Weighted Pseudo-Labels and Bounding Boxes for Semisupervised SAR Target Detection

Zhuangzhuang Tian, Wei Wang, Kai Zhou, Xiaoxiang Song, Yilong Shen, and Shengqi Liu

***Abstract*—Synthetic aperture radar (SAR) image target detection methods based on semisupervised learning, such as the mean teacher framework, have shown promise in diminishing the issue of limited labeled data. However, several challenges exist in current methods. First, data augmentation techniques designed for optical images may not be suitable for SAR images due to differences in imaging methods. In addition, the contribution of pseudo labels remains constant during the initial retraining stage can lead to degradation in prediction results. Moreover, the low quality of predicted bounding boxes poses a challenge for effective retraining. To address these challenges, we propose an end-to-end semisupervised detection method based on the mean teacher framework. To enhance the robustness of training, we first introduce SAR-specific data augmentation techniques, including multiplicative noise, which effectively increase the diversity of training samples. Second, we propose a method that weights the losses of pseudo-labeled data using a hard-sigmoid function, gradually improving the importance of pseudo-labeled data during retraining, thereby alleviating their potential negative impact on the training process. Finally, we propose an IoU-aware subnetwork to incorporate high-quality pseudo-labeled bounding boxes into retraining, allowing them to contribute to network adjustments while mitigating the impact of low-quality samples. Experimental evaluations on publicly available SAR image datasets demonstrate the effectiveness of our proposed method in improving the target detection capability of semisupervised SAR target detection.**

***Index Terms*—Convolutional neural networks (CNNs), deep learning, object detection, semisupervised learning, synthetic aperture radar (SAR).**

I. Introduction

SYNTHETIC aperture radar (SAR) is an active microwave remote sensor that provides high-resolution and super-wide remote sensing images in all-day and all-weather conditions. SAR automatic target recognition (ATR) has rapidly developed alongside advancements in SAR imaging technology. Accurate target detection in SAR images is a significant and valuable research area.

Manuscript received 1 November 2023; revised 27 December 2023 and 23 January 2024; accepted 23 January 2024. Date of publication 7 February 2024; date of current version 28 February 2024. This work was supported by the Science and Technology Innovation Program of Hunan Province under Grant 2023RC3019. *(Corresponding author: Wei Wang.)*

Zhuangzhuang Tian, Kai Zhou, Xiaoxiang Song, and Yilong Shen are with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang 471000, China.

Wei Wang and Shengqi Liu are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: wangwei_nudt@hotmail.com).

Digital Object Identifier 10.1109/JSTARS.2024.3363491

In recent years, deep learning-based methods, particularly convolutional neural networks (CNNs), have demonstrated superior performance in computer vision and natural language processing, thanks to the increase in data size and computational capabilities. Similarly, CNN-based methods have shown promising results in various fields of SAR images due to their powerful feature extraction capabilities [1], [2].

Andrews [3] explored the use of active learning to selectively label "most helpful" samples, thereby reducing the amount of data required for training. Jahan et al. [4] utilized a cross-modal knowledge distillation framework for learning SAR image classification from an *electro-optical* image classification model, and designed a sampling strategy to balance the instance and class sampling, thus to improve the performance on tail classes. Inkawhich [5] proposed to obtain a global representation model by self-supervised learning on a large pool of diverse and unlabeled SAR data, and the model is then used as a fixed feature extractor. A classifier is trained to partition the feature space given the few-shot support samples. Cui et al. [6] proposed incorporating attention mechanisms into FPN and adopting dense connections to improve detection performance in complex scenes. Some researchers have made advancements in target detection for SAR ships by improving different variants of YOLO [7], [8]. In addition, for inshore SAR ships with complex background, especially ports that are closely distributed and arbitrarily oriented, rotated target detection has also received the attention of researchers. Liu et al. [9] integrated the global multiscale features with an attention mechanism, and proposed an rotation target detection method.

Although the above CNN-based methods have shown positive results in SAR target detection, most of them rely on fully supervised learning, which requires large-scale labeled training samples. However, labeled SAR images at the target-level are scarce in real-world situations, despite the abundance of SAR images. Labeling SAR images requires experienced laborers and material resources. The lack of labeled training data can degrade detection performance. As a result, semisupervised learning methods have recently gained attention from researchers [10]. These methods require only a small amount of labeled samples and focus on utilizing unlabeled or weakly labeled samples to improve detection performance.

In the context of semisupervised target detection, we are interested in semisupervised approaches that generate pseudo-labels and employ data augmentation. In this approach, detectors are initially trained using limited labeled samples and then used to predict pseudo-labels for unlabeled samples. The detectors are

subsequently retrained based on these pseudo-labeled unannotated samples. Data augmentations are applied to improve the generalization and robustness of the detectors, with the mean teacher framework [11], [12] enhancing training stability by gradually evolving the teacher model and guiding the learning of the student model.

However, the mean teacher frameworks still face three challenges in SAR image target detection. First, SAR imaging methods differ from those used in optical images, making data augmentations designed for optical images potentially unsuitable for SAR images. Therefore, it is necessary to find suitable data augmentations for SAR images.

Second, the existing method [13], [14], [15] treats the loss of both real and pseudo labels with a fixed ratio during the retraining phase. However, we have observed that the inclusion of pseudo-labeled data can deteriorate prediction results at the beginning of retraining. Therefore, the utilization of a fixed ratio between real and pseudo labels poses a potential risk of compromising the training process.

Finally, the low quality of predicted bounding boxes poses a dilemma for retraining. Some methods do not consider the pseudo label of the bounding boxes during retraining, resulting in the network being unable to adjust its parameters based on the losses of the bounding boxes. Conversely, directly incorporating the bounding boxes can degrade the retraining process.

