remote sensing

MDPI

Article

Contrastive Learning Network Based on Causal Attention for Fine-Grained Ship Classification in Remote Sensing Scenarios

Chaofan Pan, Runsheng Li *, Qing Hu, Chaoyang Niu, Wei Liu and Wanjie Lu

Institute of Data and Target Engineering, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China; 631505020222@mails.cqjtu.edu.cn (C.P.); huq1115@163.com (Q.H.); ncy_100@163.com (C.N.); greatliuliu@163.com (W.L.); lwj285149763@163.com (W.L.)
* Correspondence: lirunsheng27@163.com; Tel.: +86-181-3777-8669

Abstract: Fine-grained classification of ship targets is an important task in remote sensing, having numerous applications in military reconnaissance and sea surveillance. Due to the influence of various imaging factors, ship targets in remote sensing images have considerable inter-class similarity and intra-class difference, which brings significant challenges to fine-grained classification. In response, we developed a contrastive learning network based on causal attention (C2Net) to improve the model's fine-grained identification ability from local details. The asynchronous feature learning mode of "decoupling + aggregation" is adopted to reduce the mutual influence between local features and improve the quality of local features. In the decoupling stage, the feature vectors of each part of the ship targets are de-correlated using a decoupling function to prevent feature adhesion. Considering the possibility of false associations between results and features, the decoupled part is designed based on the counterfactual causal attention network to enhance the model's predictive logic. In the aggregation stage, the local attention weight learned in the decoupling stage is used to carry out feature fusion on the trunk feature weight. Then, the proposed feature re-association module is used to re-associate and integrate the target local information contained in the fusion feature to obtain the target feature vector. Finally, the aggregation function is used to complete the clustering process of the target feature vectors and fine-grained classification is realized. Using two large-scale datasets, the experimental results show that the proposed C2Net method had better fine-grained classification than other methods.

Keywords: ship targets; fine-grained classification; remote sensing images; causal attention; comparative learning

check for updates

Citation: Pan, C.; Li, R.; Hu, Q.; Niu, C.; Liu, W.; Lu, W. Contrastive Learning Network Based on Causal Attention for Fine-Grained Ship Classification in Remote Sensing Scenarios. *Remote Sens.* **2023**, *15*, 3393. https://doi.org/10.3390/rs15133393

Academic Editor: Andrzej Stateczny

Received: 10 May 2023
Revised: 26 June 2023
Accepted: 27 June 2023
Published: 3 July 2023

Copyright: © 2023 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 is an important type of remote sensing target and its classification and recognition are of great significance to both the military and civilian fields. In the military field, ship target identification is important in the deployment of maritime forces for reconnaissance and provides important intelligence support for command and decision-making. In civilian applications, accurate ship type identification plays a vital role in maritime rescue, fishery management, and anti-smuggling. Therefore, ship target recognition has many application prospects and values.

Ship target classification tasks can be roughly divided into target-level, coarse-grained, and fine-grained. Research on target-level classification [1,2] mainly focuses on distinguishing the target and background, which is essentially a binary task. Coarse-grained classification [3] focuses on categories with significant differences, such as aircraft carriers, warships, and merchant ships. Fine-grained classification [4–8] further distinguishes each category into sub-categories. Inter-class similarity and intra-class difference caused by different imaging factors (e.g., illumination, perspective change, and deformation) can make the fine-grained classification task more challenging, as shown in Figure 1. Therefore,

the high-precision, fine-grained classification of ship targets has become a research hotspot in computer vision.

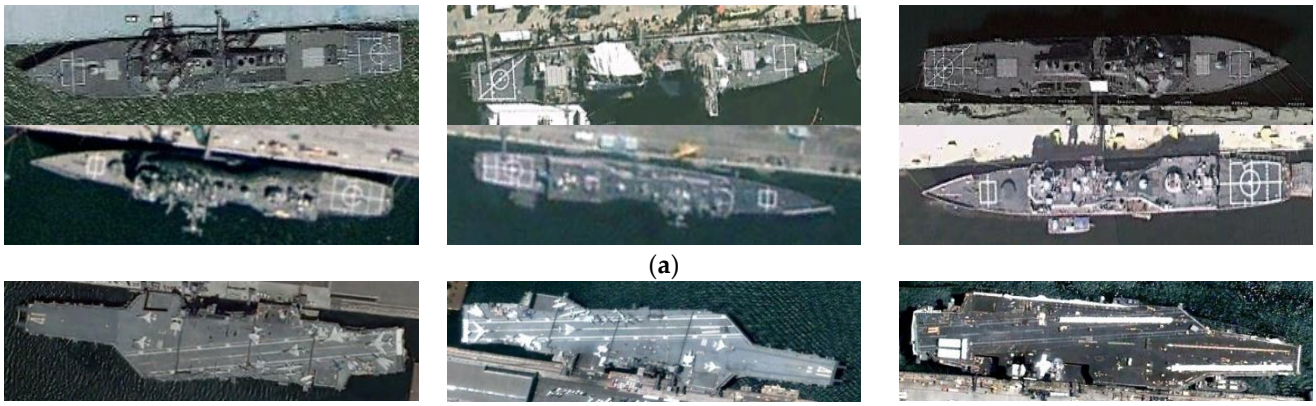


Figure 1. (**a**) Inter-class similarity: Destroyer (**first row**), Frigate (**second row**); (**b**) Intra-class difference: aircraft carrier.

Some strongly supervised methods use target local location tags [9] or key location tags [10] to train the model's localization and feature extraction ability for local areas. For example, Zhang et al. [11] used additional label information to train attributes to guide branches to achieve multi-level feature fusion. Chen et al. [12] conducted ship recognition by training a multi-level feature extraction network. This strong supervision method can improve the fine-grained classification effect. However, its high training information conditions, which require the local annotation of the dataset or the target area localization and slicing, results in low efficiency and limited applicability.

In contrast, the weak supervision method does not require additional supervision information and only uses image-level annotation to complete the model's training. Lin et al. [13] proposed the use of bilinear pooling to integrate features extracted from the same position by parallel branches to obtain more representational features. Huang et al. [14] connected the convolutional neural network (CNN) structure and the self-focused transformer structure in parallel to focus on the multi-scale features of ships. The global-to-local progressive learning module (GLPM) designed by Meng et al. [15] enhances the fine-grained feature extraction capability by promoting the information exchange between global and local features. In general, the weak supervision method greatly reduces dependence on auxiliary supervision information and is more practical.

The above methods focus on the discriminative details of the target. The strong supervision method reduces the model's learning difficulty and guides it to carry out targeted training by adding auxiliary training information through manual labeling. The weak supervision method realizes the automatic discriminative feature learning through the model structure's design. They are typical black-box deep learning models that ignore the real intrinsic causal relationship between the predicted results and the attention area. This causes the prediction results to rely on the false correlation between the two, hindering the model from learning the target's discriminative features.

To solve this problem, some studies have introduced causal reasoning into deep learning and explored causal reasoning networks. Early causal reasoning was mainly applied to statistical models with fewer variables and was not suitable for computer vision. Pearl et al. [16] pointed out that data fitting at the statistical level cannot establish a causal relationship between input and output; only models with counterfactual reasoning ability can make logical decisions similar to human brains. With the deepening of the research, the method combining causal reasoning and deep learning has become widely applied in computer vision. Rao et al. [17] used counterfactual causality to learn multi-head attention and analyze its influence on prediction. Xiong et al. [18] focused on the predictive logic and

interpretability of networks, concentrating on specific parts of ship targets by combining counterfactual causal attention and convolution filters and visualizing the decision basis. Counterfactual causal reasoning gives the model the ability to make decisions at the logical level and enhances the attention to local details.

The supervision information of the above methods comes from the output, and the model's training process is guided through the output loss. Chen et al. [19] pointed out that the fine-grained classification process is a push-and-pull process, such that different feature classes are separated while similar classes are aggregated. The features can only be synchronously pushed and pulled by attaching the loss function to the output end. Features between similar subclasses cannot be completely separated, resulting in certain adhesion, which can adversely affect feature aggregation. To address the limitations of the synchronous push-pull mode, a contrast learning-based network was proposed in which the separation and aggregation stages of features are separated in an asynchronous mode and the loss supervision is carried out for each stage. Contrast learning is an unsupervised learning paradigm [20], which uses a double-branch structure to construct homologous and non-homologous image pairs using different image processing, guiding the push-and-pull process by comparing similarities between image features. Although the current comparative learning method has lower requirements for the annotation of training data and is more convenient in practical applications, it focuses mainly on global image features and rarely considers the local feature details crucial for fine-grained classification.

To solve these problems, a contrastive learning network based on causal attention (C2Net) is proposed by making full use of the local detailed features of ship targets. C2Net can take into account the asynchronous learning of local and global features to improve local feature quality.

The main contributions of this paper are as follows:

1. To improve local feature quality, a causal attention model was developed based on feature decoupling (FD-CAM). A FD-CAM uses a decoupling function to guide the feature separation operation, eliminate the adhesion between the local features, and it uses counterfactual causal inference architecture to learn the true association between the features and classification results to reduce the influence of false association problems in data-driven deep learning.
2. A feature aggregation module (FAM) is proposed for the weighted fusion and re-association of local features. A FAM weights the features extracted from the trunk network using the local attention weight from the FD-CAM's learning to obtain the locally decoupled fusion features. Fusion features are input into the feature recoupling module (FRM) to realize the re-association between local features; the aggregation function is used to guide the clustering process of feature vectors.
3. Extensive experiments were conducted on two publicly available ship datasets to evaluate the performance of the proposed approach. The experimental results showed that our method achieved better results compared to other methods, showing strong fine-grained classification ability.

