Hierarchical Similarity Alignment for Domain Adaptive Ship Detection in SAR Images

Jun Zhang, Simin Li, Yongfeng Dong, Bin Pan and Zhenwei Shi

Abstract

Ship detection from Synthetic Aperture Radar (SAR) images is a hot topic, but the difficulty in collecting labeled SAR images may hinder the development of deep learning based detection methods. Inspired by the idea of domain adaptation, in this paper, we propose a Hierarchical Similarity Alignment neural Network (HSANet) for ship detection in SAR images, which is a domain adaptive approach with optical remote sensing images as training samples. The kernel target of HSANet is to mine and align both the global structure and the local instance information between SAR and optical images, where two modules, Structural Alignment Module (SAM) and Prototype Alignment Module (PAM), are designed to respectively conduct two hierarchies of alignment process. In general, SAM attempts to extract the global structure similarity which exists in image-level feature representation, while PAM tends to extract the local shape similarity which is instance-level representation. To be specific, SAM is developed by Fourier-based feature alignment, which tries to describe the similar structural relation between optical and SAR images. Meanwhile, PAM is proposed based on the conjoint confidence analysis where the instance-level ship representations of the source and target domains is aligned. SAM and PAM work together to construct a hierarchical domain adaptation network for SAR ship detection. Experiments on several public datasets may indicate the effectiveness of the proposed method.

Index Terms

SAR ship detection, remote sensing, domain adaptation.

I. Introduction

With the advantages of high resolution and all-weather imaging, synthetic aperture radar (SAR) has become an important data source for various remote sensing applications. Among the research on SAR images, ship detection is one of the hot topics. Traditional SAR ship detectors are usually developed using Constant False Alarm Ratio (CFAR) [1], where the objects are detected within a certain threshold range by distinguishing between input signal and clutter. However, CFAR based methods may suffer problems in identifying the accurate classes of objects.

In the past decade, deep learning based methods have accelerated the development of object detection task [2]–[7]. Meanwhile, designing a high-performance ship detector remains limited by such issues as inshore complex background and ship scale variation that exist widely in remote sensing images [8]–[10]. Some recent studies have attempt to solve the problems by inserting the feature fusion and attention mechanisms [11]–[15]. For example, literature [11] proposes a multi-scale feature attention module that suppresses the interference caused by complex coast surroundings. Literature [13] adopts the Path Argumentation Fusion Network to improve the fusion of different feature maps, raising the detection rate of small ships.

However, deep learning based methods are highly dependent on a large number of samples with their labels for training. Because of the characteristics of SAR imaging, SAR images are costly acquired and time-consuming for labeling. Therefore, it is a little difficult to train a supervised ship detection model using SAR images.

Recently, domain adaptation methods have been developed [16]–[18], which introduce a new manner to address the problem of lacking labeled samples. Domain adaptation methods aim at using easily-available source domain data to train a model, and make it work well in hard-to-access target domain data. These methods try to reduce the inter-domain gap so as to resolve the inconsistent distribution of source and target domains. Inspired by the idea of domain adaptation, we attempt to develop a domain adaptation based ship detection method for SAR images. In the past few years, domain adaptive methods have been investigated in remote sensing applications [19]–[24]. However, these methods mainly target at the classification and segmentation tasks, rather than object detection. Domain adaptive object detection is challenging because it has to predict both the class of objects and their locations. For natural scenes object detection, there have been some domain adaptation based methods, including reconstruct-based [25]–[27], adversarial-based [28]–[33], and feature disentangle-based [34], [35].

Compared with natural scene images, different remote sensing sources usually share similar contextual structure and object shape, but the details information varies a lot. This situation is quite obvious between SAR and optical remote sensing images. In general, the similarity among SAR and other remote sensing sources can be roughly divided into 2 parts: 1) contextual

This work was supported by the National Key R&D Program of China under the Grant 2019YFC1510905 and 2019YFC1904601, the National Natural Science Foundation of China under the Grant 62001251, and the Beijing-TianjinHebei Basic Research Cooperation Project under the Grant F2021203109. *(Corresponding author: Bin Pan.)*

Jun Zhang, Simin Li and Yongfeng Dong are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China, and also with the Hebei Province Key Laboratory of Big Data Calculation, Tianjin 300401, China (e-mail: zhangjun@scse.hebut.edu.cn.)

Bin Pan (corresponding author) are with the School of Statistics and Data Science, KLMDASR, LEBPS, and LPMC, Nankai University, Tianjin 300071, and also with the Key Laboratory of Pure Mathematics and Combinatorics, Ministry of Education, China. (e-mail: panbin@nankai.edu.cn).

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

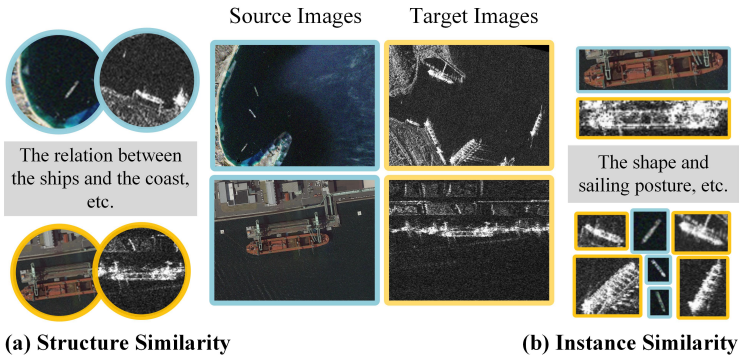


Fig. 1. Illustration of the images from different data sources. Left: Samples of the optical images. Right: Samples of the SAR images. It can be found that there exists similar characteristics between the sea surface and the coast, the shape of the ships and their sailing postures.

structure similarity which is global and image-level, and 2) ship shape similarity which is local and instance-level. Through the information extraction and abstraction of the remote sensing data in a hierarchical manner, the similarities including the position of ships on the sea surface, the structural relation between the sea surface and the coast, the shape of the ships and their sailing postures are shown in Fig 1. Therefore, it is meaningful to design specific domain adaptation approaches for ship detection in SAR images. There are several related works about domain adaptation for object detection from remote sensing images. Literature [36] uses the Unity3D engine to combine real-world and semi-synthetic data to obtain a virtual large-scale annotated dataset. Literatures [37], [38] implement the domain adaptation based on adversarial training for optical and SAR images, respectively. Literature [39] proposes a domain adaptive Faster R-CNN to train a model with a small amount of labeled SAR image data. However, these methods usually directly improve a classical domain adaptation method, which may not consider the characteristics of remote sensing scenes as shown in Fig. 1.

In this paper, we propose a Hierarchical Similarity Alignment neural Network (HSANet) for domain adaptive ship detection in SAR images, which aims at mining and aligning both the global structure and the local instance information. Similar to the literature [39], we use highly accessible optical remote sensing images as source domain data, and complete ship detection on unlabeled SAR images. HSANet consists of two modules, Structural Alignment Module (SAM) and Prototype Alignment Module (PAM), which respectively conduct two hierarchies of alignment process. SAM is developed as a Fourier-based feature alignment module, which tries to describe the similar structural relation between optical and SAR images. Meanwhile, PAM is proposed based on the conjoint confidence analysis where the instance-level ship representations of the source and target domains are aligned. Overall, SAM focuses on extracting the global structure similarity which exists in image-level feature representation, while PAM attempts to extract the local shape similarity which is instance-level representation. SAM and PAM together constitute a hierarchical domain adaptation network for SAR ship detection. The contributions in this paper can be summarized as follows.