To address these issues, we propose an end-to-end semisupervised detection method based on the mean teacher framework. First, considering the characteristics of SAR images, we introduce data augmentation methods such as multiplicative noise. Data augmentation is applied to the input data of the student network to enhance training robustness. Second, due to the instability of pseudo-labeled data in the initial retraining stage, we propose a method to weight the losses of pseudo-labeled data using a hard-sigmoid function. This gradually improves the importance of pseudo-labeled data during retraining. Finally, to incorporate high-quality pseudo-labeled bounding boxes into retraining, we propose an IoU-aware subnetwork that adjusts the participation of the bounding boxes based on their qualities. This allows high-quality pseudo-labeled samples to contribute to network adjustments while reducing the impact of low-quality ones. Experiments on publicly available SAR image datasets demonstrate that our proposed method effectively improves the target detection capability of semisupervised target detection.

II. Related Work

A. Target Detection

CNN-based target detection methods can be categorized into two types based on their detection process: 1) two-stage detectors and 2) single-stage detectors. Two-stage detectors, such as faster R-CNN [16] and FPN [17], follow a sequential approach. They first extract region proposals, which are candidate regions, from the input image. These detectors then classify the objects and regress the bounding boxes based on these regions. On the other hand, single-stage detectors, including the YOLO series [18], [19], [20] and single shot multibox detector (SSD) [21], directly predict the classification and bounding boxes without the need for region proposals. These methods have gained widespread adoption in SAR images and have demonstrated promising results.

Moreover, the transformer-based target detection method known as DETR [22], [23], [24], [25] has gained significant popularity as a research focus in recent years. DETR tackles object detection as a direct set prediction problem, employing a Transformer encoder–decoder architecture [26]. It encompasses a set-based global loss that enforces unique predictions through bipartite matching. By leveraging a fixed, small set of learned object queries, DETR analyzes the relationships between objects and the overall image context, enabling it to generate the final set of predictions in a parallel manner. This parallel nature endows DETR with remarkable speed and efficiency.

B. Semisupervised Learning in Classification

Semisupervised learning methods leverage both labeled and unlabeled data during the training process. In the domain of image classification, these methods can be broadly categorized into two main approaches: 1) consistency regularization and 2) self-training. Consistency regularization assumes that the model's predictions should remain consistent when small perturbations are introduced to the unlabeled data. Commonly used perturbations include image augmentations [27], [28], [29], adversarial training [30], [31], and model-level perturbations [11], [32]. On the other hand, self-training methods treat the predictions on unlabeled data as pseudo labels, which are then incorporated into the retraining process. Commonly used methods, such as MixMatch [33], FixMatch [34], and debiased self-training [35], employ the aforementioned steps to accomplish their objectives.

C. Semisupervised Learning in Detection

Semisupervised target detection methods draw inspiration from semisupervised classification and can also be classified into consistency regularization and self-training approaches. In current detection frameworks, such as STAC [13], Unbiased-Teacher [14], [36], Soft-Teacher [15], and LabelMatch [37], these two techniques are often combined. STAC employs a pretrained model to generate highly confident pseudo labels, and subsequently fine-tunes the network model by enforcing consistency using strong data augmentation. Mean Teachers [11], on the other hand, update the teacher model by employing an exponential moving average (EMA) of the student model's predictions over different iterations. This EMA technique enhances the stability of the teacher model and improves the quality of the predicted pseudo labels. Soft Teacher predicts the confidence of the pseudo-labels and selects retraining samples while weighting their losses based on the confidence scores. LabelMatch framework introduces redistribution mean teacher and label assignment mechanism to address the label mismathch problems during self-training. However, it is worth noting that the existing research on these methods primarily focuses on the classification task, with limited attention given to the localization

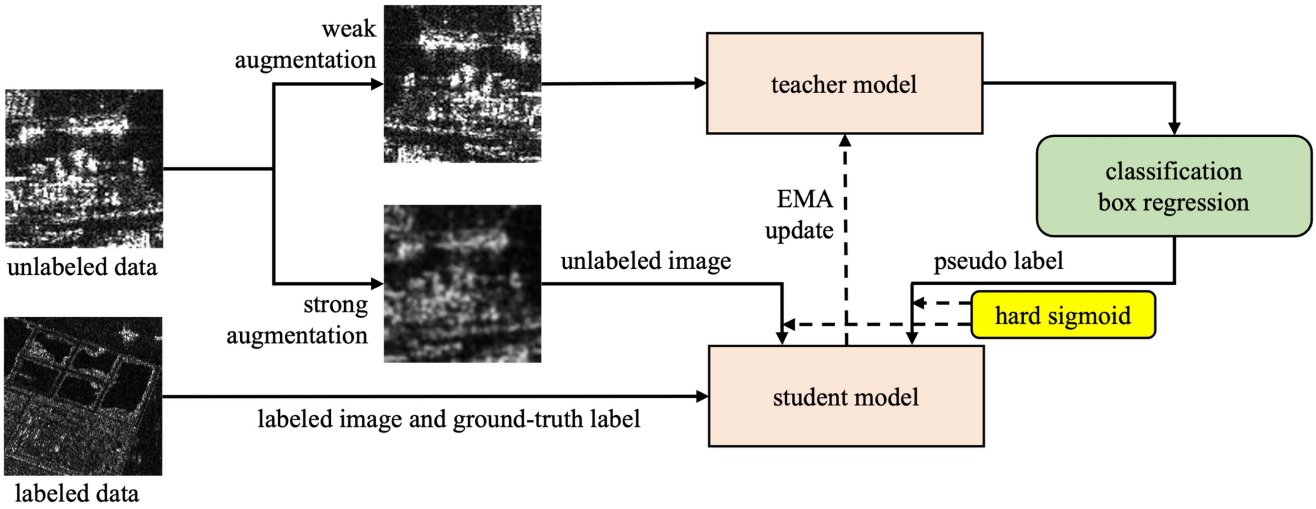


Fig. 1. Overview of the proposed semisupervised end-to-end detection framework.

task, which is an integral part of target detection and cannot be overlooked.

