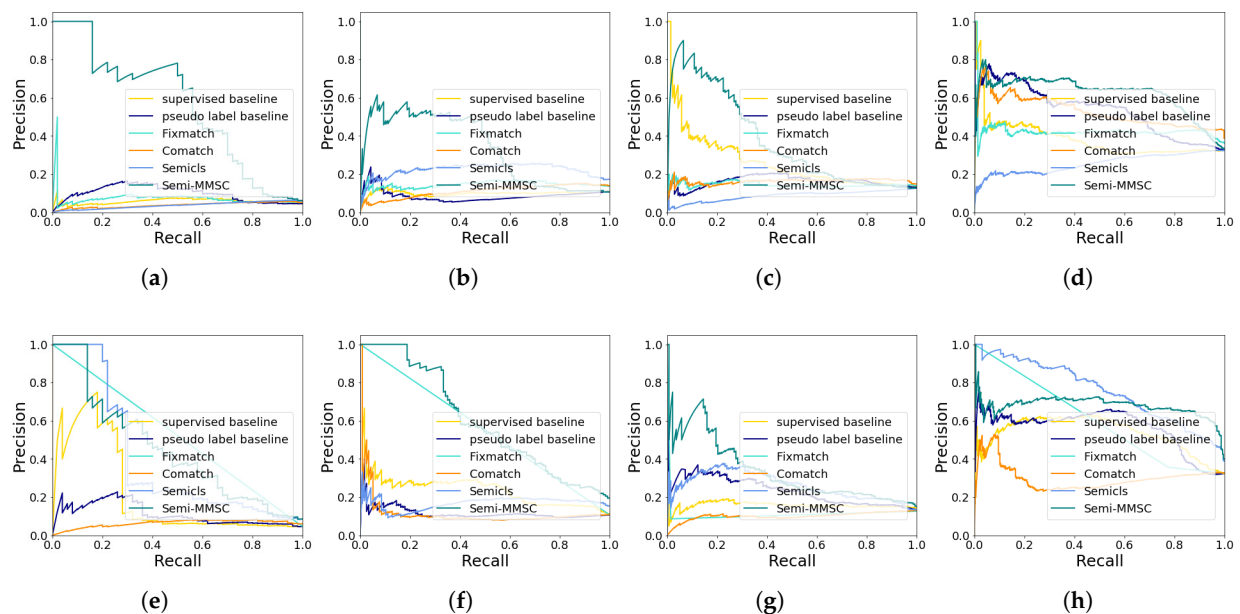


Figure 3. precision–recall curves of our approach and other competitors on two label fraction settings: (a–d) are the precision–recall curves of the classes ‘tug boat’, ‘medium passenger ship’, ‘medium other ship’, and ‘small boat’ on the first label fraction setting, respectively; (e–h) are the precision–recall curves of the classes ‘tug boat’, ‘medium passenger ship’, ‘medium other ship’, and ‘small boat’ on the second label fraction setting.

The confusion matrices of the proposed approach on the VAIS dataset are depicted in Figure 4. As displayed in Figure 4a, the diagonal elements of the confusion matrix denote the correct sample numbers of each class under the first label fraction setting. The class specific accuracy (%) is shown in Figure 4b, where the major confusion occurs between class medium passenger ships and class medium ‘other’ ships. We observe that images of these two classes are similar to each other and the sample number of two classes is also close. In addition, the confusion matrices under the second label fraction setting are shown in Figure 4c,d, the class specific accuracy of each class becomes also imbalanced, classes with a large amount of samples have higher accuracy (i.e., small boats, sailing ships) and classes with few samples have lower accuracy (i.e., tug boats).