The remainder of this paper is organized as follows: Section 2 introduces the C2Net overall structure and details the principle and fusion modes of feature decoupling and the aggregation branches. Section 3 describes the study datasets, evaluation indicators, implementation details, and the performance of C2Net, while Section 4 summarizes the work of this paper.

2. Proposed Method

2.1. Overview of the Method

The structure of the contrastive learning network based on causal attention (C2Net) proposed in this paper is shown in Figure 2. The input image is divided into two branches, each randomly performing a different image transformation operation. The transformed image is input into the CNN backbone network to extract features and to obtain the feature map $\mathbf{F} \in \mathbb{R}^{W \times H \times C}$, where W, H, and C are the feature map's width, height, and

channel number, respectively. **F** is then input into the causal attention model based on feature decoupling (FD-CAM) to learn the local feature attention diagram of the ship target. The details are in Section 2.2. Finally, the feature aggregation module (FAM) uses the feature attention mapping from the FD-CAM to fuse the features and the re-association and clustering of the fused features to achieve fine-grained target recognition. The feature aggregation module is described in detail in Section 2.3.

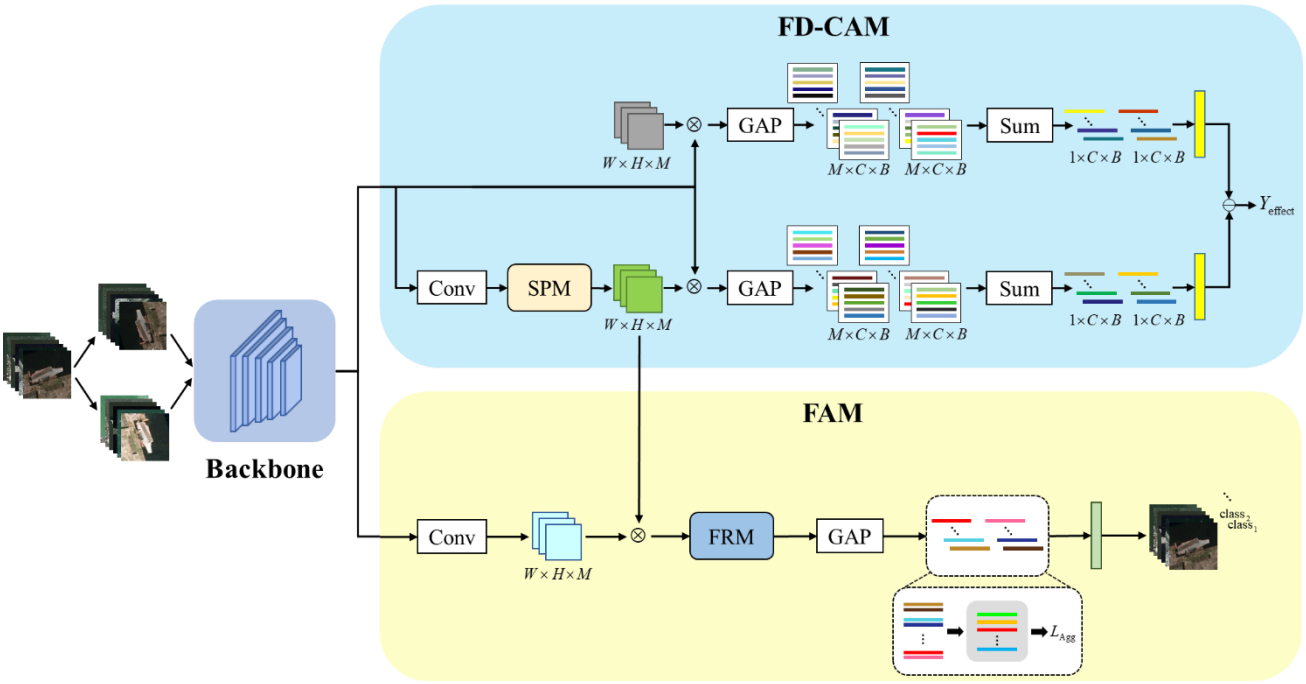


Figure 2. C2Net structure. GAP means global average pooling operation. SPM is a strip pooling module. FRM indicates the feature re-association module. Y_{effect} shows the true effect of the bulls' attention on the forecast outcomes. L_{Agg} is an aggregation loss function used to guide the feature polymerization process.

2.2. Causal Attention Model Based on Feature Decoupling

The counterfactual attention model, based on causal reasoning [16], guides the learning process of the attention diagram by establishing the causal relationship between the features and predicted results, making the model's prediction logic more transparent and avoiding the unexplainability of the black box model in deep learning. Since the attention map represents the intensity of attention on different areas of the image, improving local attention learning is vital for accurate classification. In this study, a causal attention model based on feature decoupling (FD-CAM) is proposed. An FD-CAM is a typical counterfactual causal inference framework [17,18] consisting of two branches: a multiple attention branch and a counterfactual attention branch.

Multiple attention branches are used to learn the local attention features of the target. For the input feature map **F**, CNNs are first used to extract advanced features. Then, context information is collected through a pooling operation for feature aggregation to obtain the multi-head attention mapping $\mathbf{A} \in \mathbb{R}^{W \times H \times M}$, where M is the number of attention channels. Considering the extreme aspect ratio of ship targets, the strip pooling module (SPM) [21] is selected to pool F, as shown in Figure 3. Strip pooling (SP) is a long pool core that can avoid the interference of other spatial dimensions while conducting long-distance feature correlation along one spatial dimension and facilitating the extraction of local features. The SPM can account for global and local features by deploying SP in horizontal and vertical directions and has good adaptability to long ship targets. The feature map obtained by the SPM's pooling is copied along a specific dimension to restore the shape of $W \times H \times M$.

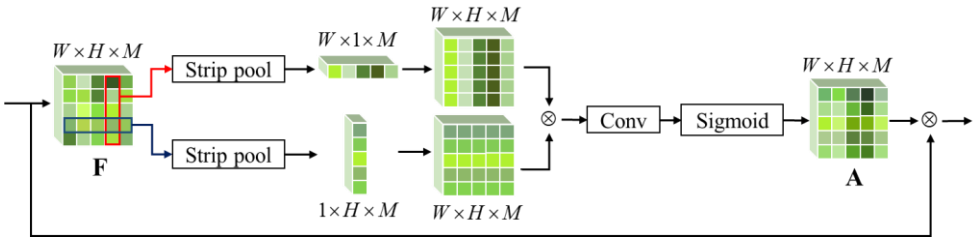


Figure 3. SPM structure.

To obtain the regional feature set $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_M\} \in \mathbb{R}^{M \times C \times 1}$, $\mathbf{A}$ and $\mathbf{F}$ are multiplied (element by element) and the global average pooling is performed using the formula:

$$\mathbf{r}_m = \{\mathbf{r}_{mc} | \mathbf{r}_{mc} = \mathrm{GAP}(\mathbf{A}_m \odot \mathbf{F}_c), c \in \{1, 2, \cdots, C\}\}, m \in \{1, 2, \cdots, M\} \tag{1}$$

where $\mathbf{r}_m$ is the m-th regional feature vector; $\mathbf{r}_{mc}$ is the value of the c-th element in the region feature vector $\mathbf{r}_m$; $\mathbf{A}_m$ is the M-th attention map in $\mathbf{A}$ and $\mathbf{F}_c$ is the c-th feature map in $\mathbf{F}$; GAP is the global average pooling; $\odot$ is the element by element multiplication. The local features of the target are obtained by summing M regional feature vectors, as shown below:

$$\mathbf{l} = \sum_{m=1}^{M} \mathbf{r}_m \tag{2}$$

$\mathbf{l}$ is essentially a feature weighted by spatial attention, which can be obtained by directly summing the multiple attention forces along the channel direction, multiplying the input features element by element, and implementing the global average pooling. The proof is as follows.

Suppose the sum of multiple attention forces along the channel direction is $\mathbf{A}_{\mathrm{sum}}$ and the calculation is as follows:

$$\mathbf{A}_{\mathrm{sum}} = \sum_{m=1}^{M} \mathbf{A}_m \tag{3}$$

Further derivation leads to the following:

$$\begin{aligned} \mathrm{GAP}(\mathbf{A}_{\mathrm{sum}} \odot \mathbf{F}) &= \mathrm{GAP}(\sum_{m=1}^{M} \mathbf{A}_m \odot \mathbf{F}) \\ &= \mathrm{GAP}(\sum_{m=1}^{M} (\mathbf{A}_m \odot \mathbf{F})) \\ &= \sum_{m=1}^{M} \mathrm{GAP}(\mathbf{A}_m \odot \mathbf{F}) \\ &= \sum_{m=1}^{M} \mathbf{r}_m \\ &= \mathbf{l} \end{aligned} \tag{4}$$