III. Methodology

The framework of our end-to-end detector is presented in Fig. 1. The detector consists of two models: 1) the teacher model and 2) the student model. The learning process can be divided into three stages: 1) supervised learning, 2) pseudo-label generation, and 3) semisupervised learning. Initially, we train the teacher model using the available labeled training samples and utilize it to generate pseudo labels for the unlabeled training samples. Subsequently, the student model is trained using a combination of labeled and pseudo-labeled training data. It is important to note that distinct data augmentation strategies are employed for pseudo-label generation and student model training. Furthermore, the teacher model is an EMA of the student model over a certain number of iterations. The following paragraphs provide detailed explanations of data augmentations, hard-sigmoid weighting and adaptive target box involvement.

A. End-to-End Detection Framework

The proposed end-to-end pseudo-labeling method is based on the mean teacher framework. The method comprises a teacher model and a student model, which facilitate the joint training of labeled and unlabeled data using a specified data sampling ratio. In each training iteration, the teacher model generates pseudo labels for weakly augmented unlabeled images. Subsequently, the student model is trained using both the labeled images with ground-truth labels and the strongly augmented unlabeled images with pseudo labels. As a result, the overall objective loss, denoted as $\mathcal{L}$, consists of a supervised loss $\mathcal{L}_l$, and an unsupervised loss $\mathcal{L}_u$. Specifically, it can be defined as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_l + \alpha\mathcal{L}_u \\ \mathcal{L}_l &= \sum_{i}^{N_l} \mathcal{L}_{\text{cls}}(x_i^l, y_i^l) + \mathcal{L}_{\text{reg}}(x_i^l, y_i^l) \\ \mathcal{L}_u &= \sum_{i}^{N_u} w_{\text{cls}}^i \mathcal{L}_{\text{cls}}(x_i^u, y_i^u) + w_{\text{reg}}^i \mathcal{L}_{\text{reg}}(x_i^u, y_i^u)\end{aligned} \tag{1}$$

where $\mathcal{L}_{\text{cls}}$ represents the classification loss, $\mathcal{L}_{\text{reg}}$ denotes the box regression loss. Meanwhile, x_l and y_l are the labeled image and the corresponding label, respectively, x_u and y_u correspond to the labeled image and its corresponding label, respectively. N_l and N_u represent the total number of labeled and unlabeled images. The weight α is used to control the contribution of the unsupervised loss. In addition, w_{cls}^i and w_{reg}^i represent the weights used to balance the classification loss and box regression loss during the unsupervised learning phase.

At the outset, both the teacher and student models are initialized with random weights. During the training process, the weights of the teacher model are progressively updated from the student model utilizing an EMA strategy. This approach ensures that the teacher model assimilates the accumulated knowledge of the student model over the course of training. The overview of semisupervised end-to-end detection framework is shown in Fig. 1.

B. Data Augmentation

The proposed detection framework employs two augmentation strategies, weak and strong augmentation, for the teacher model and student model, respectively. This approach introduces a form of consistency regularization widely utilized in semisupervised learning algorithms. Weak augmentation involves a

standard flip augmentation, randomly flipping images horizontally with a 50% probability. In contrast, strong augmentation offers greater diversity.

Numerous strong data augmentation methods have been developed for optical images and achieved good results. The commonly used methods include exposure adjustment, sharpness adjustment, color jitter, defocus, and so on. However, their effectiveness may not extend to SAR images due to the fundamental differences in imaging mechanisms between SAR and optical sensors. SAR images possess distinct characteristics that set them apart from optical images, rendering the aforementioned augmentation methods less applicable.

In SAR imaging, electromagnetic waves scatter in various directions and interfere with each other after hitting the target surface. This interaction of back-scattered electromagnetic waves engenders variations in pixel intensity, resulting in speckle noise. Typically, speckle noise is modeled as multiplicative noise. In consideration of the unique properties of SAR images, we propose the following data augmentation methods for the student model.

1) Pixel Dropout: In SAR images, the echo waves within the same resolution cell typically exhibit different phases, resulting in enhanced or weakened signals during coherent integration. Consequently, speckle noise manifests as random variations at the pixel level. Inspired by the concept of cutout [38], we propose the utilization of pixel dropout as a data augmentation technique. Specifically, pixel dropout involves randomly setting a fraction of pixels in the images to zero.

Cutout and pixel dropout differ primarily in terms of the occluded size in the spatial domain of the image. Cutout involves removing contiguous sections of the input images, resulting in rectangular occluded areas of relatively larger size. In contrast, pixel dropout selectively eliminates certain pixels in the input images with a given probability, resulting in randomly dispersed occluded areas.

Pixel dropout serves two key purposes. First, it simulates the variations caused by coherent integration, replicating the speckle noise phenomenon. Second, it encourages the model to reduce reliance on specific prominent features, enabling generalization to more complex scenarios. By introducing pixel dropout as a data augmentation technique, we seek to enhance the robustness and adaptability of the model.

2) Multiplicative Noise: Speckle noise in SAR images can be effectively modeled as multiplicative noise [39], where the resulting SAR image can be seen as the product of the original signal and the speckle noise. The presence of speckle noise directly influences the gray values of the image, with larger variances indicating a greater impact on the gray values. To incorporate this characteristic into data augmentation, we introduce multiplicative noise.

By applying multiplicative noise as a data augmentation technique, we aim to fine-tune the signal-to-noise ratio of SAR images. This process involves multiplying all pixels in the image by random values within a predetermined range. This injection of multiplicative noise serves two purposes: first, it encourages the network to learn more robust features by randomly adjusting pixel values with a given probability. Second, it enables the student model to adapt to SAR images with varying noise intensities, enhancing its generalization capability.

(a) (b) (c)

Fig. 2. Impacts of the introduced data augmentation methods. (a) Original example image. (b) Effect after pixel dropout processing. (c) Effect after applying multiplicative noise processing.