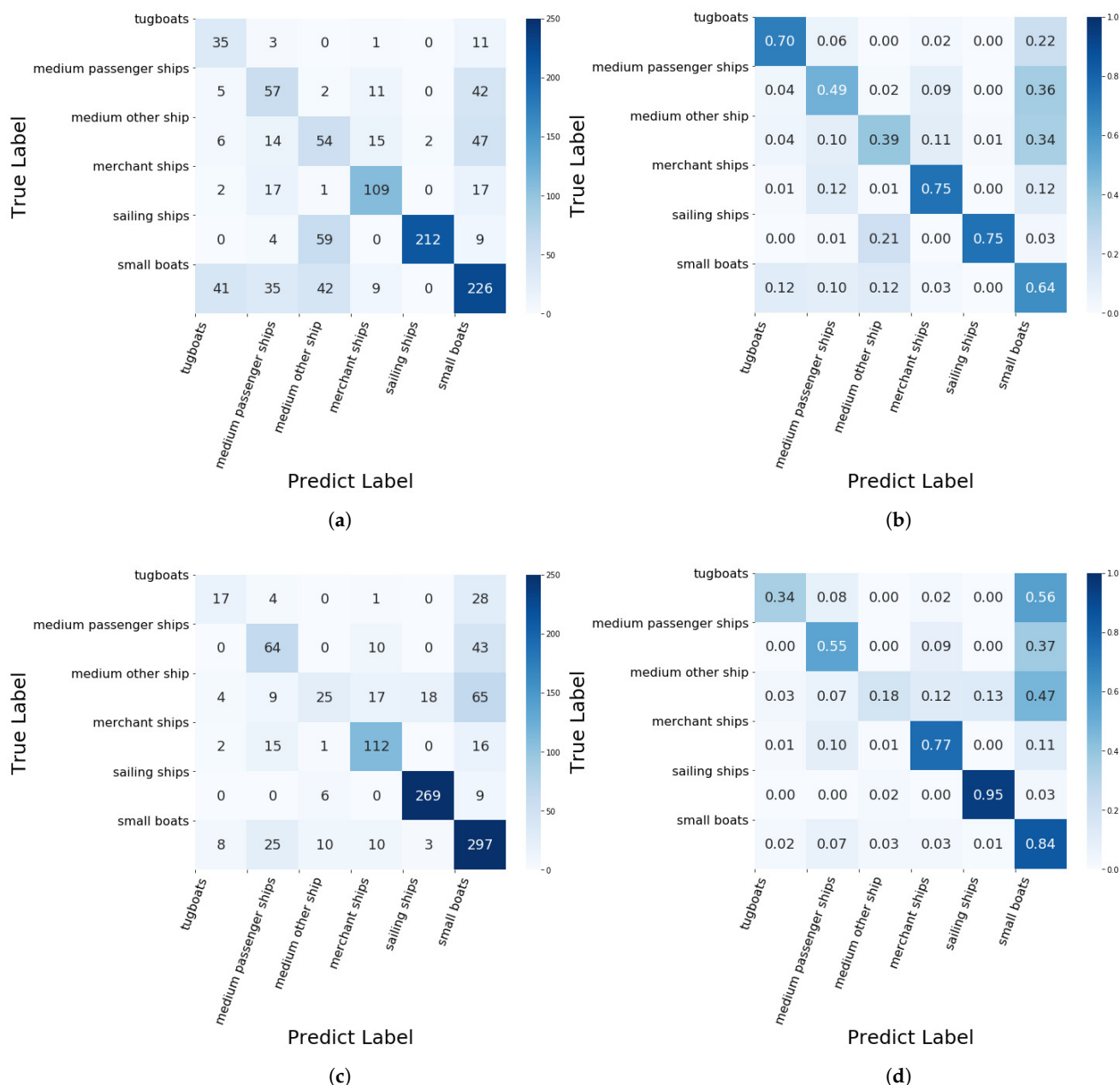


Figure 4. The confusion matrices of the proposed approach on two label fraction settings, (a,c) denote the confusion matrices of correct sample numbers of each class; (b,d) denote the confusion matrices of the class specific accuracy.

4.7. Ablation Study

An extensive ablation study is performed to verify the effectiveness of different components in the proposed approach.

Effectiveness of Multi-level Components: The proposed approach performs the cross-modal feature correlation to learn complementary information. To prove the effectiveness of local-level and global-level components, the corresponding experiment is designed. The local-level fusion module only simply concentrates features of different modalities based on the pix-level reconstruct loss. The global-level interactive module implements the feature interaction among the global features of the two modalities with local-level fused features as a medium. As shown in Table 4, comparing the employment of only local-level feature fusion, the proposed approach achieves 57.26% and 65.99% on the Top-1 accuracy metric, respectively, on two label fraction settings. Based on the multi-level cross-modal interactive architecture, Semi-MMSC further improves Top-1 accuracy by 6.43% and 4.87%.

Table 4. Effectiveness of local-level and global-level components.