The fully connected layer is selected as the classifier $\mathrm{C}(\cdot)$ and $\mathbf{l}$ is input into $\mathrm{C}(\cdot)$ to obtain the predicted output of the multiple attention branches:

$$Y_A = \mathrm{C}(\mathbf{l}) \tag{5}$$

Counterfactual attention branches are used to explore the influence of multiple attention branches on the final classification results by counterfactual intervention; the input is a randomly generated false attention diagram. The regional feature vector and the local feature vector of the false attention map are calculated in the same way as the multi-head attention using the following formulas:

$$\hat{\mathbf{r}}_m = \left\{\hat{\mathbf{r}}_{mc} \middle| \hat{\mathbf{r}}_{mc} = \mathrm{GAP}(\hat{\mathbf{A}}_m \odot \mathbf{F}_c), c \in \{1, 2, \cdots, C\}\right\}, m \in \{1, 2, \cdots, M\} \tag{6}$$

$$\hat{\mathbf{I}} = \sum_{m=1}^{M} \hat{\mathbf{r}}_m \tag{7}$$

The prediction results obtained according to the counterfactual local feature vector are:

$$Y_{\hat{\mathbf{A}}} = \mathrm{C}(\hat{\mathbf{I}}) \tag{8}$$

By calculating the difference between $Y_{\mathbf{A}}$ and $Y_{\hat{\mathbf{A}}}$, we can quantify the impact of the multi-head attention features on the prediction results:

$$Y_{\text{effect}} = Y_{\mathbf{A}} - Y_{\hat{\mathbf{A}}} \tag{9}$$

Y_{effect} can be understood as the learning objective of the local attention mechanism, that is, the internal correlation between the multiple attention and the predicted results after eliminating the interference of false attention. The cross-entropy loss function $\mathrm{CE}(\cdot)$ is used to guide the learning of association, given by the expression:

$$L_{\text{effect}} = \mathrm{CE}(Y_{\text{effect}}, y_{label}) \tag{10}$$

where y_{label} is the truth tag.

The coupling between the features in counterfactual causal reasoning networks is an important factor affecting local attention learning. Inspired by the work of Chen et al. [19], we draw on the feature separation idea of contrast learning, carrying out separation through manual intervention to realize the decoupling of regional features of input images and local features between non-homologous images to obtain more distinctive attention features and prevent the mutual influence of local features in attention learning.

Regional feature decoupling is carried out within each feature set, as shown in Figure 4a. The feature vector in the regional feature set represents the filtering feature of local attention on input feature map **F**. The essence of the decoupling feature vector is to decouple local attention, which can improve the learning of local classification features by reducing the correlation between the channels of multiple attention maps. In decoupling, combining the vectors in the region feature set is necessary for feature separation. To avoid double computation and ensure computational efficiency, cyclic shift is adopted to sort through the vector set, and decoupling is performed with the vector at the corresponding position in the original set, as shown below:

$$\begin{aligned} \mathbf{R}' &= \mathrm{CycShift}(\mathbf{R}, s) \\ &= \{\mathbf{r}'_1, \mathbf{r}'_2, \cdots, \mathbf{r}'_M\} \\ &= \{\mathbf{r}_{1+s}, \mathbf{r}_{2+s}, \cdots, \mathbf{r}_s\}, s \in \{1, 2, \cdots, M-1\} \end{aligned} \tag{11}$$

$$L_{\mathbf{R}-\mathrm{Dec}} = \sum_{i=1}^{M} \sum_{j=1}^{M} \mathrm{Dec}(\mathbf{r}_i, \mathbf{r}'_j) \tag{12}$$

where $\mathrm{CycShift}(\cdot)$ is the cyclic shift; s is the shift step size; $\mathbf{R}'$ and $\mathbf{r}'$ are the regional feature set and feature vector obtained by the cyclic shift, respectively; $L_{\mathbf{R}}$ is the decoupling loss of regional features. $\mathrm{Dec}(\cdot)$ is the decoupling function, defined as follows:

$$\mathrm{Dec}(v_1, v_2) = -\log[0.5 \cdot (1 - \frac{v_1 \cdot v_2}{\|v_1\|_2 \cdot \|v_2\|_2})] \tag{13}$$

where v_1 and v_2 are one-dimensional vectors. Since the input images are transformed in two different ways, the feature aggregation of regional feature vectors between homologous images must be considered. Since two homologous maps correspond to the same target, the attention feature maps of the same channel should have the same focusing region, and the

regional feature vectors filtered by them share the same local features. Figure 4b presents the feature aggregation process. The aggregation function is defined as follows:

$$\mathrm{Agg}(v_1, v_2) = -\log[0.5 \cdot (1 + \frac{v_1 \cdot v_2}{\|v_1\|_2 \cdot \|v_2\|_2})] \tag{14}$$

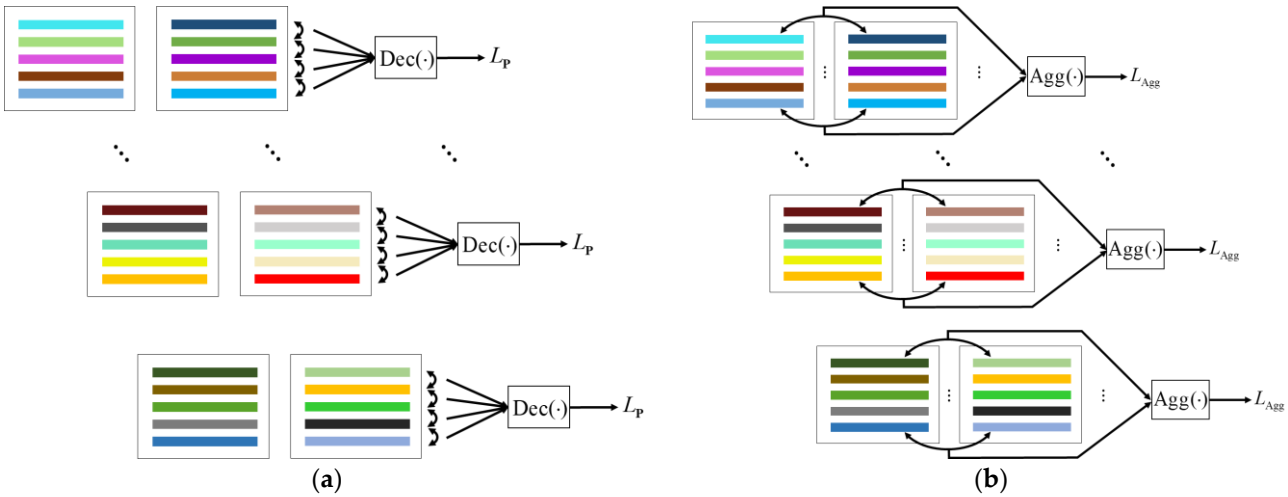


Figure 4. Schematic diagram of regional feature decoupling and aggregation. (**a**) Regional feature decoupling; (**b**) Homology image region feature aggregation.

The loss function used to guide the polymerization process is given by the expression:

$$L_{\mathbf{R}-\mathrm{Agg}} = \sum_{m=1}^{M} \mathrm{Agg}(\mathbf{r}_m^1, \mathbf{r}_m^2) \tag{15}$$

where $\mathbf{r}^1$ and $\mathbf{r}^2$ represent the regional feature vectors of two homologous maps. Algorithm 1 shows the pseudo-code.

Algorithm 1. Regional feature decoupling and polymerization

Input: Homology image region feature set $\mathbf{R}^1$ and $\mathbf{R}^2$. The number of attention channels M.
Output: Regional feature decoupling loss $L_{\mathbf{R}}$. Homologous map local feature aggregation loss L_{Agg}.
Initialization: $L_{\mathbf{R}} = 0$, $L_{\mathrm{Agg}} = 0$.
/* The cyclic shift operation is performed on the region feature set **R** to ensure that the region feature vectors are paired once in the decoupling process. Since the decoupling takes place within the regional feature set, $\mathbf{R}^1$ and $\mathbf{R}^2$ are no longer represented separately. */
for s in range$(M-1)$ **do**
  /* Cyclic shift s steps. */
  $\mathbf{R}' = \mathrm{CycShift}(\mathbf{R}, s)$
  /* The vectors are decoupled in pairs. */
  for m in range(M) **do**
    $L_{\mathbf{R}} = L_{\mathbf{R}} + \mathrm{Dec}(\mathbf{r}_m, \mathbf{r}'_m)$
  end
end
/* Homologous map feature aggregation. */
for m in range(M) **do**
  $L_{\mathrm{Agg}} = L_{\mathrm{Agg}} + \mathrm{Agg}(\mathbf{r}_m^1, \mathbf{r}_m^2)$
end
Return $L_{\mathbf{R}}$, L_{Agg}

Unlike regional feature decoupling, the local feature decoupling between non-homologous images is de-correlated for instance-level targets. By separating the target from its subclass cluster, we can eliminate the feature adhesion between the targets, avoid feature confusion, and reduce the difficulty of learning causal attention. As shown in Figure 5, the local features of the input image are combined in pairs for the decoupling operation; the decoupling loss is as follows:

$$L_{\mathbf{l}} = \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} \mathrm{Dec}(\mathbf{l}_i, \mathbf{l}_j) \tag{16}$$

where B indicates the input batch size. Note that the local features of the pairwise combination originate from the same input branch and the corresponding input image uses the same transformation operation. Algorithm 2 shows the pseudo-code for the local feature decoupling.