Fig. 2 depicts the impacts of pixel dropout and multiplicative noise on a sample SAR image, showcasing the processing effects of the introduced data augmentation method.

In addition to the aforementioned methods, SAR images from different sources may undergo contrast enhancement through different methods. To account for this variability, we further include commonly used brightness jitter and contrast jitter as data augmentation techniques. These methods help the student model adapt to SAR images with diverse contrast characteristics.

C. Hard-Sigmoid Weight

Semisupervised learning incorporates both real labeled data and pseudo-labeled data. In the initial stages of training, the quality of pseudo-labels is typically lower compared to the real labels. The loss value based on pseudo-labels stays consistent throughout the semisupervised learning progress can result in a deterioration of the training effect. Kihyuk Sohn et al. [34], argued that by screening the generated pseudo-labels, the model can acquire more high-confidence pseudo-labels during training, thereby obtaining a natural curriculum "for free." However, this method still overlooks the balance between real labels and pseudo-labels, as well as the network's adaptation to the data source during fine-tuning. Hence, it is necessary to devise a weighting method for pseudo-label data.

In designing the weighting method, we adhere to the following guidelines. First, since pseudo-label data is the predominant training data in semisupervised learning, it should be involved in the training process from the outset. Second, the accuracy of pseudo-labels should progressively improve as training progresses, and their weights should also increase accordingly, without surpassing those of the real labeled data. Finally, the generation of weights should be computationally efficient and not overly resource-intensive

$$\alpha = \begin{cases} \alpha_1, & t < t_1 \\ \alpha_1 + \frac{(1-\alpha_1)(t-t_1)}{t_2-t_1}, & t_1 \le t \le t_2 \\ 1, & t > t_2. \end{cases} \tag{2}$$

Based on the aforementioned considerations, we propose the utilization of a hard-sigmoid weight, represented by the formula shown in (2), where t indicates the tth iteration. The hard-sigmoid weight function divides the entire training process into

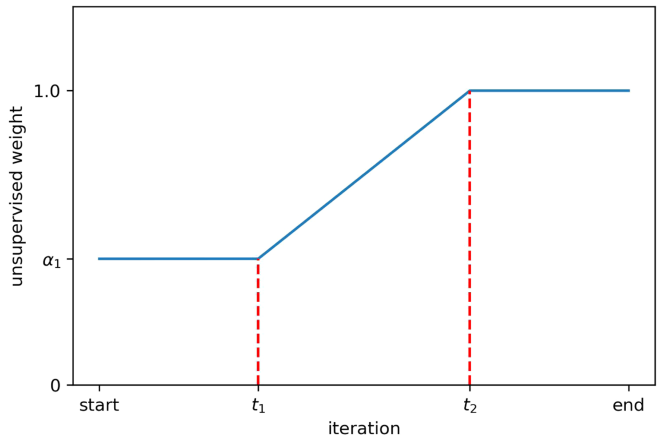


Fig. 3. Diagram of the hard-sigmoid weight.

three distinct stages. During the first stage, denoted by $t < t_1$, the hard-sigmoid function assigns a fixed low weight, denoted as α_1, to the loss values derived from the pseudo-labeled data. This allows for the inclusion of pseudo-labeled data from new data sources in the training process, while minimizing its impact on the training results compared to the original real-labeled data. Consequently, this mitigates the potential degradation of training caused by the lower quality of pseudo-labeling in the initial stages. As the training progresses into the second stage, which occurs when t falls within the range of t_1 to t_2, we assume that the network has undergone gradual fine-tuning to adapt to the unlabeled data source, thereby leading to an improvement in the quality of pseudo-labels. The hard-sigmoid function linearly increases the weight value assigned to the pseudo-labeled data. Once the weight value reaches 1, it no longer undergoes further increment and the training enters the third stage. During this stage, we posit that the network has fully adapted to the unlabeled data source, and therefore, both the pseudo-label data and the real-label data should be assigned equal weight. The weight value remains fixed at 1 until the conclusion of the training process. The variation of the hard-sigmoid weight in relation to the epochs is depicted in Fig. 3.

The hard-sigmoid weight in our proposed method allows for manual adjustments of the initial weight values, as well as the rise epoch and duration epochs. These adjustments can be made based on the specific training scenario and the dissimilarities between the two data sources. It is worth noting that the hard-sigmoid weight does not rely on the outcomes of each epoch, which enables its predefinition and direct application during training, resulting in minimal computational resource requirements. However, one drawback of this approach is the absence of adaptive adjustment capabilities.

D. Soft Box Weight

The performance of the detection method heavily relies on the quality of the pseudo-labels. Empirically, we anticipate that high-quality pseudo-labels would exert more influence during training, thereby yielding superior training results. However, most existing semisupervised target detection frameworks, derived from recognition tasks, often evaluate pseudo-label quality solely based on the utilization of classification scores. Furthermore, it has been observed in [15] that employing the intersection over union (IoU) between student-generated box candidates and teacher-generated pseudo boxes to assign foreground and background labels, with reference to real-label data, can inadvertently misclassify certain foreground box candidates as negatives. This misclassification can hinder the training process and impair overall performance. In practice, we have also noted that directly leveraging the regression loss of target boxes during semisupervised training can lead to a reduction in the final detection performance.

Some target detection frameworks choose not to incorporate the regression loss of the bounding boxes in semisupervised training, thereby mitigating the impact of inaccurate pseudo-labels. However, since the detection task encompasses both classification and localization, this approach prevents the target boxes from actively participating in the semisupervised learning process, potentially diminishing the efficacy of the localization task.

Through our experiments, we have observed that the weight value assigned to the target box loss function significantly influences the detection performance. For instance, a weight value of 2 results in pronounced degradation effects, whereas a weight value of 0.5 considerably mitigates these effects. Consequently, we propose utilizing weight values to adjust the loss function of the target boxes. Given the varying quality of the boxes, weight values should be set individually. Higher quality boxes are assigned higher weight values to encourage their greater involvement in training, and vice versa. Traditionally, IoU has been employed as the evaluation criterion for box quality. However, as previously mentioned, the accuracy of teacher-generated pseudo boxes cannot be guaranteed, rendering the obtained IoU unsuitable for reflecting box quality and potentially exacerbating results.