- We propose a new hierarchical domain adaptation network, HSANet, for SAR ship detection, which takes full advantage of the structure and instance similarities between SAR and optical remote sensing images.
- We develop two aligning modules, SAM and PAM, to respectively mine the similarities from contextual structure and ship shape, where SAM generates pseudo-SAR images with Fourier transform and PAM conducts instance-level ship prototypes alignment.

II. Methodology

Suppose we have access to a source dataset which consists of n_s labeled optical image samples $\{x_i^s\}_{i=1}^{n_s}$ and corresponding labels $\{y_i^s\}_{i=1}^{n_s}$. x_i^s follows source distribution $\mathbb{P}_{\mathcal{S}}$. Similarly, the target data which consist of n_t unlabeled SAR image samples $\{x_i^t\}_{i=1}^{n_t}$ drawn from the target distribution $\mathbb{P}_{\mathcal{T}}$. The source and target domains follow different data distributions but share a common label space (ship class only). The goal is to utilize the knowledge from source domain to learn a ship detector that can perform-well in target domain images.

A. Framework Overview

The overall architecture of the proposed HSANet is shown in Fig. 2, which contains the domain adaptive (DA) base detector and two alignment modules. The DA base detector improves based on Faster R-CNN, which enables the novel detector with the benefits of flexibility and reliability. Compared with conventional two-stage methods, the DA base detector attaches the feature encoder E at the back-end of the backbone, which aims at decomposing domain-invariant features F_{di} from base features F_{base}. F_{di} keeps the same size as F_{base} and the distribution of the source and target domains on it is closer. Then, Region Proposal Network (RPN) generates class-agnostic region proposals for the input images which are mapped to F_{base} and F_{di} respectively with RoIAlign. Finally, the RoI heads predict the categories and locations of the proposals. It is worth noting that

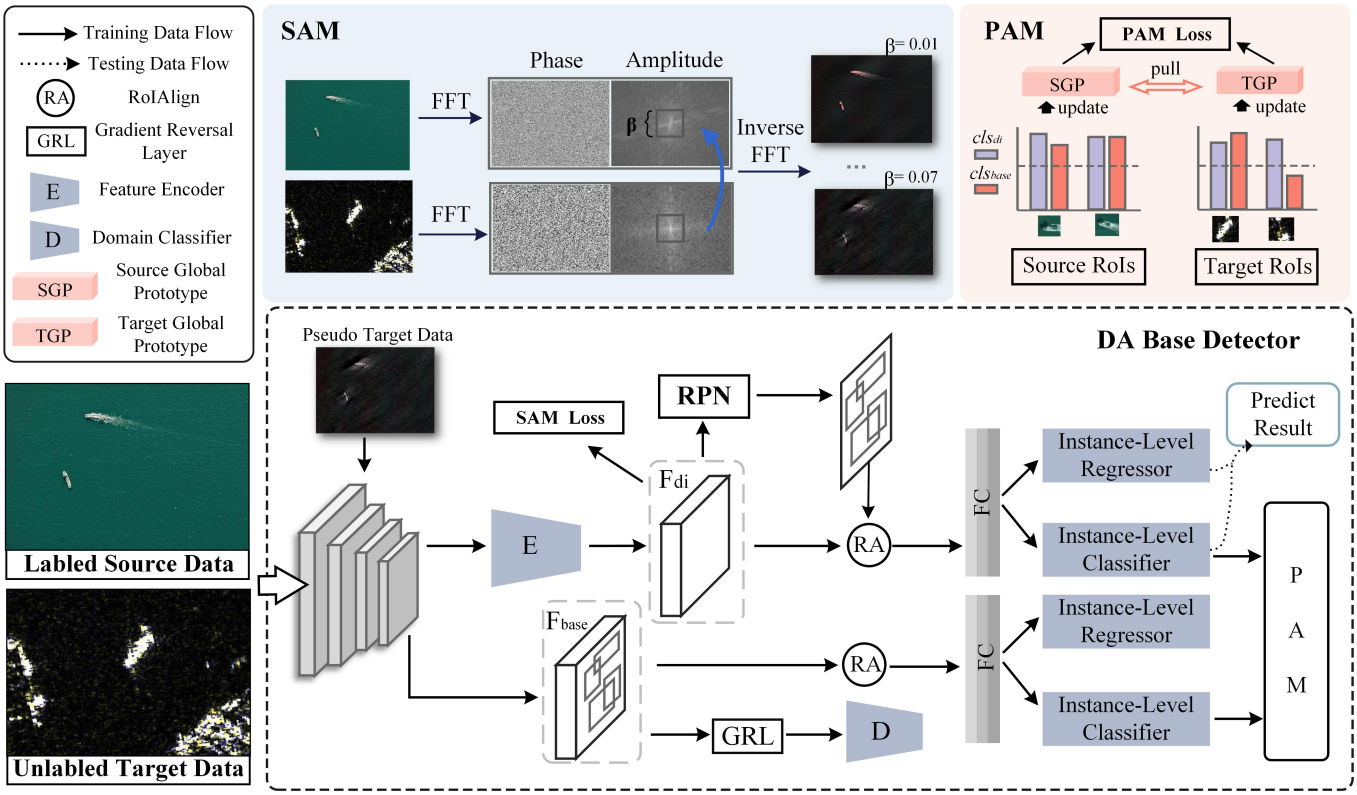


Fig. 2. An overview of our proposed HSANet. HSANet mines the structural similarity by SAM and performs instance alignment by PAM. 'FC' is the fully-connected layers. The detector has two RoI heads which connect to base features F_{base} and domain-invariant features F_{di}, respectively. The RoI heads do not share parameters.The region proposals generated by RPN are mapped to F_{base} and F_{di} respectively with RoIAlign. During the test, we only use the prediction of the RoI head connected to the domain-invariant features as the detection results.

each RoI head is connected to different features and does not share parameters. For an image x^i from source domain, the detection loss can be written as:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{RCNN}} \tag{1}$$

where $\mathcal{L}_{\text{RPN}}$ and $\mathcal{L}_{\text{RCNN}}$ are the loss functions for the RPN and the RoI heads, respectively.