Algorithm 2. Local feature decoupling of non-homologous images

Input: Local features $\mathbf{l}^1$ and $\mathbf{l}^2$ of the two branches' input images. The batch size B of the input image.
Output: Regional feature decoupling loss $L_{\mathbf{l}}$.
Initialization: $L_{\mathbf{l}} = 0$.
/* Iterate over all local features. */
for i in range(B) **do**
  for j in range(B) **do**
    # Determine whether it is a local feature of non-homologous image. If it is, perform feature decoupling; otherwise, skip it
    if i is equal to j **then**
      continue
    else
      $L_{\mathbf{l}} = L_{\mathbf{l}} + \mathrm{Dec}(\mathbf{l}_i^1 + \mathbf{l}_j^1) + \mathrm{Dec}(\mathbf{l}_i^2 + \mathbf{l}_j^2)$
    end
  end
end
Return $L_{\mathbf{l}}$

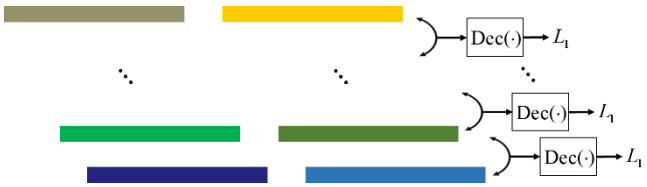


Figure 5. Local feature decoupling.

2.3. Feature Aggregation Module

In Section 2.2, based on feature separation in contrast learning, the regional feature representations of homologous maps and the local features of non-homologous maps are separated to decouple the regional features and classes. In this section, the clustering operation in contrast learning is introduced and the feature aggregation module (FAM) is used to pull the FD-CAM-separated features back into the corresponding subclass cluster.

The FAM structure is shown in Figure 2. To input the main stem feature $\mathbf{F}$ and multi-head attention diagram, a 1×1 convolution is used to reduce the dimension $\mathbf{F}$, and $\mathbf{F}' \in \mathbb{R}^{W\times H\times M}$ is determined by multiplying the feature map after dimensionality reduction with multiple attention map $\mathbf{A}$ pixel by pixel. Each channel $\mathbf{F}'$ focuses on different local areas and are weakly correlated with each other. To realize the re-association between the target's local features, we designed a feature re-association module (FRM) based on an SP, as shown in Figure 6. Different from the SPM, the FRM strips the pool of the three

dimensions of $\mathbf{F}'$ and extends the output dimensions of the three dimensions to obtain the re-association attention diagram $\mathbf{A}_r \in \mathbb{R}^{W\times H\times M}$, as follows:

$$\begin{cases} \mathbf{F}'_{\text{row}} = \text{Expand}(\text{SP}_{\text{row}}(\mathbf{F}')) \\ \mathbf{F}'_{\text{col}} = \text{Expand}(\text{SP}_{\text{col}}(\mathbf{F}')) \\ \mathbf{F}'_{\text{channel}} = \text{Expand}(\text{SP}_{\text{channel}}(\mathbf{F}')) \\ \mathbf{A}_r = \text{Conv}[\text{Conv}(\mathbf{F}'_{\text{row}})] + \text{Conv}[\text{Conv}(\mathbf{F}'_{\text{col}})] + \text{Conv}[\text{Conv}(\mathbf{F}'_{\text{channel}})] \end{cases} \tag{17}$$

where $\text{SP}_{\text{row}}(\cdot)$ indicates the average pooling of the row dimensions, $\text{SP}_{\text{col}}(\cdot)$ indicates the average pooling of the column dimensions, and $\text{SP}_{\text{channel}}(\cdot)$ indicates the average pooling of the channel dimensions. $\text{Expand}(\cdot)$ indicates a dimension replication operation. $\mathbf{F}'_{\text{row}}$, $\mathbf{F}'_{\text{col}}$, and $\mathbf{F}'_{\text{channel}}$ are the feature maps of the row dimension, column dimension, and channel dimension after strip pooling.

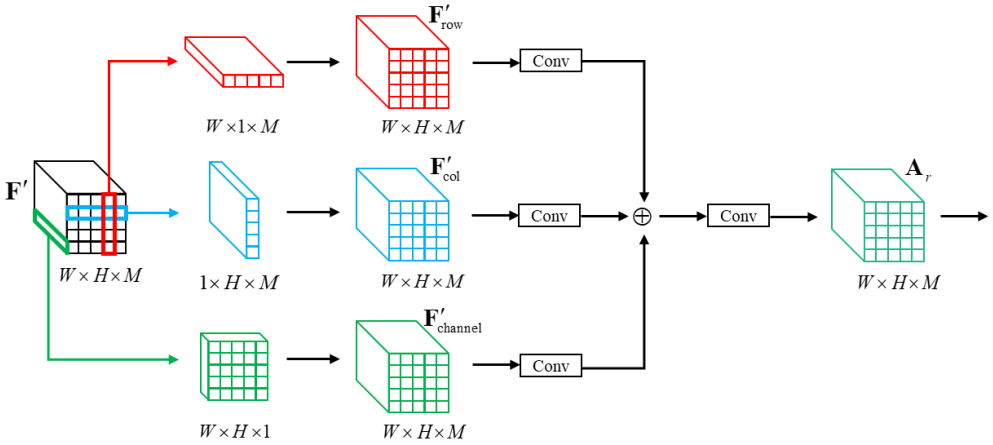


Figure 6. FRM structure.

The re-association feature map $\mathbf{F}_r$ is obtained by using the pair weighting $\mathbf{A}_r$ and $\mathbf{F}'$. Through multi-dimensional information integration, the FRM can consider both the spatial and channel domains to achieve better feature re-association. Global average pooling for $\mathbf{F}_r$ is implemented to obtain the feature vector of the target:

$$\mathbf{f} = \text{GAP}(\mathbf{F}_r) \tag{18}$$

To cluster the feature vectors, the clustering centers of various categories must be defined. We adopt the method based on agent loss to realize feature clustering by setting learnable explicit agents for each category and pulling similar feature vectors toward their corresponding proxy vectors. The principle is shown in Figure 7. Considering that a single display agent has certain limitations for fine-grained recognition tasks, this paper follows the method by Chen et al. 2020 [19], setting multiple proxy vectors as the clustering centers for various categories to reduce the clustering difficulty and improve efficiency. Specifically, let the set of proxy vectors be $\mathbf{P} = \{\mathbf{p}_1, \cdots, \mathbf{p}_n\}$, where n is the number of proxy vectors. The aggregation function is used to pull the feature vector to its proxy vector set:

$$L_{\mathbf{f}} = \sum_{i=1}^{n} \text{Agg}(\mathbf{f}, \mathbf{p}_i) \tag{19}$$

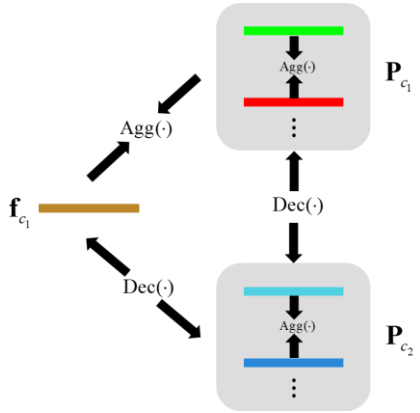


Figure 7. Principle of feature clustering.

To ensure the separability of the agents between classes and the relevance of the agents within classes, we perform separation operations on the agents of different classes and aggregation operations on the agent vectors of the same class:

$$L_{\mathbf{P}-\text{Sep}} = \sum_{i,j\in\{1,\cdots,n\}} \text{Dec}(\mathbf{p}_i^{c_1}, \mathbf{p}_j^{c_2}) \tag{20}$$

$$L_{\mathbf{P}-\text{Agg}} = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \text{Agg}(\mathbf{p}_i, \mathbf{p}_j) \tag{21}$$

where c_1 and c_2 represent two categories.

Finally, the full-connection layer is used as the classifier to complete the fine-grained classification of ship targets:

$$y_{pred} = \text{Fc}(\mathbf{f}) \tag{22}$$

where y_{pred} is the predicted result of the category.

2.4. Loss Function

The loss function in this paper mainly includes the loss of the causal attention branch, the loss of the feature aggregation branch, and the loss of classification, which are defined as follows:

$$L_{\text{CAM}} = L_{\mathbf{R}-\text{Dec}} + L_{\mathbf{R}-\text{Agg}} + L_{\mathbf{l}} + L_{\text{effect}} \tag{23}$$

$$L_{\text{FAM}} = L_{\mathbf{f}} + L_{\mathbf{P}-\text{Sep}} + L_{\mathbf{P}-\text{Agg}} \tag{24}$$

$$L_{\text{cls}} = \text{CE}(y_{pred}, y_{label}) \tag{25}$$

L_{CAM} is used to guide the feature separation and causal attention learning of the causal attention branches. L_{FAM} is used to supervise the feature re-association of the feature aggregation branches and the clustering process of the proxy vectors. L_{cls} uses cross-entropy loss to punish the final classification prediction. The total loss function is defined as:

$$L = L_{\text{CAM}} + L_{\text{FAM}} + L_{\text{cls}} \tag{26}$$

3. Experiments

This section evaluates the proposed method on a public dataset. First, we introduce the fine-grained classification datasets FGSC-23 [11] and FGSCR-42 [22] used in this paper. Secondly, we elaborate on the specific implementation details used during the experiment, including the environmental version, hardware configuration, training parameters, and dataset processing. We then discuss the ablation experiment design used to assess the

performance of each part of the model. Finally, we compare the proposed method with other fine-grained classification methods.