Methods	10 Samples per Class		10% Samples per Class	
	Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc
Local-level fusion	57.2610	100.00	65.9926	100.00
Global-level fusion	55.2390	100.00	67.1875	100.00
Semi-MMSC	63.6949	100.00	72.0588	100.00

Effectiveness of Different Backbones: To prove the robustness of the ResNet backbone, we select the transformer architecture [47] as the backbone and conduct the comparison experiment, which is shown in Table 5. Compared with the ResNet backbone, the transformer backbone increases Top-1 accuracy under the first label fraction setting by 13.6949% while decreasing Top-1 accuracy under the second label fraction setting by 2.9412%. It is worth noting that the proposed approach can outperform the previous methods no matter which backbone is chosen. Meanwhile, the ResNet backbone can capture more detailed information directly from input images while the transformer backbone relies on labeled samples to learn global semantic information. This indicates that the transformer architecture captures multidimensional information such as visual features, spatial structure, and styles of images, while ResNet captures important information such as shape, texture, objects, and scene in images through convolutional neural networks. When more labeled samples are provided for training, the transformer can exhibit better performance. For adapting the condition of a few samples, we select the ResNet as the backbone.

Table 5. Effectiveness of different backbones.

Backbones	10 Samples per Class		10% Samples per Class	
	Top-1 acc	Top-5 acc	Top-1 acc	Top-5 acc
Transformer	50.0000	100.00	75.0000	100.00
ResNet	63.6949	100.00	72.0588	100.00

Then, we give an in-depth analysis of our design for the semi-supervised contrastive learning strategy, and demonstrate the benefits of them through comparative experiments.

Effectiveness of Contrastive Representations: Table 6 provides some results about self-supervised contrastive learning strategy. If we train the model with only pseudo label learning to predict unlabeled training data, we can obtain Top-1 accuracy of 39.61% and 44.39%. If we use the contrastive head module in the feature space, Top-1 accuracy improves by 18.65% and 25.09%. In other words, pseudo-label-based self-supervised learning demonstrates poor classification performance, which is the motivation of the self-supervised contrastive learning strategy. We believe this will help the model learn class-wise complementary information during training, and it can be seen that our method gains 5.42% and 2.57% benefits with the whole function from the results reported in Table 5.

Table 6. The influence of different semi-supervised contrastive learning modules based on the VAIS dataset.

Adding Module	Top-1 Accuracy on the First Label Fraction	Top-1 Accuracy on the Second Label Fraction
baseline	39.6140	44.3934
add contrastive head module	58.2721	69.4853
add memory module	63.6949	72.0588

Memory Feature Selection Strategy: In this part, we discuss the impact of different memory selection strategies. Three alternative sample selection methods are listed as follows:

- using correlated features of different modalities as the input of the memory mechanism, including ω^1 , ω^2 , and ω^f .
- using the pair-wise features, including three pairs (ω^1, ω^f) , (ω^2, ω^f) , and (ω^1, ω^2) .
- using the feature similarity representations, which are the output of the contrastive head.

As shown in Table 7, selecting the combination of features (i.e., (ω^1, ω^f)) performs even worse than the single-modal feature. The reason is that local-level fused features have much texture information and less semantic information, which shows that the simple combination between features of different modalities has negative effects on performance. Additionally, using contrastive representations is superior to using similarity representations by 7.16% and 5.51%, which demonstrates that the former can store more class information.

Table 7. Impact of different memory feature selection strategy on the model.

Feature Selection Strategy	Accuracy (%)	
	10 Samples per Class	10% Samples per Class
using single-modal features	56.0662	63.2353
	52.2978	64.3382
	59.2831	65.0397
using pair-wise features	54.4118	61.3971
	52.2978	64.1544
	51.6544	65.6541
using the feature similarity representations	56.5257	62.0404
	53.9522	62.9596
	55.5147	66.5441
using contrastive representations	63.6949	72.0588

Evaluation of Weight Assignment for Loss Functions: We analyze the impact of several hyper-parameters, which are, respectively, unsupervised contrastive loss weight α and self-supervised divergence loss weight β . The values of loss weights are selected from 0.001, 0.01, 0.1, 1, 10, 100 correspondingly, and the results are shown in Figure 5. The two losses are both the main optimization objective as they determines the quality of unsupervised learning and have an important impact on the final prediction. It can be observed that the optimal performance can be obtained when the values of (α, β) are set to (10, 1) and (1, 1), respectively, on two different label fraction settings.