To address this issue, we present a box quality evaluation method based on predicted IoU. Specifically, we introduce an additional IoU prediction branch besides the classification and localization subnetworks, enabling the prediction of IoU between the predicted box and the real box. The IoU prediction branch network shares similarities with the classification and localization subnetworks, and its specific composition is depicted in Table I. To optimize parameter usage during training,

TABLE I
COMPOSITION OF THE PROPOSED IOU PREDICTION BRANCH

Layer	Type	Input/Output Dimension
FC1_1	fully connection layer	12544/1024
FC1_2	ReLU activation	
FC2_1	fully connection layer	1024/1024
FC2_2	ReLU activation	
FC3_1	fully connection layer	1024/1
FC3_2	sigmoid activation	

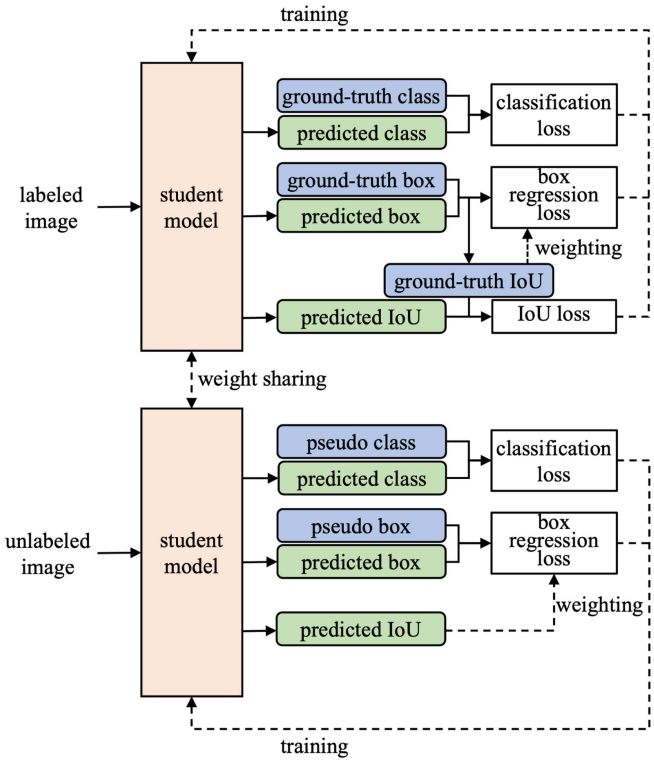


Fig. 4. Diagram of the soft bbox weight.

we reuse the first two fully-connected layers in conjunction with the classification and localization subnetworks, thereby reducing the overall number of parameters required.

During supervised training, we utilize the IoU between the predicted box and the real box as the training label. Subsequently, we employ the focal loss to calculate the loss function between the predicted IoU and the real IoU label. This loss function effectively addresses the imbalance inherent in difficult and easy samples. Focal loss $\mathcal{L}_f$ is calculated, as shown in (3), where IoU_p is the predicted IoU, IoU is the real IoU, γ is an adjustable factor that regulates the contribution of boxes with different qualities to the loss, and $\mathcal{L}_b$ is the binary cross entropy loss. As IoU_p gets closer to IoU, it decreases the loss function value, and vice versa. By reducing the loss for easier regression samples, the training process becomes more focused on challenging samples with larger prediction errors. Simultaneously, the real IoU value is employed to weight the regression loss of the target box. Conversely, when training on unlabeled data, we directly employ the predicted IoU values to weight the regression loss. The diagram of the soft box weight is shown in Fig. 4

$$\mathcal{L}_f(\text{IoU}_p, \text{IoU}) = -\beta|\text{IoU} - \text{IoU}_p|^{\gamma}\mathcal{L}_b(\text{IoU}_p, \text{IoU}). \quad (3)$$

The proposed method offers two key advantages. First, it enables the weighting of the regression loss of the target box based on the accuracy of the predicted target box. This approach retains the target box in the training process, reducing the impact of training degradation caused by low-quality boxes. Second, the proposed method incorporates the accuracy of the target box as one of the objective functions during training, resulting in improved target box accuracy.

IV. Experimental Results and Analysis

This section presents the experimental results to evaluate the detection performance of the proposed method. To provide a comprehensive assessment, we introduce the dataset utilized, the evaluation criteria employed, and share relevant implementation details. Subsequently, we present the experimental results obtained and provide a thorough analysis to evaluate the performance of the proposed method.

A. Datasets

The dataset utilized in our experiments is partitioned into three nonoverlapping subsets: 1) the labeled training set, 2) the unlabeled training set, and 3) the test set. The labeled training set is exclusively employed for early supervised learning, while both the labeled and unlabeled training sets are jointly utilized for subsequent semisupervised learning. Finally, the test set is employed to evaluate the overall performance of our method.

The dataset employed in our experiments is derived from the publicly available HRSID dataset from the University of Electronic Science and Technology of China [40]. This dataset is collected from Sentinel-1B, TerraSAR-X, and TanDEM. The image size is 800×800 and the resolutions contain 0.5, 1, and 3 m. There are 3642 images in the training set and 1962 images in the test set. To ensure a fair and unbiased evaluation, we randomly divided the training set into two parts in a ratio of 3:7, designating them as the labeled training set and unlabeled training set, respectively.

B. Evaluation Criteria

In order to comprehensively evaluate the performance of the different methods, we adopt the average precision (AP) and average recall (AR) as the evaluation metrics.