3.1. Datasets

3.1.1. FGSC-23

The FGSC-23 dataset was mainly collected from panchromatic remote sensing images of Google Earth and GF-2 satellites, including about 4080 sliced ship images with a resolution of 0.4–2 m and a size of 40 × 40–800 × 800 pixels. The labeling information of each sample includes a category label, aspect ratio label, and direction label. The classification task is only conducted based on the category label. All samples were divided into training and test sets using a 4:1 ratio. There were 23 categories (Figure 8), 1 non-ship target and 22 ship targets; namely, aircraft carrier (AC), destroyer (DE), landing craft (LC), frigate (FR), amphibious transport dock (ATD), cruiser (CR), Tarawa-class amphibious assault ship (TAA), amphibious assault ship (AS), command ship (CS), submarine (SU), medical ship (MS), combat boat (CB), auxiliary ship (AU), container ship (CT), car carrier (CC), hovercraft (HC), bulk carrier (BC), oil tanker (OT), fishing boat (FB), passenger ship (PS), liquefied gas ship (LG), and barge (BA).

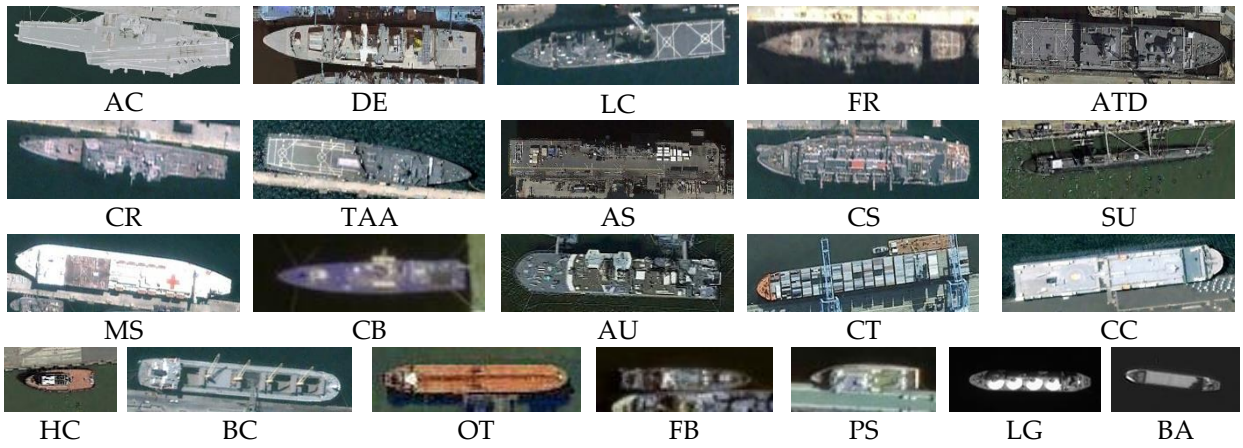


Figure 8. Samples of the FGSC-23 dataset.

3.1.2. FGSCR-42

The FGSCR-42 dataset comprised 7776 images of common types of ships, ranging from 50 × 50 to 500 × 1500 pixels and obtained from Google Earth and other remote sensing datasets, such as DOTA [23] and HRSC2016 [24]. Under 10 main categories, the ship type consisted of 42 sub-categories; namely, Nimitz class aircraft carrier (NAC), KittyHawk class aircraft carrier (KHCA), Midway class aircraft carrier (MCA), Kuznetsov class aircraft carrier (KNCA), Charles de Gaulle aircraft carrier (CGA), INS Vikramaditya aircraft carrier (IVA), INS Virrat aircraft carrier (IVKA), Ticonderoga class cruiser (TCC), Arleigh Burke class destroyer (ABCD), Akizuki class destroyer (AKCD), Asagiri class destroyer (ASCD), Kidd class destroyer (KDCD), Type 45 destroyer (T45), Wasp class assault ship (WCS), Osumi class landing ship (OCL), Hyuga class helicopter destroyer (HCHD), Lzumo class helicopter destroyer (LCHD), Whitby Island class dock landing ship (WICDL), San Antonio class transport dock (SACTD), Freedom class combat ship (FCC), Independence class combat ship (ICC), Sacramento class support ship (SCS), Crane ship (CS), Abukuma class frigate (ACF), Megayacht (MY), Cargo ship (CGS), Murasame class destroyer (MCD), Container ship (CTS), Towing vessel (TV), Civil yacht (CY), Medical ship (MS), Sand carrier (SC), Tank ship (TS), Garibaldi aircraft carrier (GAC), Zumwalt class destroyer (ZCD), Kongo class destroyer (KGCD), Horizon class destroyer (HCD), Atago class destroyer (ACD), Maestrale class frigate (MCAA), Juan Carlos I Strategic Projection Ship (JCSPS), Mistral class amphibious assault ship (MCF), and San Giorgio class transport dock (SGTD).

A sample example is shown in Figure 9. We trained and tested the model by dividing the dataset into training and test sets in a 1:1 ratio.

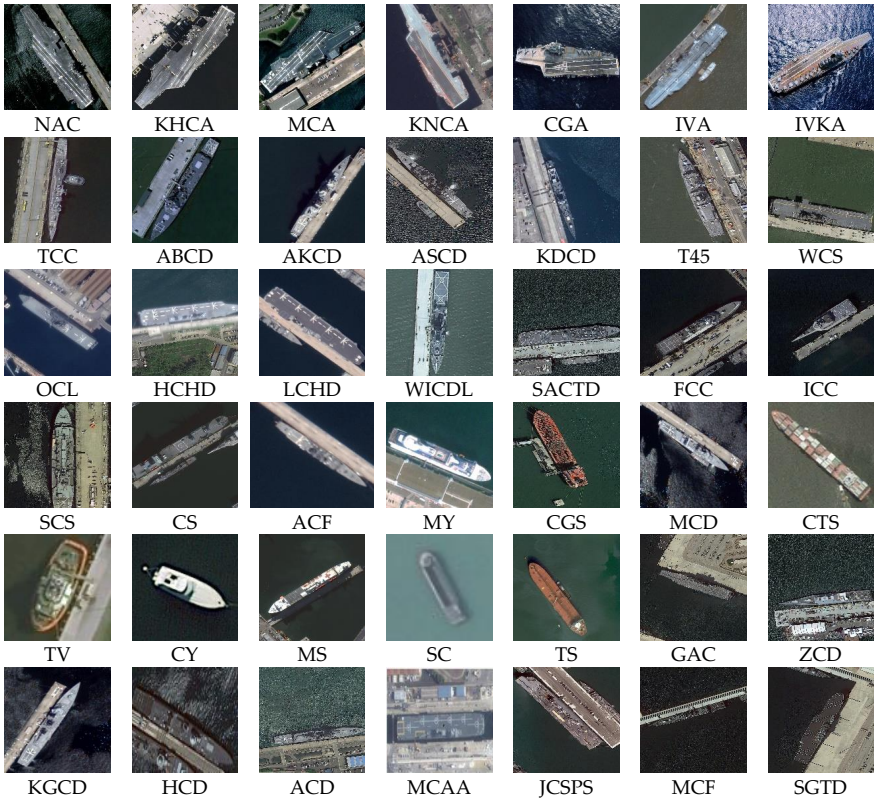


Figure 9. Samples of the FGSCR-42 dataset.

3.2. Evaluation Metrics

Overall accuracy (OA) and average accuracy (AA) were selected as the main metrics for the model's performance. Assessing the forecast accuracy from an overall perspective, OA determines the proportion of correctly predicted samples from the total samples without classifying them and is given by the expression:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{27}$$

where TP, TN, FP, and FN are the number of true positive samples, true negative samples, false positive samples, and false negative samples, respectively. The higher the OA value, the higher the prediction accuracy and the better the model's effect.

In comparison, AA focuses on the prediction of each category and evaluates the model's performance by calculating the average recall rate of all categories using the following:

$$\begin{cases} \mathrm{R}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} \\ \mathrm{AA} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{R}_c \end{cases} \tag{28}$$

where C is the number of categories and R_c is the recall of class c, which is used to represent the recognition accuracy of the class. The prediction results were then visualized through confusion matrix (CM) and 2D feature distribution to provide a more intuitive assessment of the model's performance.

3.3. Implementation Details

The experiments were conducted using the Windows 10 operating system and the development platform was Anaconda Pytorch 1.6.0 + CUDA10.1. All experiments were

performed on a laptop equipped with an Inter(R) Core(TM) i7-10875H @ 2.30GHz CPU and a NVIDIA GeForce RTX 2080 Super with Max-Q Design (8G of video memory) GPU. We used the stochastic gradient descent (SGD) [25] as the optimizer and adjusted the learning rate using the cosine annealing algorithm [26]. The model trained 200 epochs on the FGSC-23 dataset with a batch size of 16,100 epochs on the FGSCR-42 dataset and 10 epochs in the warm-up training. The backbone network was the ResNet50 [27], pre-trained on ImageNet [28]. The initial learning rate and weight decay were set to 0.001 and the momentum to 0.9. The proxy vector was initialized with Xavier Uniform [29]. To ensure that the input image was not distorted, we used the edge-filling method to scale all the image sizes to 224 × 224 pixels and carried out image enhancement operations by random horizontal flip, random vertical flip, and random rotation, among others.

3.4. Ablation Studies

We then designed a series of ablation experiments to study the effects of various parts of C2Net on classification accuracy, including the FD-CAM and FAM branches. The FD-CAM analysis focused on the causal attention structure and feature decoupling, while the FAM analysis focused on the FRM and feature aggregation. The experiment was conducted on the FGSC-23 datasets and FGSCR-42 datasets; the results are shown in Tables 1 and 2. In addition, we also explored the influence of some hyperparameters and functional forms on the model's performance.

Table 1. The ablation experiments of the different parts of the C2Net on the FGSC-23 dataset.