Furthermore, the image-level domain classifier D aligns the base feature distributions of the source and target domains based on adversarial training. The domain classifier aims at distinguishing whether the input features come from the source or target domain, while the purpose of the backbone network F is to extract features that are sufficient to deceive it. Their training strategies are adversarial. For this reason, we utilize Gradient Reversal Layer (GRL) [40] during back propagation. The loss functions of the domain classifier D are summarized as:

$$\mathcal{L}_{\text{da}}^{s} = -\frac{1}{n_s}\sum_{i=1}^{n_s}\left(1 - D\left(F\left(x_i^s\right)\right)\right)\log\left(D\left(F\left(x_i^s\right)\right)\right) \tag{2}$$

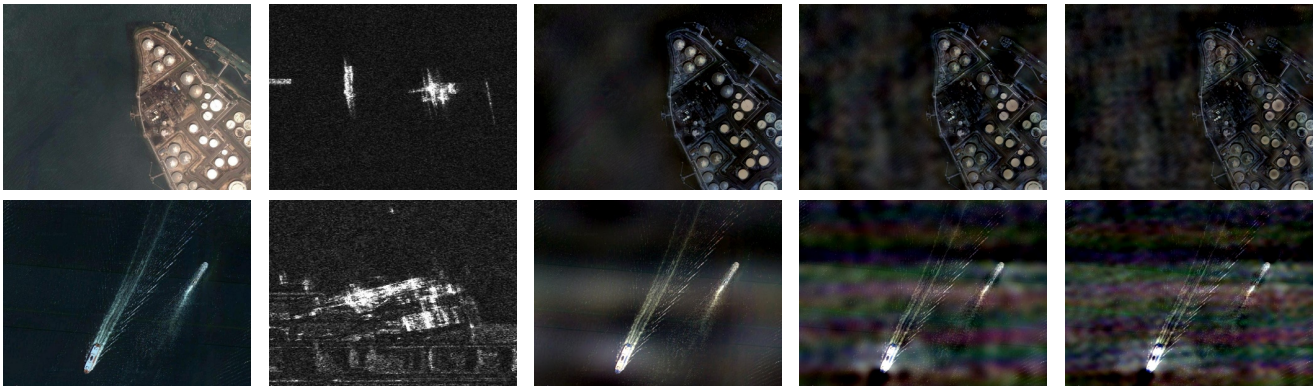
$$\mathcal{L}_{\text{da}}^{t} = -\frac{1}{n_t}\sum_{i=1}^{n_t}\left(1 - D\left(F\left(x_i^t\right)\right)\right)\log\left(D\left(F\left(x_i^t\right)\right)\right) \tag{3}$$

$$\mathcal{L}_{\text{da}} = \frac{1}{2}\left(\mathcal{L}_{\text{da}}^{s} + \mathcal{L}_{\text{da}}^{t}\right) \tag{4}$$

As mentioned in the introduction, the optical and SAR images share similar contextual structure and object shape. The ability to extract these similar features is exactly what we expect to obtain through domain adaptation. Thus, in this paper, we employ two non-adversarial alignment based modules to mine and align both the global structure and the local instance information between the source and target domains. We will describe them in Section II-B and Section II-C.

B. Structural Alignment Module

The proposed Structural Alignment Module (SAM) is designed for the similarity of the contextual structure shared by the SAR and and optical images. With the help of adversarial training, the distribution in the base feature space of the source and target domain is approached. However, the model could not understand the structural information explicitly on target domain. To this end, HSANet decomposes the domain-invariant contextual features from the base features with encoder. Then, the SAM



(a) optical image (b) SAR image (c) $\beta = 0.01$ (d) $\beta = 0.04$ (e) $\beta = 0.07$

Fig. 3. Illustration of the results of style transfer by Fourier Transform. β denotes the edge of the replaced central area of amplitude spectrum. We set $0.01 \leq \beta \leq 0.07$ and randomly tune the value of β to vary the effect of style transfer.

further aligns the domain-invariant features of a pair of optical and pseudo-SAR images, learning to extract the contextual features from the features close to the SAR image distribution.

Specifically, we generate pseudo-SAR images corresponding one-to-one with the optical images using the Fourier Transform (FFT). These images keep the same content information as the initial optical images while being presented as the appearance of the SAR images. FFT maps an image from image space to spectral signals (phase and amplitude), where the amplitude corresponds to the gray-scale variation of the image. Areas with more drastic gray-scale changes, such as the texture and edges of the image, are mostly high-frequency components of the spectral, while areas with flatter gray-scale changes, such as the background of the image, are represented as low-frequency components of the spectral. We denote the amplitude and phase components as $\mathcal{F}^A$ and $\mathcal{F}^P$. For an RGB image x, we compute it with Fourier algorithm independently for each channel [41]:

$$\mathcal{F}(x)(m,n) = \sum_{h,w} x(h,w) e^{-j2\pi}(\frac{h}{H}m + \frac{w}{W}n) \tag{5}$$

Currently, the study of [42] proposes that the low-level amplitudes can vary significantly with no effect on perceiving high-level semantics. Here we follow this research by replacing the low-level frequencies of the SAR images with the optical images before reconstituting the image via the inverse Fourier Transform. First, we define a mask M_β, whose value is zero besides the center area.

$$M_\beta(h,w) = \mathbb{I}_{(h,w)\in[-\beta H:\beta H,-\beta W:\beta W]} \tag{6}$$

where $\beta \in (0,1)$ denotes the ratio of the center area to the width and height of an image. Here, the center of the image is $(0,0)$. Then, we replace the center area of amplitude of the optical image with the corresponding region of the SAR image amplitude. Finally, a pseudo-SAR image is generated with the inverse Fourier Transform:

$$x^{s\to t} = \mathcal{F}^{-1}([M_\beta \circ \mathcal{F}^A(x^t) + (1 - M_\beta) \circ \mathcal{F}^A(x^s), \mathcal{F}^P(x^s)]) \tag{7}$$

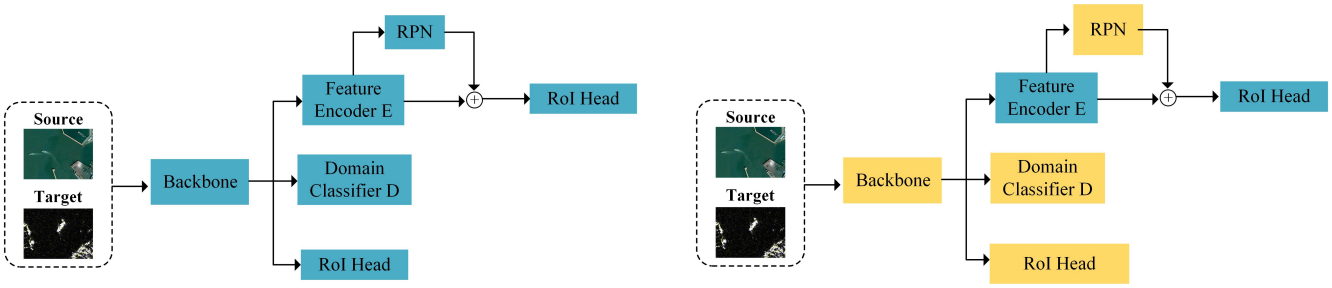
We set $0.01 \leq \beta \leq 0.07$ and randomly tune the value of β to vary the effect of style transfer. Fig. 3 shows the generated pseudo-SAR images for different β values. Note that the optical image x^s and SAR image x^t are randomly selected in each batch. Details of the steps are described in Algorithm 1. The Fourier-based feature alignment method saves the overhead of training and storing the style transfer network. The SAM aligns the domain-invariant feature distribution of the optical image x_i^s and its pseudo-SAR image $x_i^{s\to t}$. The loss function of SAM is summarized as:

$$\mathcal{L}_{\text{SAM}} = \left\| E\left(F\left(x_i^s\right)\right) - E\left(F\left(x_i^{s\to t}\right)\right) \right\| \tag{8}$$

where $\|\cdot\|$ indicates L1-norm. Minimizing the SAM loss could promote the feature encoder to learn more contextual structure features from both domains.