The detection results can be grouped into three categories, including true positive (TP), false positive (FP), and false negative (FN). Specifically, TP means that the IoU between the predicted box and the true box exceeds the threshold, and FP means that the IoU does not exceed the threshold. If the real box does not have a corresponding predicted box, then it is defined as FN. Thus, the precision $\mathcal{P}$ measures the proportion of TP in all detection results, and the Recall $\mathcal{R}$ measures the proportion of correctly identified positives in all positives. $\mathcal{P}$ and $\mathcal{R}$ are defined as

$$\mathcal{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\mathcal{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (4)$$

The AP computes the average value of $\mathcal{P}$ over the interval from $\mathcal{R} = 0$ to $\mathcal{R} = 1$. A high AP value means better detection performance and vice versa. According to the IoU between the predicted and real target boxes and the size of the target, AP can be further subdivided into AP_{50}, AP_{75}, AP_{S}, AP_{M}, and AP_{L}. Among them, AP_{50} and AP_{75} mean that a detected box can be

regarded as target when the IoU is greater than 0.5 and 0.75, respectively. The higher the IoU, the higher the requirement for the accuracy of the prediction box position. The $AP_{50:95}$ means that the IoU threshold is taken from 0.5 to 0.95 at intervals of 0.05 and averaged. This can more comprehensively reflect the results of the accuracy of the detection box under different IoU thresholds. AP_S, AP_M, and AP_L represent the detected targets of different sizes, where AP_S counts the results of the target area less than 32×32, AP_M counts the results of the target area between 32×32, and 96×96, and AP_L counts that greater than 96×96.

The AR signifies the average value of the recall metric, denoted as $\mathcal{R}$, across the range of IoU thresholds from 0.5 to 0.95. Specifically, AR_{100} represents the AR value calculated when 100 detections are provided per image. In addition, AR_S, AR_M, and AR_L correspond to the AR values specifically computed for small, medium, and large objects, respectively.

C. Implementation Details

The faster R-CNN equipped with FPN is adopted as the default detection framework in the proposed method. The backbone network of faster R-CNN is the ResNet-50 [41] pretrained on ImageNet dataset [42]. The strides of anchor are 4, 8, 16, 32, and 64, and the ratios of anchor are 0.5, 1.0, and 2.0. RoI pooling uses the RoIAlign with an output size of 7×7. The network adopts the stochastic gradient descent algorithm with momentum. In the initial supervised learning phase, the learning rate, momentum and weight decay are 0.0025, 0.9, and 0.0001, respectively. In semisupervised learning phase, the learning rate is reduced to 0.00125, and the momentum and weight decay remain unchanged. The total number of parameters in our networks amounts to 41.349 M, with the backbone network accounting for 23.508 M parameters. During semisupervised training, we maintain a ratio of 1:1 between pseudo-labeled samples and real labeled samples.

D. Comparison With the State of the Art

In order to evaluate the performance of the proposed method, we conducted a comparative analysis with several representative semisupervised target detection methods. The selected methods for comparison are as follows: STAC, Unbiased Teacher, Soft Teacher, and LabelMatch.

STAC [13] is a semisupervised learning framework along with data augmentation. STAC chooses the highly confident pseudo-labels of localized objects from the unlabeled image and updates the model by enforcing consistency via strong augmentation.

Unbiased Teacher [14] trains teacher and student network jointly. Two networks are given different augmented input image. The student gradually updates the teacher network via EMA. By applying EMA and focal loss, Unbiased Teacher solves the preudo-labeling bias caused by class-imbalance.

Based on the teacher–student framework, Soft Teacher [15] assesses the reliability of each box candidate generated by student network to be a real background. The reliability is then used to weigh the corresponding background classification loss. In addition, Soft Teacher samples jittered boxes around pseudo box candidates, and regresses them several times in teacher network. The box regression variance is defined as localization reliability and used to select the training sample for student network.

LabelMatch [37] recognizes that the semisupervised detection framework faces challenges related to label mismatch at both the distribution level and the instance level. To address this problem, LabelMatch introduces a redistribution mean teacher and a proposal self-assignment scheme. These mechanisms aim to align labels at the distribution level and assign appropriate labels to instances.

As a baseline, we employ the widely used faster R-CNN framework, which shares the same network architecture as our proposed methods. The key distinction lies in the fact that the baseline network is trained only using labeled data, without any semisupervised learning.

All of the aforementioned methods utilize the same backbone network and parameters as the proposed method. The main differences lie in their respective semisupervised training strategies.

The target detection results of the proposed method and the compared detection methods on SAR images are presented in Table II. It is evident from the table that all semisupervised learning methods exhibit improvements in Precision and Recall when compared to networks trained solely with labeled data.

Regarding Precision, enhancements are observed across different target sizes. Specifically, LabelMatch demonstrates higher Precision for large targets, while the proposed method excels in detecting small and medium-sized targets. Overall, the proposed method achieves the highest detection performance in terms of AP_{50}, AP_{75}, and $AP_{50:95}$, with improvements of 0.059, 0.103, and 0.075, respectively, compared to the supervised method.

In terms of Recall, improvements are also observed for targets of various sizes compared to the supervised learning method. Similar to Precision, LabelMatch exhibits better recall rates for large targets, while the proposed method performs well for small and medium-sized targets. The proposed method achieves an overall recall rate improvement of 0.058 compared to the supervised method.

E. Ablation Study

In ablation study, we conduct detailed experiments to verify our key designs. All ablation studies are conducted on the same datasets.