Backbone	FD-CAM		FAM		OA	AA
	Counterfactual Causal Attention Structure	Feature Decoupling	FRM	Feature Polymerization		
ResNet50					88.83	88.66
	√				90.78	90.47
		√			90.66	90.37
	√	√			90.90	90.63
	√		√	√	90.90	91.02
		√	√	√	91.75	91.32
			√		91.02	90.60
				√	90.53	90.24
			√	√	91.26	90.96
	√	√	√		91.02	91.39
	√	√		√	91.14	91.13
	√	√	√	√	**92.48**	**91.43**

Table 2. The ablation experiments of the different parts of C2Net on the FGSCR-42 dataset.

Backbone	FD-CAM		FAM		AA
	Counterfactual Causal Attention Structure	Feature Decoupling	FRM	Feature Polymerization	
ResNet50					92.47
	√				93.76
		√			93.07
	√	√			94.19
	√		√	√	94.48
		√	√	√	94.76
			√		94.13
				√	93.98
			√	√	94.22
	√	√	√		95.01
	√	√		√	94.94
	√	√	√	√	**95.42**

3.4.1. Effectiveness of the FD-CAM

The FD-CAM was used to determine the local attention diagram of channel decoupling. We conducted ablation experiments on the counterfactual causal attention structure and the feature decoupling operation in the FD-CAM using the control variable method to study the influence of each part on the classification performance. As shown in Tables 1 and 2, compared with the backbone network, the OA and AA, using only the counterfactual causal attention structure, increased by 1.95% and 1.81%; the OA and AA, using only the feature decoupling operation, increased by 1.83% and 1.71%, respectively. The complete FD-CAM achieved 90.90% OA and 90.63% AA on the FGSC-23 and 94.19% AA on the FGSCR-42. The results in Table 2 show that when the FAM-only classifier was used as the baseline model, the AA of the FD-CAM increased by 1.20%. The experimental results indicate that the counterfactual causal attention structure improved the learning quality of local attention by changing the model's discriminant mode, while feature decoupling eliminated feature redundancy by separating the local features; thus, improving the classification performance. In addition, the classification accuracy of the combined FD-CAM and FAM was higher compared to using only the FAM, confirming the effectiveness and feasibility of the FD-CAM.

To assess the model's effect more directly, we visualized the confusion matrix and two-dimensional feature distribution diagrams of the stochastic neighbor embedding (t-SNE) algorithm [30], as shown in Figures 10 and 11. The horizontal axis of the confusion matrix shows the prediction results, the vertical axis presents the actual results, and the diagonal refers to the accuracy of each class. From Figure 10, the confusion matrix with the most chaotic color distribution occurred when only the trunk network was used. When the FD-CAM was used, the color distribution became more concentrated in the diagonal line, and the recall rate for most kinds improved. Comparing the feature distributions in Figure 11, the FD-CAM had a higher aggregation degree than the trunk network, and C2Net had a higher aggregation degree than only the FAM. The results suggest that the local feature decoupling of the FD-CAM can reduce the mutual interference between the discriminant features of different target parts, providing high-quality local features and effectively improving classification performance.

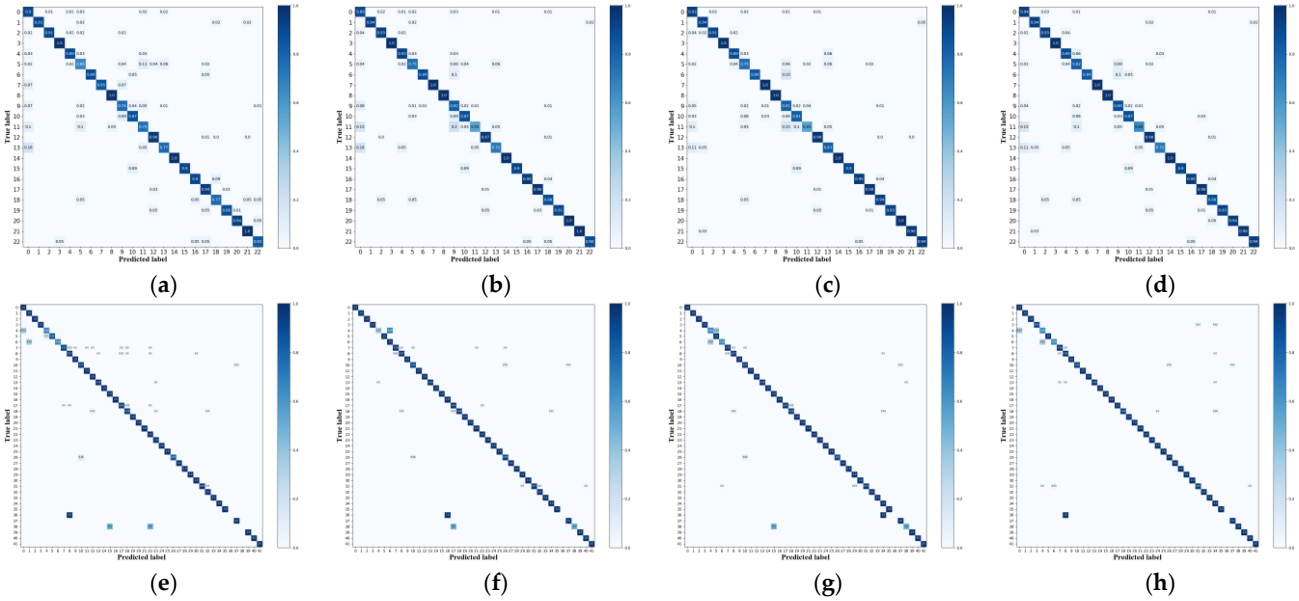


Figure 10. Confusion matrix. (**a**–**d**) Confusion matrices of the main trunk network, backbone network +FD-CAM, backbone network +FAM, and C2Net on the FGSC-23 test set, respectively; (**e**–**h**) are the confusion matrices of the FGSCR-42 test set.

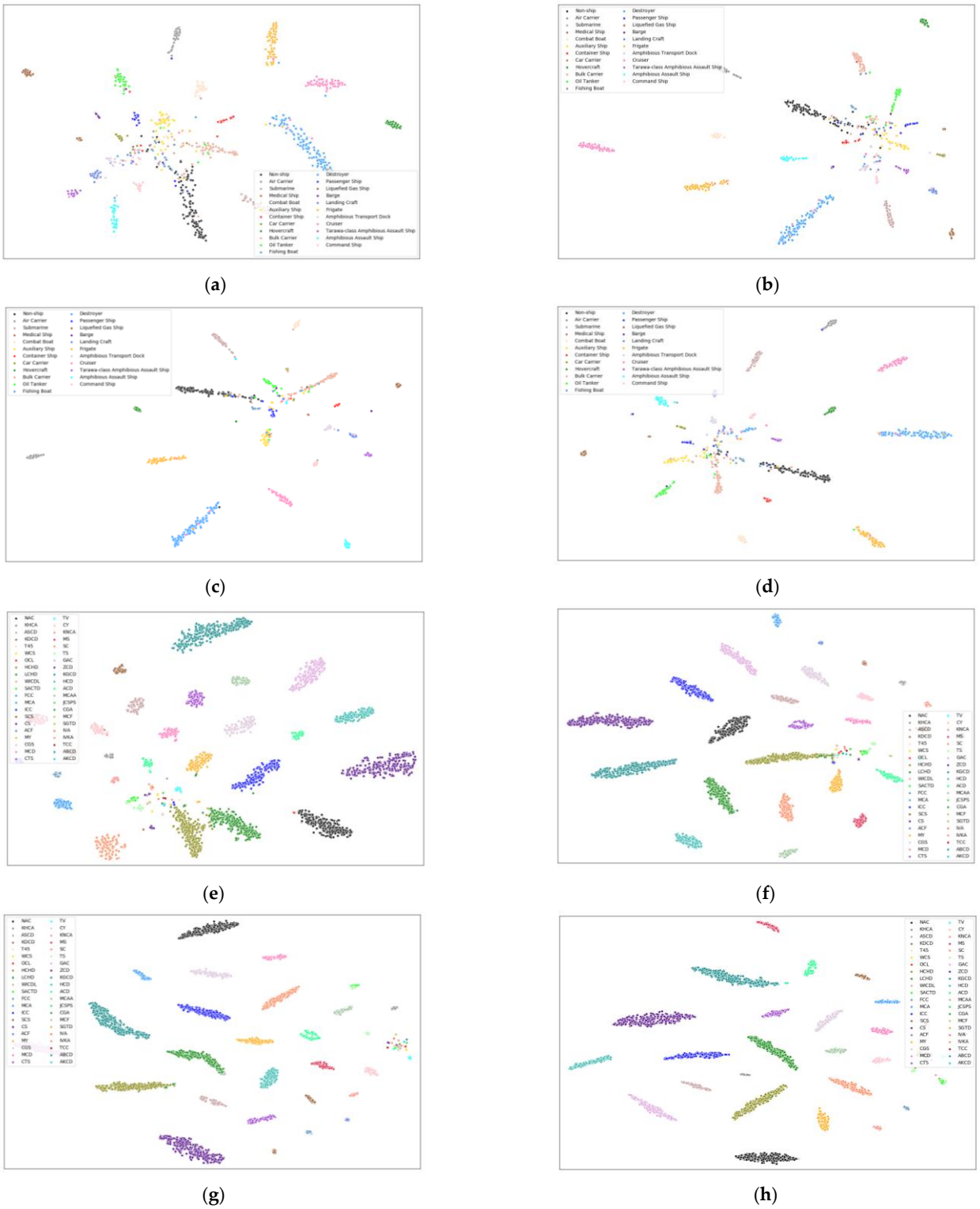


Figure 11. 2D feature distribution. (**a**–**d**) Feature distributions of the main trunk network, backbone network +FD-CAM, backbone network +FAM, and C2Net on the FGSC-23 test set, respectively; (**e**–**h**) feature distributions of the FGSCR-42 test set.