C. Prototype Alignment Module

Noting the similarity of ship objects in optical and SAR images, we propose Prototype Alignment Module (PAM) to align the feature distribution at local instance-level, which builds the ship prototypes of the source and target domains separately and bridges the gap between them. Traditional methods commonly compute the prototype vectors with pseudo-labels for target domain. Since the target domain images have no label information, the model is biased naturally towards the labeled source domain, leading to more false-positives. Calculating the ship prototypes using the feature vectors of false-positives will result in large deviations with the actual prototype. The proposed PAM exploits the dual detection head to select more accurate region proposals and alleviate the potential negative effects of inaccurate pseudo-label when calculating ship prototypes.



(a) First Step (b) Second Step

Fig. 4. Illustration of the two-step optimization process. In the two-stage optimization process, we update only the parameters in the blue blocks and fix the parameters in the yellow blocks.

Specifically, for an image, the j-th $(1 \leq j \leq N)$ region proposal r_j is mapped to the base features and domain-invariant features, and receives its foreground confidence scores cls^j_{base} and cls^j_{di} from the RoI heads respectively. N indicates the number of proposals. By combining the base features and domain-invariant features, we can obtain the foreground confidence of a region from different perspectives. Only if both RoI heads determine the box belongs to the foreground, it will be added to the calculation for the prototypes. We filter out the less reliable boxes based on the conjoint confidence analysis and build the more accurate ship prototypes. Furthermore, we re-weight each foreground region according to its confidence scores. Formally, the local ship prototypes P in this batch can be defined as:

$$w_j = \begin{cases} \frac{1}{2}(cls^j_{\text{base}} + cls^j_{\text{di}}) & cls^j_{\text{base}} \geq 0.5\ ,\ cls^j_{\text{di}} \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

$$P = \frac{\sum_{j=1}^{N} w_j f_j}{\sum_{j=1}^{N} w_j} \tag{10}$$

Here, we suppose f_j denotes the feature vector of r_j outputs from the second fully-connected (FC) layer in the RoI head connected to the domain-invariant features. w_j indicates the average confidence of r_j. The weights of the proposals determined to be false-positives will be set to zero. Since ships in the images are variable in size and rotation angle, the representation of ship instances also changes. The prototype obtained from a mini-batch can not properly summarize the entire category. Therefore, we maintain the global ship prototypes GP at each mini-batch dynamically to make the prototypes more representative.

$$\alpha = \frac{1}{2}(1 + \frac{(P^{(i)})^{\mathrm{T}} \cdot (GP^{(i-1)})}{\left\|P^{(i)}\right\| \left\|GP^{(i-1)}\right\|}) \tag{11}$$

$$GP^{(i)} = \alpha P^{(i)} + (1-\alpha) GP^{(i-1)} \tag{12}$$

where α denotes the cosine similarity between $P^{(i)}$ and $GP^{(i-1)}$. $P^{(i)}$ and $GP^{(i)}$ represent the local prototypes and the global prototypes at i-th iteration, respectively. HSANet aligns the instance-level ship feature representations by minimizing the Euclidean distance between the source global prototypes $GP^{S(i)}$ and target global prototypes $GP^{T(i)}$. The loss function of PAM is summarized as:

$$\mathcal{L}_{\text{PAM}} = \left\| GP^{S(i)} - GP^{T(i)} \right\|_2^2 \tag{13}$$

D. Network Optimization

For completing ship detection on unlabeled SAR data, we conduct two hierarchies of alignment process at both image-level and instance-level. Including the proposed novel modules and the supervised detection loss on the source domain, the overall loss can be described as:

$$\mathcal{L}_1 = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{da}} + \lambda_1 \mathcal{L}_{\text{SAM}} + \lambda_2 \mathcal{L}_{\text{PAM}} \tag{14}$$

where λ is a parameter to balance each module. In particular, $\mathcal{L}_{\text{RCNN}}$ takes the average of the loss of two RoI heads. We use $\mathcal{L}_1$ to optimize the entire model, as shown in the left part of Fig. 4.

In addition, according to the theory of learning disentangled representations [43], to keep the decomposed components independent, we continue the detached optimization strategy in [35]. In each mini-batch, after optimizing the entire model with $\mathcal{L}_1$, we update the parameters of the feature encoders using the orthogonal loss. The process is shown in the right part of Fig. 4. Concretely, we take the difference between the base features and the domain-invariant features extracted by the encoder as domain-specific features. This approach enables the decomposed components to contain all the input information without reconstruction loss, saving parameter storage and computational costs. Simultaneously, to keep the disentangle components independent, we trained the encoder with orthogonal loss based on vector decomposition.

For an input image, RPN generates several object proposals, which are mapped to the base features and domain-specific features respectively after RoIAlign. Then, we obtain the the outputs $P_{\mathrm{di}} \in \mathbb{R}^{n\times c}$ and $P_{\mathrm{ds}} \in \mathbb{R}^{n\times c}$ of global average pooling of the RoIAlign results, where n and c indicate the number of proposals and the number of channels. The process of orthogonal loss is as follows:

$$M = (\|P_{\mathrm{di}}\|_2^2) \odot (\|P_{\mathrm{ds}}\|_2^2) \tag{15}$$

$$\mathcal{L}_{\perp} = \frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{c} M[i,j]\right| \tag{16}$$

where $\odot$ and $|\cdot|$ separately indicate element-wise product and the absolute value operation, and $M[i,j]$ is the value of $M \in \mathbb{R}^{n\times c}$ at the position (i,j). Finally, $\mathcal{L}_2$ can be defined as follows:

$$\mathcal{L}_2 = \mathcal{L}_{\mathrm{det}} + \mathcal{L}_{\perp} \tag{17}$$

It is worth noting that we update the encoder and RoI classifier by $\mathcal{L}_2$, while the other parameters are fixed. After the second training step, the decomposed features will be kept independent with no effect on the detection performance of the model. The training details are shown in Algorithm 1.

Algorithm 1 HSANet Algorithm.

Input:
source images $\{X_s, Y_s, B_s\}$; target images $\{X_t\}$;
backbone F; feature encoder E.
Output:
An adaptive detector.
Calculate the initial global prototypes $GP^{S(0)}$ and $GP^{T(0)}$ using the pretrained detector.
while not converged **do**
  Sample a mini-batch from $\{X_s, Y_s, B_s\}$ and $\{X_t\}$;
  First Step:
  Compute $\mathcal{L}_{\mathrm{det}}$ and $\mathcal{L}_{\mathrm{da}}$ according to Eq. (1) - Eq. (4);
  Generate a pseudo-SAR image according to Eq. (7);
  Calculate P according to Eq. (10);
  Update GP^S and GP^T according to Eq. (12);
  Compute $\mathcal{L}_{\mathrm{SAM}}$ and $\mathcal{L}_{\mathrm{PAM}}$ according to Eq.(8) and Eq.(13);
  Compute $\mathcal{L}_1$ according to Eq. (14);
  Update the whole detection model by $\mathcal{L}_1$;
  Second Step:
  Compute $\mathcal{L}_2$ according to Eq. (15) - Eq. (19);
  Update E, Classifier and Regressor by $\mathcal{L}_2$;
end while

III. Experiments

In this section, we first introduce a concise description of the datasets, evaluation metrics and implementation details. Furthermore, to validate the proposed HSANet, the results of extensive experiments and visualization analysis on several tasks are presented. Finally, we make a discussion on the failure detection. Table I shows the detailed information of the adopted datasets.