Effect of Different Data Augmentation Methods: In addition to the classical brightness adjustment, contrast adjustment, and Gaussian blur, the proposed method also utilizes pixel dropout and multiplicative noise as strong augmentation of the student network. We compare the effects of different data augmentation methods on the final detection performance, and the results are shown in Table III. The results presented in Table III demonstrate the impact of the introduced data augmentation methods on the AP_{50}, AP_{75}, $AP_{50:95}$, and AR_{100} metrics. Without the two proposed augmentation methods, the values for AP_{50}, AP_{75}, $AP_{50:95}$, and AR_{100} are 0.738, 0.557, 0.485, and 0.564, respectively. However, with the inclusion of pixel dropout, we observe improvements of 0.008, 0.004, 0.006, and 0.001 in AP_{50}, AP_{75},

TABLE II
OVERALL EVALUATION OF DIFFERENT TARGET DETECTION METHODS

Methods	Average Precision						Average Recall			
	$AP_{50:95}$	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_{100}	AR_S	AR_M	AR_L
Baseline	0.430	0.701	0.478	0.453	0.358	0.059	0.520	0.513	0.591	0.364
STAC	0.446	0.739	0.490	0.465	0.400	0.195	0.538	0.528	0.628	0.487
Unbiased Teacher	0.456	0.728	0.508	0.467	0.436	0.204	0.562	0.552	0.653	0.476
Soft Teacher	0.488	0.746	0.554	0.497	0.488	0.212	0.562	0.549	0.673	0.489
LabelMatch	0.480	0.739	0.540	0.490	0.487	**0.229**	0.555	0.541	0.668	**0.501**
The Proposed Method	**0.505**	**0.760**	**0.581**	**0.516**	**0.514**	0.154	**0.578**	**0.566**	**0.678**	0.495

The bold entities mean that these values are the best among the results of all methods.

TABLE III
EFFECT OF DIFFERENT DATA AUGMENTATION METHODS

PD	MN	RS	DS	$AP_{50:95}$	AP_{50}	AP_{75}	AR_{100}
				0.485	0.738	0.557	0.564
✓				0.491	0.746	0.561	0.565
✓	✓			0.495	0.748	0.565	0.568
✓	✓	✓		0.484	0.743	0.545	0.559
✓	✓	✓	✓	0.475	0.740	0.533	0.549

TABLE IV
EFFECT OF PSEUDO-LABELED DATA WEIGHT

Pseudo-labeled Data Weight	$AP_{50:95}$	AP_{50}	AP_{75}	AR_{100}
weight=0	0.475	0.740	0.533	0.549
weight=0.5	0.496	0.754	0.564	0.570
weight=1.0	0.491	0.749	0.554	0.566
weight=2.0	0.487	0.745	0.554	0.563
Linear Increase	0.484	0.748	0.544	0.560
Hard Sigmoid	0.500	0.753	0.569	0.575

$AP_{50:95}$, and AR_{100}, respectively. Moreover, the application of multiplicative noise further enhances these metrics by 0.002, 0.004, 0.004, and 0.003, respectively.

In addition to the two proposed methods, we also tried random shadow and down scale. Random shadow simulates shadows randomly in the image, while down scale decreases the image quality by first downscaling and then upscaling the image. However, both methods deteriorate the detection effect, as shown in Table III.

In Table III, PD denotes pixel dropout, MN denotes multiplicative noise, RS denotes random shadow, and DS denotes down scale.

Effect of Pseudo-Labeled Data Weight: To understand the effect of the weight value of the pseudo-labeled data on the training, we compare the detection results for the cases where the weight values are 0, 0.5, 1.0, and 2.0, respectively. Meanwhile, we also compare the linear increase strategy, namely, the weight value is increased linearly with the epoch.

In contrast to the linear increase strategy, which changes the weight values of pseudo-labeled data throughout the training process without finer adjustments for different stages, our proposed hard-sigmoid offers greater flexibility in weight adjustment, catering to the specific requirements of various training stages. The experimental results are shown in Table IV. Our findings indicate that the optimal results were obtained when the weight value was set to 0.5. Conversely, using larger constant values for the weights, along with a linear increase strategy, negatively impacted the detection results. The proposed hard sigmoid weight exhibited superior detection results in comparison, underscoring the efficacy of our proposed method.

TABLE V
EFFECT OF BBOX LOSS WEIGHT

Bbox Loss Weight	$AP_{50:95}$	AP_{50}	AP_{75}	AR_{100}
weight=0	0.500	0.753	0.569	0.575
weight=0.5	0.496	0.750	0.563	0.570
weight=1.0	0.489	0.745	0.551	0.561
weight=2.0	0.480	0.752	0.533	0.559
Soft Box Weight	0.505	0.760	0.581	0.578

Effect of Bbox Loss Weight: To assess the influence of different box weights on the training outcomes, we varied the box weights. Specifically, we set the weights to 0, 0.5, 1.0, and 2.0, respectively, and evaluated their impact on the training results. The experimental outcomes are presented in Table V. We observed that the most favorable results were achieved when the box weight was fixed at 0. This finding suggests that employing a fixed positive weight may have a detrimental effect on test performance. In contrast, the proposed soft box weight exhibited an improvement in the training of semisupervised learning, thereby providing a best detection results.

Detection Result Instances: Fig. 5 illustrates two distinct scenarios: 1) the sea surface scenario in the top row and 2) the near-shore scenario in the bottom row. In the figure, correctly detected targets are represented by blue boxes, incorrectly recognized targets by red boxes, and undetected missed targets by yellow boxes. It is evident that the proposed method demonstrates effective ship detection in the simpler sea scene. However, in the