3.4.2. Effectiveness of the FAM

Ablation experiments were conducted on two FAM parts: the FRM and characteristic polymerization. In Table 1, when the baseline model was the primary trunk network, the FRM increased in OA from 88.83% to 91.02% and in AA from 88.66% to 90.60%. The

experimental results in Table 2 also show that the classification accuracy of the FRM improved after the addition. When the baseline model was only the FD-CAM, the FRM increased in AA from 94.19% to 95.01%, indicating that multi-dimensional information integration can more comprehensively establish the correlation between features.

After the FRM completes the feature association, the clustering of target features is implemented through the feature aggregation operation. Without the use of the FD-CAM, the improvement in classification accuracy by feature aggregation is limited. When combined with the FD-CAM, the feature aggregation reclusters locally decoupled features, improving the OA and AA by 2–3%. The FAM and FD-CAM, mainly composed of the FRM and feature aggregation, realize the asynchronous feature learning process together. When the FD-CAM was used alone, the OA decreased by 2.18% and the AA decreased by 0.5–1.5% compared with C2Net. As shown in Figure 11, the FAM can effectively pull the features of similar targets to the corresponding clustering center, resulting in a relatively compact feature distribution.

3.4.3. Effectiveness of Attention Channels

Multi-head attention mapping improves the model's fine-grained recognition ability by learning the detailed local features of different parts of the ship target. To explore the influence of attention density on the model's classification ability, ablation experiments were conducted on different channel numbers; the results are shown in Table 3. When the number of channels was 16, the OA reached 92.48% and the AA reached 91.43%. When the number of channels is too small, the model cannot learn enough detailed features. Conversely, if the number of channels is too large, there will be feature redundancy among the channels, resulting in considerable feature ambiguity during decoupling and affecting the learning of the model's local attention.

Table 3. The performance of attention channels M on the FGSC-23 dataset.

M	OA	AA
4	90.66	90.28
8	90.90	90.75
16	**92.48**	**91.43**
32	91.02	90.71

3.4.4. Effectiveness of the Proxy Vector Number

As the attribute agent in the target category, the display agent plays the clustering center in the feature aggregation stage. In this paper, multiple proxy vectors are set for each category as a clustering center group, which enlarges the range of the clustering center and reduces the dependence on a single central point. This can improve the effect and efficiency of clustering to a certain extent. In order to find the optimal number of agent vectors, the OA and AA of different proxy vector numbers were compared; the results are presented in Table 4. The best classification accuracy was achieved when $n = 2$, with an OA of 92.48% and an AA of 91.43%. The effectiveness of the clustering center group can be verified by comparing it with the experimental results of a single agent. However, when the number of agents is too large, the central cluster becomes loose, which is not conducive to feature aggregation. In addition, using large numbers of agents would increase the training burden, resulting in the inadequate learning of the agent vector and adversely impacting feature aggregation.

Table 4. The performance of proxy vector number n on the FGSC-23 dataset.

n	OA	AA
1	90.53	90.08
2	**92.48**	**91.43**
3	90.78	90.53
4	90.53	90.26

3.4.5. Effectiveness of the Decoupling Function and Aggregation Function

Is this what you mean: Both the decoupling and aggregation functions are compound logarithmic and cosine distance functions. Logarithmic loss is a common loss function in logistic regression tasks; its domain is a finite interval (0,1). The value range for the cosine distance is also a finite interval $(-1,1)$, so the vector regression problem can be transformed into a binary classification problem. The advantages of logarithmic function in probability distribution characterization can be used to improve feature decoupling and aggregation. To evaluate the effectiveness of the logarithmic form, a set of ablation experiments were employed to compare the influence of the two proposed compound functions and the ordinary cosine distance on the experimental results. As shown in the results in Table 5, both $\mathrm{Dec}(\cdot)$ and $\mathrm{Agg}(\cdot)$ helped to improve the model's classification. When used separately, $\mathrm{Dec}(\cdot)$ increased the OA by 0.85% and the AA by 1.61%, while $\mathrm{Agg}(\cdot)$ increased the OA by 0.25% and the AA by 0.94%. When used together, the OA increased by 1.95% and the AA increased by 1.72%, confirming the method's effectiveness.

Table 5. The performance of $\mathrm{Dec}(\cdot)$ and $\mathrm{Agg}(\cdot)$ on the FGSC-23 dataset.

Method	Dec$(\cdot)$	Agg$(\cdot)$	OA	AA
C2Net			90.53	89.71
	√		91.38	91.32
		√	90.78	90.65
	√	√	**92.48**	**91.43**

3.5. Comparisons with Other Methods

The proposed C2Net was then compared with other methods to analyze the model's fine-grained classification performance. The experiments were first conducted on the FGSC-23 dataset, calculating the recognition accuracy, OA, and AA of 23 sub-categories; the results are shown in Table 6.

Table 6. Accuracy (%) of the different methods on the testing set of the FGSC-23 dataset. The overall accuracy (OA), average accuracy (AA), and the accuracy of each category are listed.

Method	FDN [5]	DCL [31]	Inception v3 [32]	DenseNet121 [33]	B-CNN [34]	MobileNet [35]	ME-CNN [6]	PMG [36]	Xception [37]	T2FTS [38]	FBNet [39]	LGFFE [40]	C2Net (Ours)
AR_0	85.57	83.50	87.63	86.60	84.54	84.54	93.81	87.60	89.69	87.63	92.80	92.78	94.85
AR_1	88.24	85.30	91.18	82.35	91.18	88.24	91.18	83.30	88.24	91.67	94.10	94.12	94.12
AR_2	85.19	84.30	89.81	89.81	86.11	88.89	87.04	91.70	91.67	94.92	91.70	95.37	93.75
AR_3	81.82	77.30	77.27	72.73	90.91	86.36	63.64	100.00	81.82	89.83	95.50	81.82	100.00
AR_4	89.83	93.20	91.53	84.75	89.83	88.16	86.44	96.60	96.61	75.36	94.90	96.61	89.66
AR_5	77.78	77.80	83.33	77.78	83.33	77.78	77.78	68.90	88.89	91.18	77.80	83.33	82.22
AR_6	81.36	86.40	76.27	72.88	76.27	86.44	76.27	84.20	86.44	96.77	89.80	83.05	84.21
AR_7	72.22	100.00	94.44	77.78	83.33	83.33	66.67	100.00	83.33	100.00	94.40	72.22	100.00
AR_8	77.42	96.80	100.00	100.00	96.77	100.00	83.87	100.00	90.32	89.66	100.00	96.77	100.00
AR_9	66.67	88.90	55.56	77.78	77.78	77.78	83.33	72.50	94.44	82.22	94.40	83.33	88.41
AR_{10}	87.50	97.90	89.58	93.75	91.67	91.67	100.00	84.80	93.75	90.91	93.80	97.92	87.88
AR_{11}	90.00	60.00	100.00	100.00	100.00	90.00	100.00	91.20	100.00	81.82	100.00	100.00	65.00
AR_{12}	93.10	82.80	93.10	93.10	93.10	93.10	93.10	60.00	82.76	77.78	100.00	96.55	98.15
AR_{13}	77.78	60.00	68.89	75.56	73.33	75.56	82.22	83.30	77.78	100.00	68.90	77.78	72.22
AR_{14}	75.00	75.00	80.00	85.00	80.00	80.00	85.00	95.00	85.00	94.44	85.00	80.00	100.00
AR_{15}	85.71	85.70	92.86	100.00	100.00	92.86	100.00	90.90	92.86	100.00	100.00	92.86	90.91
AR_{16}	100.00	100.00	100.00	100.00	100.00	100.00	100.00	94.40	100.00	85.00	100.00	100.00	95.45
AR_{17}	73.91	87.00	76.81	75.36	72.46	75.36	78.26	100.00	76.81	92.86	78.30	85.51	98.31
AR_{18}	72.73	72.70	66.67	78.79	75.76	69.70	81.82	94.90	81.82	100.00	84.80	81.82	88.89
AR_{19}	45.00	35.00	40.00	55.00	55.00	45.00	60.00	77.80	70.00	50.00	60.00	65.00	93.22
AR_{20}	77.78	66.70	72.22	66.67	61.11	55.56	66.67	86.40	77.78	77.78	66.70	77.78	94.44
AR_{21}	100.00	95.00	90.91	100.00	100.00	100.00	100.00	88.90	100.00	95.00	100.00	100.00	96.77
AR_{22}	90.91	100.00	87.63	90.91	90.91	90.91	100.00	93.50	90.91	100.00	100.00	90.91	94.44
OA	82.30	83.60	83.88	84.00	84.00	84.24	85.58	87.20	87.76	88.73	89.30	89.45	**92.48**
AA	81.54	82.23	82.86	84.20	84.93	83.53	85.09	88.08	87.87	88.91	89.69	88.07	**91.43**
Time(ms)	7.92	-	6.89	7.05	11.17	**5.92**	-	-	9.17	-	-	6.85	6.17

Inception v3 [32], DenseNet121 [33], MobileNet [35], and Xception [37] are general CNNs and classification is carried out by extracting high-level features from images; their main limitation is that only global features are available, while detailed features are ignored. To address the problem of limited samples, FDN [5], DCL [31], and B-CNN [34] improve the

recognition accuracy of remote sensing targets through multi-feature fusion and pseudo-tag training; however, they do not consider the fusion of different receptive field features, resulting in a low utilization rate of local information. ME-CNN [6] combines the CNN, Gabor filter, LBP operator, and other means, in extracting multiple features, providing more information than the FDN. T2FTS [38] solves the long-tail recognition problem in remote sensing images using a hierarchical distillation framework. FBNet [39] uses a feature-balancing strategy to strengthen the representation of weak details and refine local features through an iterative interaction mechanism, addressing the problem of fuzzy or missing target details. PMG [36] and LGFFE [40] have been proposed for fine-grained classification tasks under weak supervision, providing more discriminant features through multi-level feature learning.