A. Datasets

1) LEVIR: The LEVIR contains a large number of 800*600 resolution Google Earth images with a spatial resolution of 0.2-1.0 m [44]. To implement domain adaptive ship detection, we selecte images containing complete ships and ignore the rest of the categories on LEVIR (aircraft and oil pipeline). Finally, we take 805 samples from LEVIR, containing a total of 1527 ship objects. During the training process, we use all samples as a training set.

2) SSDD: SAR Ship Detection Dataset (SSDD) is a publicly available SAR ship dataset produced by [45]. SSDD contains 1160 images and 2456 ships in total, constructing the dataset following a procedure similar to PASCAL VOC. The variable conditions of the environment of the dataset such as sea state, resolution, image size, and sensors make it difficult for detectors to obtain high performance on it. We split the dataset into a training set, a test set and a validation set in the proportion of 6:2:2.

TABLE I
EXPERIMENTAL DATA

No.	Dataset	Data Source	Sensor	Size	Resolution	Images	Objects
1	LEVIR (Ship-only)	Optical	Google Earth	800×600	0.2m∼1.0m	805	1527
2	HRRSD (Ship-only)	Optical	Google Earth Baidu Map	1k×1k∼2k×2k	0.15m∼1.2m	2165	3975
3	SSDD	SAR	RadarSat-2 TerraSAR-X Sentinel-1	300×300∼500×500	1m∼15m	1160	2456
4	SAR-Ship-Dataset	SAR	GF-3 Sentinel-1	256×256	1.7m∼25m	43819	59563

TABLE II
RESULTS OF SHIP DETECTION ADAPTING ON SSDD FROM TWO OPTICAL IMAGE DATASET (%). NOTE THAT THE BACKBONE OF THESE METHODS ARE RES-101.

Method	Dataset	
	LEVIR→SSDD	**HRRSD→SSDD**
Faster R-CNN	30.2	28.4
DA Faster	47.8	46.2
SW Faster	53.1	54.3
DA-CR	53.6	52.6
SW-CR	54.9	54.4
HTCN	54.4	51.2
SW-VDD	52.6	56.6
SCL	55.2	55.5
Ours	**58.2**	**58.1**

3) HRRSD: High-Resolution Remote Sensing Detection (HRRSD) is a dataset released in 2019 by the University of Chinese Academy of Sciences. HRRSD contains thousands of images from Google Earth and Baidu Maps, with a spatial resolution size of 0.15m-1.2m. HRRSD is a multi-category dataset and we select 2057 images, which contains 3897 ship objects. As with LEVIR, all samples are used for training.

4) SAR-Ship-Dataset: SAR-Ship-Dataset (SAR) is a publicly available SAR image ship detection dataset released by Chinese Academy of Sciences [46]. This dataset is constructed using Chinese Gaofen-3 SAR data as the main data source to build a high-resolution SAR ship deep learning sample bank. The SAR-Ship-Dataset contains 43819 ship chips which have diverse backgrounds and is suitable for a wide range of SAR image applications. Next experiments select 2000 samples as the test set, 2000 as the validation set, and the remaining samples are used for training.

B. Evaluation Metrics

Similar with the object detection task, we use the average precision (AP) to evaluate the performance of our proposed network. By setting the Intersection over Union (IoU) threshold, the detection results can be divided into true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The formulas of accuracy, recall, and AP are as follows.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{18}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

$$AP = \int_0^1 P(r)dr \tag{20}$$

The AP for each category is the average of the precision in the interval from recall 0 to recall 1. We use AP as the evaluation metrics in the following experiments.

C. Implementation Details

We use ResNet-101 pretrained with ImageNet as the backbone network, and in all experiments, each input image is adjusted to make the length of its short edge 600 pixels. Each mini-batch contains two images, from the source and target domains, respectively. The model weights are updated by the SGD optimizer with the weight decay of 0.0005 and the momentum of 0.9. We finetune the network with a learning rate of 0.001 for 10 epoch and then reduce the learning rate to 0.0001 for another 10 epoch. In addition, we empirically set λ_1 to 0.1 and λ_2 to 0.01. We report the AP with the IoU threshold of 0.5 during evaluation. All experiments are done based on Pytorch.

TABLE III
RESULTS OF SHIP DETECTION ADAPTING ON SAR FROM LEVIR (%)

Method	Dataset
	LEVIR→SAR
Faster R-CNN	32.4
DA Faster	43.0
DA-CR	47.2
HTCN	50.9
SW-VDD	50.7
Ours w/o SAM	**53.2**
Ours w/o PAM	**54.0**
Ours	**54.6**

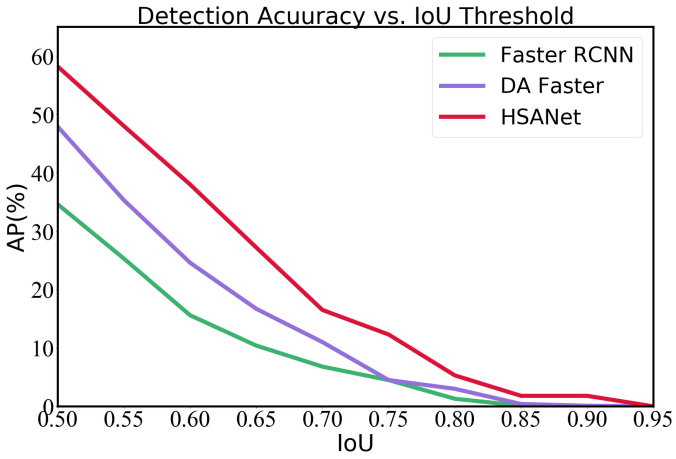


Fig. 5. The performance with the different IoU thresholds on adaptation task LEVIR→SSDD.

D. Experimental Results

To validate the effectiveness of the proposed HSANet, we perform contrast experiments under three groups of adaptation datasets. We compare the HSANet with several SOTA cross-domain object detection algorithms: DA [28], SW [29], CR [30], HTCN [31], VDD [35] and SCL [32]. ResNet-101 is used as the backbone of all networks involved in the comparison. For fairness, we ensure the consistent division of the datasets, and the evaluation results are shown in Tables II and III.

Table II shows the results of adaptation from two optical image datasets, LEVIR and HRRSD, respectively, to SSDD. 'Faster R-CNN' indicates the model is only trained by source domain images and tested directly on the test set of target domain. Column 2 of Table II shows the domain adaptation results from LEVIR to SSDD, and column 3 shows the domain adaptation results from HRRSD to SSDD. Our method outperforms the highest SOTA method by 3.0-10.4% on both datasets. All these results demonstrate that our model has a more significant advantage.

The adaptation results of LEVIR→SAR are shown in Table III. HSANet outperforms all comparison methods and achieves 3.7% AP improvement. This proves the effectiveness of extracting the similar features hierarchically of the optical and SAR images.

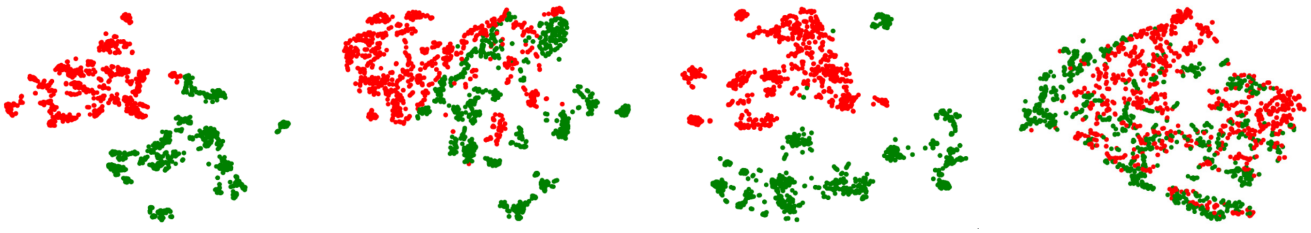
E. Futher Analysis

1) Ablation Study: In this section, we conduct a series of ablation studies on adaptation to test the effectiveness of each component. To ensure the fairness of the experiments, we keep the same hyperparameters during the validation and use ResNet-101 as the backbone network for all networks.