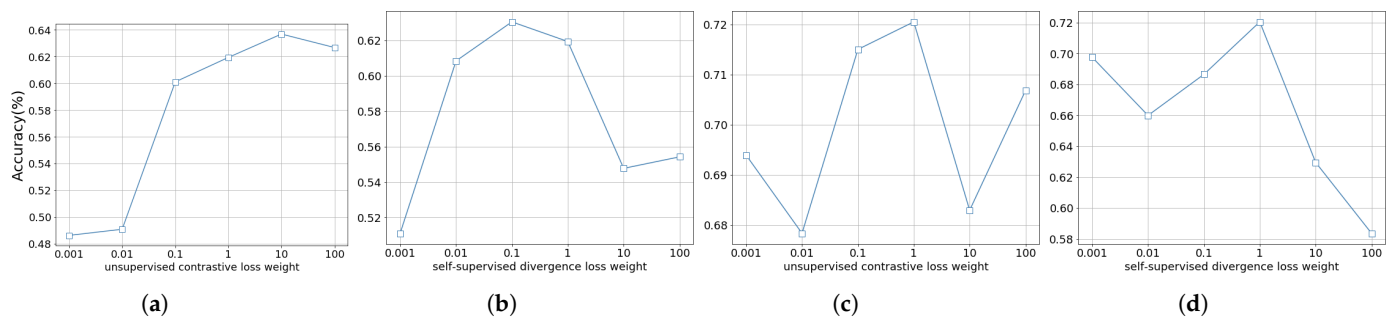


Figure 5. The analysis of default hyper-parameters on two different label fraction settings, unsupervised contrastive loss weight α and self-supervised divergence loss weight β . (a,c) Varying weight α , which controls the unsupervised contrastive learning. (b,d) Varying weight β for the self-supervised learning.

5. Conclusions

In this paper, we propose a novel Semi-supervised Multi-Model Ship Classification approach, which consists of the multi-level interactive network and semi-supervised contrastive learning strategy. The multi-level cross-modal interactive network encoders deep correlated features containing cross-modal complementary information in a multi-level feature interaction manner. Then, for taking advantage of unlabeled data in the training stage, the semi-supervised contrastive learning strategy is designed to enrich supervised signals of unlabeled data to guide the network training. Extensive experiments on the public datasets demonstrate that our approach obtains a state-of-the-art performance. In the future, we will focus on applying this approach in the field of the fishery industry and further extend the research by designing more refined feature extraction modules like transformers.

Author Contributions: X.S. and Z.C. conceived this study, designed the computational algorithms, wrote the program code, and wrote the manuscript. F.Z., J.G., J.Z. and P.L. proposed some valuable suggestions and guided the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Science and Technology Planning Project of Liaoning Province (2023JH2/101300092), and the Dalian Science and Technology Innovation Foundation (2023JJ13SN065).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Zhang, T.; Zhang, X.; Ke, X.; Liu, C.; Xu, X.; Zhan, X.; Wang, C.; Ahmad, I.; Zhou, Y.; Pan, D.; et al. HOG-ShipCLSNet: A Novel Deep Learning Network with HOG Feature Fusion for SAR Ship Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5210322. [\[CrossRef\]](#)
2. Li, Y.; Lai, X.; Wang, M.; Zhang, X. C-SASO: A Clustering-Based Size-Adaptive Safer Oversampling Technique for Imbalanced SAR Ship Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5231112. [\[CrossRef\]](#)
3. Song, Y.; Li, J.; Gao, P.; Li, L.; Tian, T.; Tian, J. Two-Stage Cross-Modality Transfer Learning Method for Military-Civilian SAR Ship Recognition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4506405. [\[CrossRef\]](#)
4. Santos, C.E.; Bhanu, B. Dyfusion: Dynamic IR/RGB Fusion for Maritime Vessel Recognition. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 1328–1332. [\[CrossRef\]](#)
5. Yang, X.; Zhang, X.; Wang, N.; Gao, X. A Robust One-Stage Detector for Multiscale Ship Detection With Complex Background in Massive SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5217712. [\[CrossRef\]](#)