In contrast, the proposed C2Net method takes into account the adhesion redundancy among the local features, improves feature quality by asynchronous feature learning, and fully utilizes the detailed local features of targets. The experimental results show that C2Net achieved the highest accuracy, with an OA of 92.48% and an AA of 91.43%, 3.03–10.18% and 3.36–9.89% higher than the other approaches. Because we do not expand the dataset to balance the number of categories, class 11 has the smallest number in the dataset and the recognition accuracy is lower than other methods, at 65.00%. To verify the generalization of the proposed method, we conducted further tests using the FGSCR-42 dataset. As shown in Table 7, C2Net achieved the highest AA of 95.42%, higher by 2.21~5.50% compared with other methods, proving the model's strength in the fine-grained classification of ship targets.

Table 7. SOTA comparison on the FGSCR-42 dataset.

Method	ProtoTree [41]	DCL [42]	HBPNet [43]	API-Net [44]	TASN [45]	GFNet [46]	P^2Net [19]	C2Net (Ours)
AA	89.92	90.24	91.32	91.47	91.85	92.03	93.21	**95.42**

4. Conclusions

This study explored the fine-grained classification of ship objects in optical remote-sensing images and developed a comparative learning network based on causal attention. In this network, the FD-CAM and FAM are used to decouple and aggregate features in an asynchronous manner to improve the quality of local features. The FD-CAM is designed based on the counterfactual causal attention model to eliminate false associations between the results and features by strengthening the predictive logic of the model. To prevent feature adhesion, the FD-CAM uses a decoupling function to separate features and obtain a high-quality local attention weight. The partial attention weight is used as the input of the FAM to weigh the trunk, and the FRM is used to achieve feature re-association. Then, the feature clustering process is guided by proxy loss to achieve a fine-grained classification. Experimental results on two common ship datasets showed that the proposed method achieved optimal accuracy and better classification performance than other methods. Its limitation is that the number of model parameters is not lightweight enough and there is a certain gap between the practical engineering application. In future work, we plan to study the decoupling and aggregation of local features further, to improve the fine-grained classification accuracy, and on this basis, explore a lighter weight classification model.

Author Contributions: Conceptualization, C.P.; methodology, C.P.; validation, C.P.; formal analysis, C.P. and R.L.; investigation, C.P. and Q.H.; writing, original draft preparation, C.P.; writing, review and editing, R.L. and Q.H.; visualization, C.P., C.N. and W.L. (Wei Liu); supervision, R.L. and W.L. (Wanjie Lu) All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Youth Science Foundation under Grant No. 41901378.

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

References

1. Wang, N.; Li, B.; Wei, X.; Wang, Y.; Yan, H. Ship detection in spaceborne infrared image based on lightweight CNN and multisource feature cascade decision. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4324–4339. [CrossRef]
2. You, Y.; Ran, B.; Meng, G.; Li, Z.; Liu, F.; Li, Z. OPD-Net: Prow detection based on feature enhancement and improved regression model in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6121–6137. [CrossRef]
3. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017; SciTePress: Vienna, Austria, 2017; Volume 2, pp. 324–331.
4. Oliveau, Q.; Sahbi, H. Learning attribute representations for remote sensing ship category classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 2830–2840. [CrossRef]
5. Shi, Q.; Li, W.; Tao, R. 2D-DFrFT based deep network for ship classification in remote sensing imagery. In Proceedings of the 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), Beijing, China, 19–20 August 2018; IEEE: New York, NY, USA, 2018; pp. 1–5.
6. Shi, Q.; Li, W.; Tao, R.; Sun, X.; Gao, L. Ship classification based on multifeature ensemble with convolutional neural network. *Remote Sens.* **2019**, *11*, 419. [CrossRef]
7. Shi, J.; Jiang, Z.; Zhang, H. Few-shot ship classification in optical remote sensing images using nearest neighbor prototype representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3581–3590. [CrossRef]
8. Xiao, Q.; Liu, B.; Li, Z.; Ni, W.; Yang, Z.; Li, L. Progressive data augmentation method for remote sensing ship image classification based on imaging simulation system and neural style transfer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9176–9186. [CrossRef]
9. Goring, C.; Rodner, E.; Freytag, A.; Denzler, J. Nonparametric part transfer for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2489–2496.
10. Branson, S.; Van Horn, G.; Belongie, S.; Perona, P. Bird species categorization using pose normalized deep convolutional nets. *arXiv* **2014**, *preprint*. arXiv:1406.2952.
11. Zhang, X.; Lv, Y.; Yao, L.; Xiong, W.; Fu, C. A new benchmark and an attribute-guided multi-level feature representation network for fine-grained ship classification in optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1271–1285. [CrossRef]
12. Chen, Y.; Zhang, Z.; Chen, Z.; Zhang, Y.; Wang, J. Fine-Grained Classification of Optical Remote Sensing Ship Images Based on Deep Convolution Neural Network. *Remote Sens.* **2022**, *14*, 4566.
13. Lin, T.Y.; Maji, S. Improved bilinear pooling with cnns. *arXiv* **2017**, *preprint*. arXiv:1707.06772.
14. Huang, L.; Wang, F.; Zhang, Y.; Xu, Q. Fine-Grained Ship Classification by Combining CNN and Swin Transformer. *Remote Sens.* **2022**, *14*, 3087. [CrossRef]
15. Meng, H.; Tian, Y.; Ling, Y.; Li, T. Fine-grained ship recognition for complex background based on global to local and progressive learning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
16. Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*; Basic Books: New York, NY, USA, 2018.
17. Rao, Y.; Chen, G.; Lu, J.; Zhou, J. Counterfactual attention learning for fine-grained visual categorization and re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 1025–1034.
18. Xiong, W.; Xiong, Z.; Cui, Y. An Explainable Attention Network for Fine-Grained Ship Classification Using Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
19. 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*, 1–16. [CrossRef]
20. Zhan, X.; Xie, J.; Liu, Z.; Ong, Y.S.; Loy, C.C. Online deep clustering for unsupervised representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6688–6697.
21. Hou, Q.; Zhang, L.; Cheng M, M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4003–4012.
22. Di, Y.; Jiang, Z.; Zhang, H. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sens.* **2021**, *13*, 747. [CrossRef]
23. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
24. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
25. Bordes, A.; Bottou, L.; Gallinari, P. SGD-QN: Careful Quasi-Newton Stochastic Gradient Descent. *J. Mach. Learn. Res.* **2009**, *10*, 1737–1754.
26. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, *preprint*. arXiv:1608.03983.
27. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.

28. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: New York, NY, USA, 2009; pp. 248–255.
29. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.
30. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
31. Zhao, W.; Tong, T.; Wang, H.; Zhao, F.; He, Y.; Lu, H. Diversity Consistency Learning for Remote-Sensing Object Recognition with Limited Labels. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–10. [CrossRef]
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
33. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
34. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
35. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, *preprint*. arXiv:1704.04861.
36. Du, R.; Chang, D.; Bhunia, A.K.; Xie, J.; Ma, Z.; Song, Y.-Z.; Guo, J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 153–168.
37. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
38. Zhao, W.; Liu, J.; Liu, Y.; Zhao, F.; He, Y.; Lu, H. Teaching teachers first and then student: Hierarchical distillation to improve long-tailed object recognition in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
39. Zhao, W.; Tong, T.; Yao, L.; Liu, Y.; Xu, C.; He, Y.; Lu, H. Feature balance for fine-grained object classification in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
40. Lv, Y.; Zhang, X.; Xiong, W.; Cui, Y.; Cai, M. An end-to-end local-global-fusion feature extraction network for remote sensing image scene classification. *Remote Sens.* **2019**, *11*, 3006. [CrossRef]
41. Nauta, M.; Van Bree, R.; Seifert, C. Neural prototype trees for interpretable fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14933–14943.
42. Chen, Y.; Bai, Y.; Zhang, W.; Mei, T. Destruction and construction learning for fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5157–5166.
43. Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; You, X. Hierarchical bilinear pooling for fine-grained visual recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 574–589.
44. Zhuang, P.; Wang, Y.; Qiao, Y. Learning attentive pairwise interaction for fine-grained classification. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13130–13137. [CrossRef]
45. Zheng, H.; Fu, J.; Zha, Z.J.; Luo, J. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5012–5021.
46. Wang, Y.; Lv, K.; Huang, R.; Song, S.; Yang, L.; Huang, G. Glance and focus: A dynamic approach to reducing spatial redundancy in image classification. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2432–2444.

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.