As shown in Table IV, the first network adopts Faster R-CNN as a baseline, which is trained by LEVIR only and tested on the test set of SSDD. Since there has no any domain adaptation module, the model achieves only 32.0% AP on the target domain. Then, we use an image-level domain classifier to roughly align the base feature distributions between the source and target domains and the experimental result is improved by 8.2%. To further extract domain-invariant features from the base features, we insert a feature encoder and the two-stage optimization strategy from [35], which make AP reaches 46.9%. This proves that filtering out domain-specific features that do not need to be aligned could improve the domain adaptation performance, and we use it as DA base detector. After aligning domain-invariant features with SAM, the results improved by 9.4%. Fig. 7 shows the visualization of base features and domain-invariant features. We find that the features extracted from the encoder have stronger instance-level information, *i.e.,* the features of the ships are clearer and the features of the coast are

TABLE IV
ANALYSIS OF HSANET IN ABLATION STUDIES(%)

Faster R-CNN	Domain Classifier	Feature Encoder	SAM	PAM	AP(%)
✓					32.0
✓	✓				40.2
✓	✓	✓			46.9
✓	✓	✓	✓		56.3
✓	✓	✓		✓	54.2
✓	✓	✓	✓	✓	**58.2**



(a) Source-only (b) SW-Faster (c) Ours (Base Feature) (d) Ours (DI Feature)

Fig. 6. Feature visualization results from LEVIR to SSDD by *t*-SNE [47]. Red indicates the samples of source domain and green is the target one. (a) and (b) represent the feature distribution results of the model trained by source images only and the SW-faster. (c) and (d) indicate the base features output from backbone and domain-invariant features output from encoder E in our model, respectively.

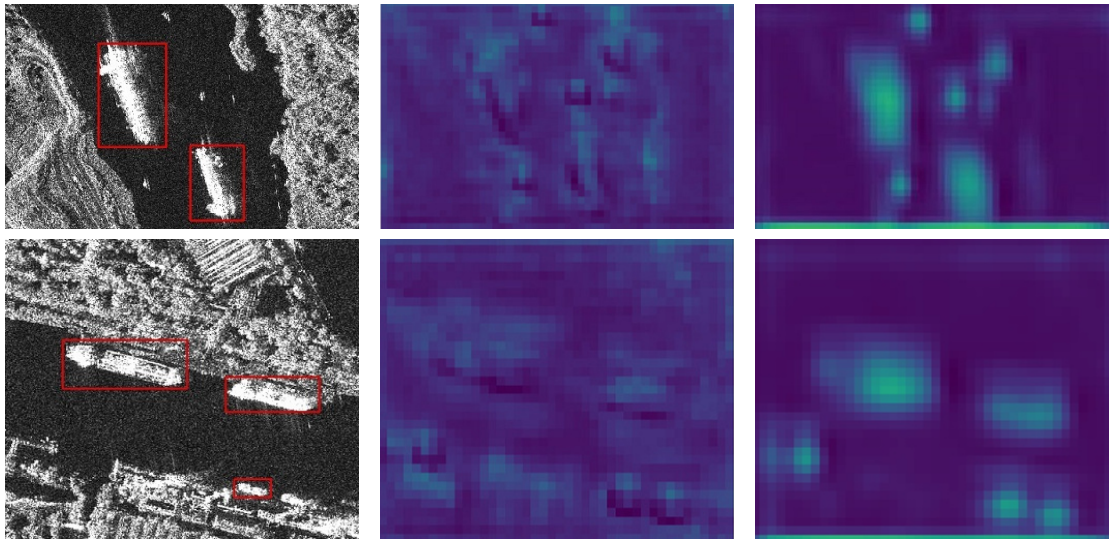
weakened. Furthermore, (c) and (d) in Fig. 6 show that the domain-invariant features output from the encoder trained by SAM are better aligned compared to the base features. For the instance-level, we add the proposed PAM to the DA base detector, and the results improved by 7.3%. These experimental results prove the effectiveness of the proposed modules. Finally, we combine SAM and PAM with the DA base detector, and the AP achieves 58.2%.

Similarly, we make an ablation analysis on SAR by evaluating variants of HSANet. As shown in Table III, all the proposed modules have a positive impact for the model. And it achieves the highest AP when both SAM and PAM were added to the DA base detector. *2) Influence of IoU threshold:* For the object detection task, IoU is the ratio of the intersection and the concatenation of the predicted boxes and ground-truth. A larger IoU means that the two boxes have more overlapping area. When the IoU value of a predicted box and the ground-truth is greater than the set threshold, the box is treated as positive. Hence the setting of the IoU threshold affects the quality of the detection. To explore the effect of different IoU thresholds for the experimental results, we conduct experiments on the LEVIR→SSDD task. The performance of our approach and other models (*i.e.,* Source Only, DA-Faster R-CNN) is shown in Fig. 5. We find the AP decreases dramatically as the IoU threshold increases and is close to zero in the end. Meanwhile, the AP of our method is greater or equal to other methods at different thresholds. *3) Visualization Experiment:* To observe the effect of feature alignment more accurately, we select the same amount of optical images and SAR images and visualize the image-level features using *t*-SNE [47]. The features are extracted with global average pooling on the output of the backbone and the encoder. As shown in Fig. 6, the feature distribution between source and target domains obtained by Source-only model displays a notable domain shift. After training with SAM, the features output from the encoder in our model are better aligned compared with SW-Faster.

In addition, Figs. 8 and 9 show the examples of detection results on adaptation task LEVIR→SSDD and LEVIR→SAR. Compared with Source-only and DA-Faster, our method detects more ship objects and predicts more accurate location information. *4) Analysis of Failed Detection:* The above experiments demonstrate the effectiveness of our proposed model. It can be observed from Fig. 7 that although the ship features are strengthened in the domain-invariant features, part of the disturbance instances is also wrongly enhanced. Fig. 10 shows some error-detected samples. We find two types of errors, *i.e.,* (1) wrong detection of ships close to the coast (as shown in Fig. 10 (a)). (2) no-detection of undersized ships (as shown in Fig. 10 (b)). We assume two reasons for these errors. Firstly, the features of near-shore ships are easily confused with the features of coast due to the lack of effective color information. Secondly, the mismatch of instance sizes in the source and target domains makes it difficult to identify small targets in SAR images. The way to solve these two problems will be the next direction of our research.