6. Huang, L.; Li, W.; Chen, C.; Zhang, F.; Lang, H. Multiple features learning for ship classification in optical imagery. *Multimed. Tools Appl.* **2018**, *77*, 13363–13389. [\[CrossRef\]](#)
7. Li, L.; Zhou, Z.; Wang, B.; Miao, L.; Zong, H. A Novel CNN-Based Method for Accurate Ship Detection in HR Optical Remote Sensing Images via Rotated Bounding Box. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 686–699. [\[CrossRef\]](#)
8. Lang, H.; Xi, Y.; Zhang, X. Ship Detection in High-Resolution SAR Images by Clustering Spatially Enhanced Pixel Descriptor. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5407–5423. [\[CrossRef\]](#)
9. Hu, J.; Zhi, X.; Jiang, S.; Tang, H.; Zhang, W.; Bruzzone, L. Supervised Multi-Scale Attention-Guided Ship Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5630514. [\[CrossRef\]](#)
10. Ren, Y.; Yang, J.; Zhang, Q.; Guo, Z. Ship recognition based on Hu invariant moments and convolutional neural network for video surveillance. *Multimed. Tools Appl.* **2021**, *80*, 1343–1373. [\[CrossRef\]](#)
11. Du, Q.; Zhang, Y.; Yang, X.; Liu, W. Ship target classification based on Hu invariant moments and ART for maritime video surveillance. In Proceedings of the 2017 4th International Conference on Transportation Information and Safety (ICTIS), Banff, AB, Canada, 8–10 August 2017; pp. 414–419. [\[CrossRef\]](#)
12. Chen, S.; Zhan, R.; Wang, W.; Zhang, J. Domain Adaptation for Semi-Supervised Ship Detection in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4507405. [\[CrossRef\]](#)
13. Xu, Y.; Lang, H.; Niu, L.; Ge, C. Discriminative Adaptation Regularization Framework-Based Transfer Learning for Ship Classification in SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1786–1790. [\[CrossRef\]](#)
14. Oliveau, Q.; Sahbi, H. From Transductive to Inductive Semi-Supervised Attributes for Ship Category Recognition. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 4827–4830. [\[CrossRef\]](#)
15. Li, D.; Liang, Q.; Liu, H.; Liu, Q.; Liu, H.; Liao, G. A Novel Multidimensional Domain Deep Learning Network for SAR Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5203213. [\[CrossRef\]](#)
16. Liu, K.Z.; Lee, P.J.; Xu, G.C.; Chang, B.H. SIFT-Enhanced CNN Based Objects Recognition for Satellite Image. In Proceedings of the 2020 IEEE International Conference on Consumer Electronics—Taiwan (ICCE-Taiwan), Taoyuan, Taiwan, 28–30 September 2020; pp. 1–2. [\[CrossRef\]](#)
17. Chendeb El Rai, M.; Giraldo, J.H.; Al-Saad, M.; Darweech, M.; Bouwmans, T. SemiSegSAR: A Semi-Supervised Segmentation Algorithm for Ship SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4510205. [\[CrossRef\]](#)
18. Wang, C.; Zhang, H.; Wu, F.; Jiang, S.; Zhang, B.; Tang, Y. A Novel Hierarchical Ship Classifier for COSMO-SkyMed SAR Data. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 484–488. [\[CrossRef\]](#)
19. Wu, F.; Wang, C.; Jiang, S.; Zhang, H.; Zhang, B. Classification of Vessels in Single-Pol COSMO-SkyMed Images Based on Statistical and Structural Features. *Remote Sens.* **2015**, *7*, 5511–5533. [\[CrossRef\]](#)
20. Ma, J.; Ma, Y.; Li, C. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion* **2019**, *45*, 153–178. [\[CrossRef\]](#)
21. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2016**, *33*, 100–112. [\[CrossRef\]](#)
22. Zhao, W.; Lu, H.; Wang, D. Multisensor Image Fusion and Enhancement in Spectral Total Variation Domain. *IEEE Trans. Multimed.* **2018**, *20*, 866–879. [\[CrossRef\]](#)
23. Li, J.; Huo, H.; Li, C.; Wang, R.; Feng, Q. AttentionFGAN: Infrared and Visible Image Fusion Using Attention-Based Generative Adversarial Networks. *IEEE Trans. Multimed.* **2021**, *23*, 1383–1396. [\[CrossRef\]](#)
24. Guo, X.; Nie, R.; Cao, J.; Zhou, D.; Mei, L.; He, K. FuseGAN: Learning to Fuse Multi-Focus Image via Conditional Generative Adversarial Network. *IEEE Trans. Multimed.* **2019**, *21*, 1982–1996. [\[CrossRef\]](#)
25. Liu, Y.; Chen, X.; Wang, Z.; Wang, Z.J.; Ward, R.K.; Wang, X. Deep learning for pixel-level image fusion: Recent advances and future prospects. *Inf. Fusion* **2018**, *42*, 158–173. [\[CrossRef\]](#)
26. Qi, S.; Ma, J.; Lin, J.; Li, Y.; Tian, J. Unsupervised Ship Detection Based on Saliency and S-HOG Descriptor from Optical Satellite Images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1451–1455. [\[CrossRef\]](#)
27. Li, W.; Chen, C.; Su, H.; Du, Q. Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3681–3693. [\[CrossRef\]](#)
28. Lin, H.; Song, S.; Yang, J. Ship Classification Based on MSHOG Feature and Task-Driven Dictionary Learning with Structured Incoherent Constraints in SAR Images. *Remote Sens.* **2018**, *10*, 190. [\[CrossRef\]](#)
29. Zhang, E.; Wang, K.; Lin, G. Classification of Marine Vessels with Multi-Feature Structure Fusion. *Appl. Sci.* **2019**, *9*, 2153. [\[CrossRef\]](#)
30. Yang, G.; Lang, H. Semisupervised Heterogeneous Domain Adaptation via Dynamic Joint Correlation Alignment Network for Ship Classification in SAR Imagery. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4508105. [\[CrossRef\]](#)
31. Jiang, M.; Chen, X.; Xu, L.; Clausi, D.A. Semi-Supervised Sea Ice Classification of SAR Imagery Based on Graph Convolutional Network. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 1031–1034. [\[CrossRef\]](#)
32. Oliveau, Q.; Sahbi, H. Semi-Supervised Deep Attribute Networks for Fine-Grained Ship Category Recognition. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 6871–6874. [\[CrossRef\]](#)

33. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. MixMatch: A Holistic Approach to Semi-Supervised Learning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; Volume 32.
34. Iscen, A.; Tolias, G.; Avrithis, Y.; Chum, O. Label Propagation for Deep Semi-Supervised Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
35. Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.E.; McGuinness, K. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [\[CrossRef\]](#)
36. Hu, Z.; Yang, Z.; Hu, X.; Nevatia, R. SimPLE: Similar Pseudo Label Exploitation for Semi-Supervised Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15099–15108.
37. Jeong, J.; Lee, S.; Kim, J.; Kwak, N. Consistency-based Semi-supervised Learning for Object detection. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; Volume 32.
38. Zhang, M.M.; Choi, J.; Daniilidis, K.; Wolf, M.T.; Kanan, C. VAIS: A Dataset for Recognizing Maritime Imagery in the Visible and Infrared Spectrums. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Boston, MA, USA, 7–12 June 2015.
39. Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. GMES Sentinel-1 mission. *Remote Sens. Environ.* **2012**, *120*, 9–24. [\[CrossRef\]](#)
40. Chen, J.; Chen, K.; Chen, H.; Li, W.; Zou, Z.; Shi, Z. Contrastive Learning for Fine-Grained Ship Classification in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4707916. [\[CrossRef\]](#)
41. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 596–608.
42. Li, J.; Xiong, C.; Hoi, S.C. CoMatch: Semi-Supervised Learning With Contrastive Graph Regularization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 9475–9484.
43. Yang, F.; Wu, K.; Zhang, S.; Jiang, G.; Liu, Y.; Zheng, F.; Zhang, W.; Wang, C.; Zeng, L. Class-Aware Contrastive Semi-Supervised Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14421–14430.
44. Xu, H.; Xiao, H.; Hao, H.; Dong, L.; Qiu, X.; Peng, C. Semi-supervised learning with pseudo-negative labels for image classification. *Knowl. Based Syst.* **2023**, *260*, 110166. [\[CrossRef\]](#)
45. Li, H.; Wu, X.J.; Durrani, T. NestFuse: An Infrared and Visible Image Fusion Architecture Based on Nest Connection and Spatial/Channel Attention Models. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9645–9656. [\[CrossRef\]](#)
46. Huang, Z.; Datcu, M.; Pan, Z.; Lei, B. Deep SAR-Net: Learning objects from signals. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 179–193. [\[CrossRef\]](#)
47. Chen, C.F.R.; Fan, Q.; Panda, R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.

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.