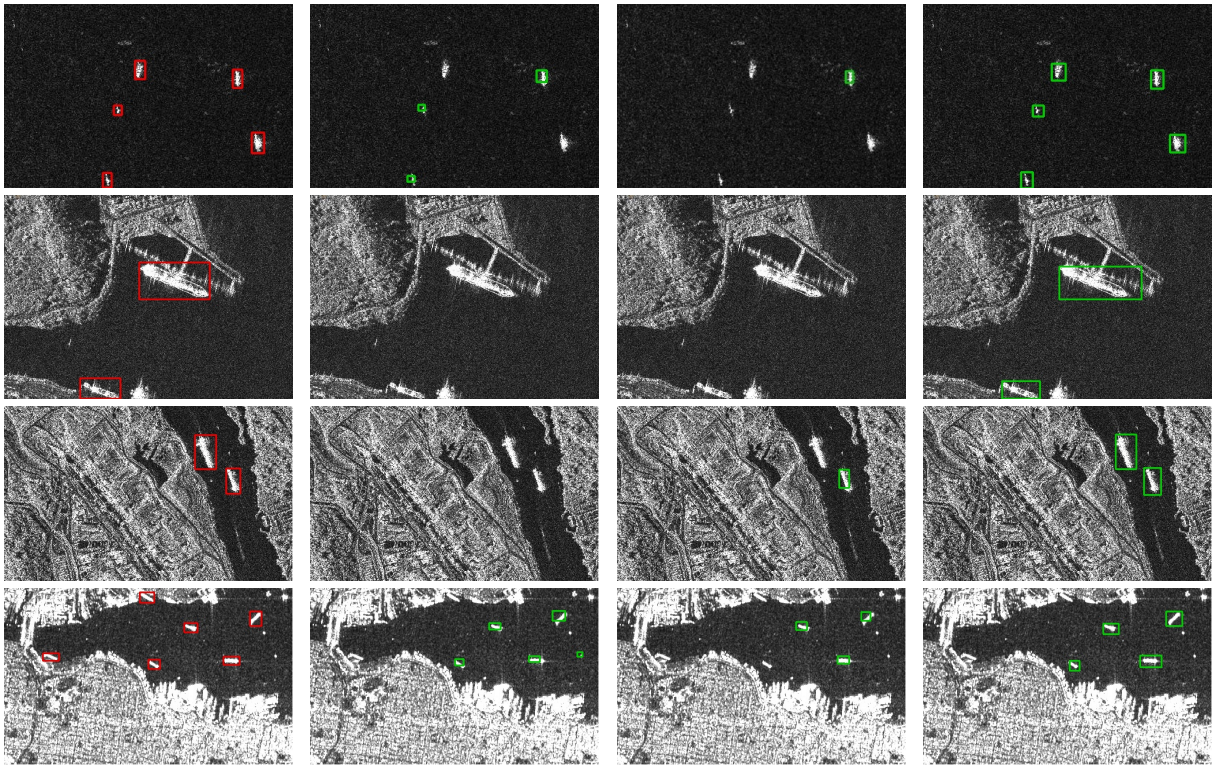
IV. CONCLUSION

In summary, we propose HSANet, a ship detection model in SAR images based on domain adaptation, to address the difficulty in training models due to unlabeled SAR data. Exploiting the similar contextual structures and ship shapes in optical and SAR images, we conduct two hierarchies of alignment process with SAM and PAM respectively. The SAM generates pseudo SAR images with Fourier Transform to describe the similar structural relation which exists in image-level feature representation. Then, the PAM builds the more accurate ship prototypes by the conjoint confidence analysis and aligns the



(a) GT (b) Base Feature (c) DI Feature

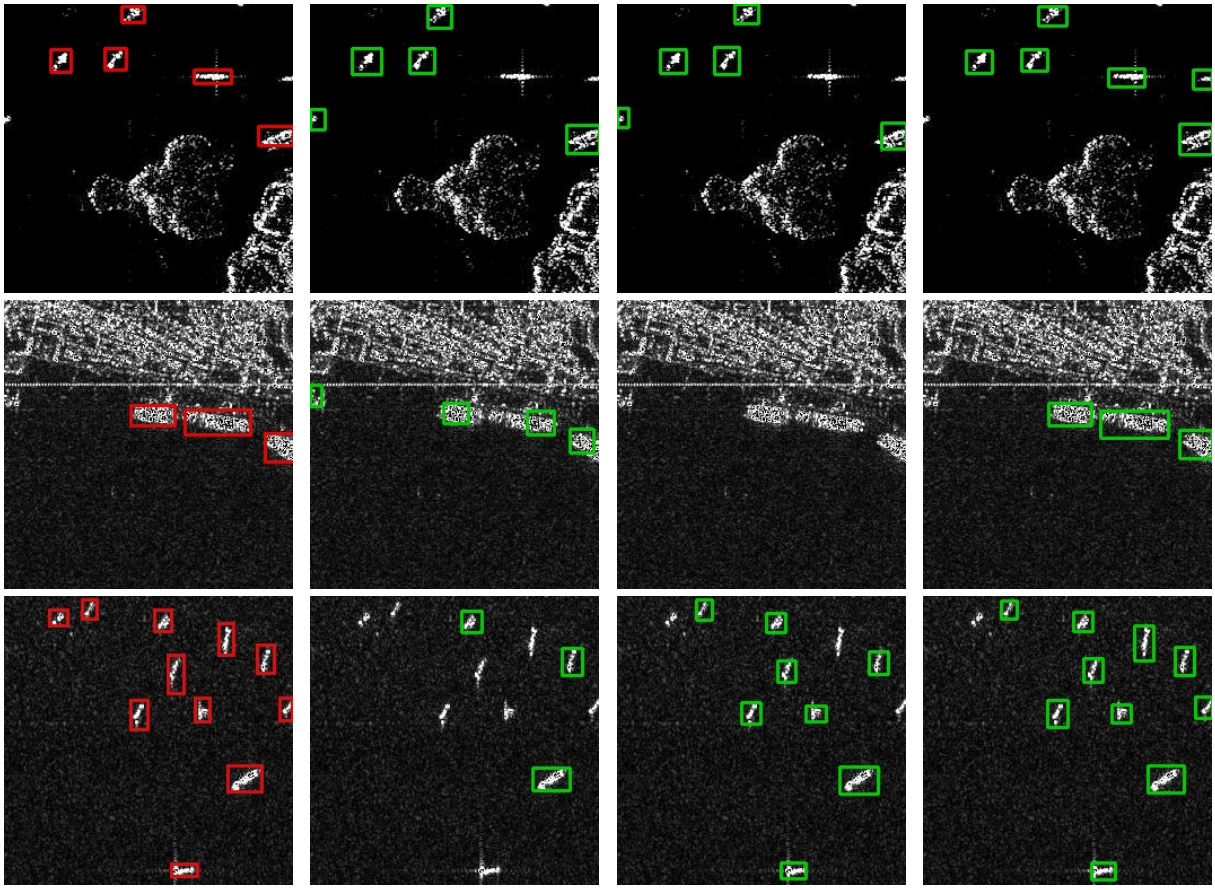
Fig. 7. Visualization of feature maps based on SSDD. 'GT' means the ground-truth, 'Base feature' and 'DI feature' represent the feature visualization results from the backbone and the encoder E, respectively.



(a) Ground Truth (b) Source Only (c) DA Faster R-CNN (d) HSANet (ours)

Fig. 8. Illustration of the detection results on SSDD. (a)-(d) indicate the ground truth and the detection results of the source-only model, the DA Faster R-CNN, and the proposed HSANet respectively.

instance-level feature representations of the source and target domains. The experimental results on several datasets and the visualization analysis show that our model has advantages compared with other existing methods.



(a) Ground Truth (b) Source Only (c) DA Faster R-CNN (d) HSANet (ours)

Fig. 9. Illustration of the detection results on SAR. Same indication as Fig. 8 for (a)-(d). Compared with Source-only and DA-Faster, our method detects more ship objects and predicts more accurate location information.

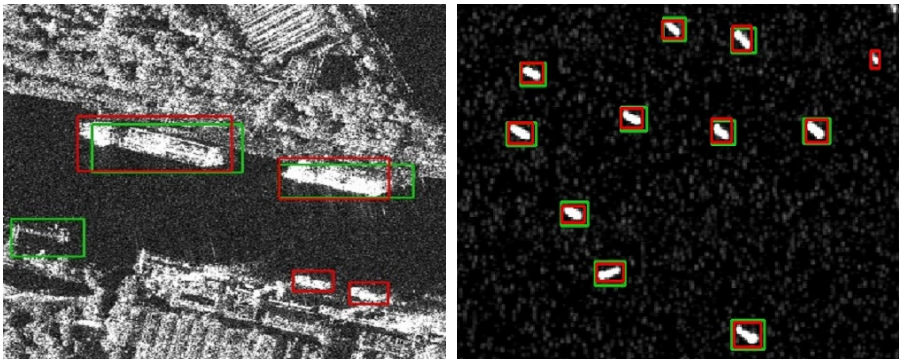


Fig. 10. Analysis of false detection results. Ground-truth and predicted ships are colored red and green, respectively. From this figure, we could find two kinds of errors: (1) Wrong detection of ships close to the coast. (2) No-detection of undersized ships.

REFERENCES

[1] J. Hu, G.-S. Xia, and H. Sun, "Target detection in SAR images via radiometric multi-resolution analysis," in *MIPPR 2013: Automatic Target Recognition and Navigation*, vol. 8918. SPIE, 2013, pp. 16–22.

[2] K. Zhang, Y. Wu, J. Wang, and Q. Wang, "A hierarchical context embedding network for object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[3] G. Guo, L. Fang, and J. Yue, "Oriented spatial correlative aligned feature for remote sensing object detection," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 5319–5322.

[4] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 310–314, 2018.

[5] X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[6] K. Zhang, Y. Wu, J. Wang, Y. Wang, and Q. Wang, "Semantic context-aware network for multiscale object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2021.

[7] J. Ding, J. Wang, W. Yang, and G.-S. Xia, "Object Detection in Remote Sensing," *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, pp. 67–89, 2021.

[8] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2738–2756, 2020.

[9] Q. Li, R. Min, Z. Cui, Y. Pi, and Z. Xu, "Multiscale ship detection based on dense attention pyramid network in SAR images," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019.

[10] C. Wang, W. Su, and H. Gu, "Two-stage ship detection in synthetic aperture radar images based on attention mechanism and extended pooling," *Journal of Applied Remote Sensing*, vol. 14, no. 4, p. 044522, 2020.

[11] Y. Du, L. Du, and L. Li, "An SAR target detector based on gradient harmonized mechanism and attention mechanism," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2021.

[12] J. Zhang, C. Xie, X. Xu, Z. Shi, and B. Pan, "A contextual bidirectional enhancement method for remote sensing image object detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4518–4531, 2020.

[13] J. Wang, Y. Lin, J. Guo, and L. Zhuang, "SSS-YOLO: Towards more accurate detection for small ships in sar image," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 2, pp. 93–102, 2021.

[14] C. Zhu, D. Zhao, Z. Liu, and Y. Mao, "Hierarchical attention for ship detection in SAR images," in *IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 2145–2148.

[15] L. Shi, L. Kuang, X. Xu, B. Pan, and Z. Shi, "CANet: Centerness-aware network for object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2021.

[16] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and K. Keutzer, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 473–493, 2022.

[17] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[18] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[19] X. Zhang, X. Yao, X. Feng, G. Cheng, and J. Han, "DFENet for domain adaptation-based remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.

[20] L. Wu, M. Lu, and L. Fang, "Deep covariance alignment for domain adaptive remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[21] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 6180–6195, 2018.

[22] Y. Zhong, L. Zhang, and W. Gong, "Unsupervised remote sensing image classification using an artificial immune network," *International Journal of Remote Sensing*, vol. 32, no. 19, pp. 5461–5483, 2011.

[23] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7920–7930, 2020.

[24] J. Zhang, J. Liu, B. Pan, Z. Chen, X. Xu, and Z. Shi, "An open set domain adaptation algorithm via exploring transferability and discriminability for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2021.

[25] V. F. Arruda, T. M. Paixão, R. F. Berriel, A. F. De Souza, C. Badue, N. Sebe, and T. Oliveira-Santos, "Cross-domain car detection using unsupervised image-to-image translation: From day to night," *International Joint Conference on Neural Networks*, 2019.

[26] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5001–5009, 2018.

[27] C. T. Lin, "Cross domain adaptation for on-road object detection using multimodal structure-consistent image-to-image translation," in *IEEE International Conference on Image Processing*. IEEE, 2019, pp. 3029–3030.

[28] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3339–3348, 2018.

[29] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-Weak distribution alignment for adaptive object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6956–6965, 2019.

[30] C. Xu, X. Zhao, X. Jin, and X. Wei, "Exploring categorical regularization for domain adaptive object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11 724–11 733, 2020.

[31] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8869–8878, 2020.

[32] Z. Shen, H. Maheshwari, W. Yao, and M. Savvides, "SCL: Towards accurate domain adaptive object detection via gradient detach based stacked complementary losses," *arXiv preprint arXiv:1911.02559*, 2019.

[33] Y. Zheng, D. Huang, S. Liu, and Y. Wang, "Cross-domain object detection through coarse-to-fine feature adaptation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13 766–13 775, 2020.

[34] A. Wu, Y. Han, L. Zhu, and Y. Yang, "Instance-invariant domain adaptive object detection via progressive disentanglement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[35] A. Wu, R. Liu, Y. Han, L. Zhu, and Y. Yang, "Vector-decomposed disentanglement for domain-invariant object detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9342–9351, 2021.

[36] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "FADA: Feature aligned domain adaptive object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.

[37] L. Li, Z. Zhou, B. Wang, L. Miao, Z. An, and X. Xiao, "Domain adaptive ship detection in optical remote sensing images," *Remote Sensing*, vol. 13, no. 16, p. 3168, 2021.

[38] S. Chen, R. Zhan, W. Wang, and J. Zhang, "Domain adaptation for semi-supervised ship detection in SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[39] Y. Guo, L. Du, and G. Lyu, "SAR target detection based on domain adaptive faster r-cnn with small training data size," *Remote Sensing*, vol. 13, no. 21, p. 4202, 2021.

[40] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *International Conference on Machine Learning*, pp. 1180–1189, 2015.

[41] M. Frigo and S. G. Johnson, "FFTW: An adaptive software architecture for the FFT," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 3. IEEE, 1998, pp. 1381–1384.

[42] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[43] K. Do and T. Tran, "Theory and evaluation metrics for learning disentangled representations," *arXiv preprint arXiv:1908.09961*, 2019.

[44] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1100–1111, 2017.

[45] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved Faster R-CNN," in *SAR in Big Data Era: Models, Methods and Applications*. IEEE, 2017.

[46] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sensing*, vol. 11, no. 7, p. 765, 2019.

[47] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.