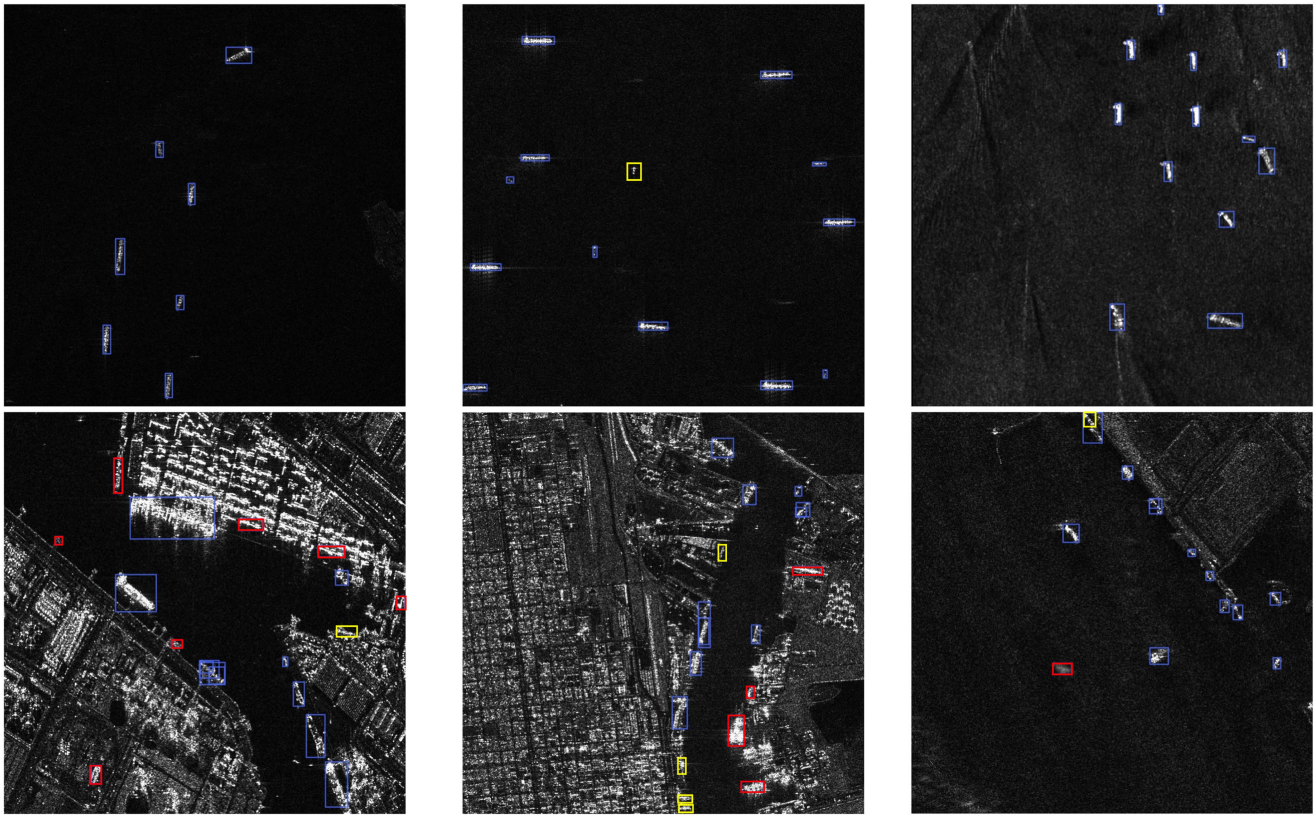


Fig. 5. Examples of the detection results on HRSID dataset.

near-shore scenario, some false alarms and missed targets are still present. This highlights the ongoing difficulty in detecting SAR ship targets in near-shore scenarios, which will serve as a crucial area for future method enhancements.

V. Discussion

Despite the demonstrated superiority of the proposed method in addressing SAR image target detection, a notable challenge remains. Currently, there is a scarcity of data augmentation techniques specifically designed for SAR images through digital image processing. In our experiments, we randomly selected labeled training samples. However, to bridge this gap, future research endeavors should prioritize the identification of appropriate training samples and the exploration of suitable data augmentation methods for SAR images. Furthermore, in the case of SAR ships, rotating boxes offer a more refined approach for target labeling compared to horizontal boxes. Consequently, future research should aim to explore the extension of the semisupervised learning method to accommodate rotating boxes.

VI. Conclusion

In this article, we proposed an end-to-end semisupervised detection method based on the mean teacher framework for SAR image. We introduce data augmentation techniques, such as multiplicative noise, tailored to the characteristics of SAR images to enhance training robustness. We also propose a weighting method that utilizes a hard-sigmoid function to gradually increase the importance of pseudo-labeled data during retraining, mitigating the instability observed in the initial stages. Furthermore, we introduce an IoU-aware subnetwork that selectively incorporates high-quality pseudo-labeled bounding boxes into retraining, enabling effective network adjustments while minimizing the influence of low-quality samples. Experimental results on publicly available SAR image datasets demonstrate the efficacy of our proposed method in enhancing the target detection capability of semisupervised target detection.

References

[1] D. Xiang, Y. Xie, J. Cheng, Y. Xu, H. Zhang, and Y. Zheng, "Optical and SAR image registration based on feature decoupling network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5235913.

[2] D. Xiang, Y. Xu, J. Cheng, Y. Xie, and D. Guan, "Progressive keypoint detection with dense siamese network for SAR image registration," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 5, pp. 5847–5858, Oct. 2023.

[3] S. Andrews, "Active learning for target detection and classification in SAR imagery," *Proc. SPIE*, vol. PC 12520, 2023, Art. no. PC1252006.

[4] C. S. Jahan and A. Savakis, "Balanced sampling meets imbalanced datasets for SAR image classification," *Proc. SPIE*, vol. 12525, 2023, Art. no. 1252506.

[5] N. Inkawhich, "A global model approach to robust few-shot SAR automatic target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 4004305.

[6] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[7] A. A. Adegun, J. V. Fonou Dombeu, S. Viriri, and J. Odindi, "State-of-the-art deep learning methods for objects detection in remote sensing satellite images," *Sensors*, vol. 23, no. 13, 2023, Art. no. 5849.

[8] K. Patel, C. Bhatt, and P. L. Mazzeo, "Deep learning-based automatic detection of ships: An experimental study using satellite images," *J. Imag.*, vol. 8, no. 7, 2022, Art. no. 182.

[9] H. Liu, L. Wang, C. Zhao, N. Wang, and J. Chen, "Rotating target detection of SAR image based on multi-scale attention module for inshore ships," in *Proc. IEEE IGARSS Int. Geosci.Remote Sens. Symp.*, 2023, pp. 7034–7037.

[10] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8934–8954, Sep. 2023.

[11] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Redhook, NY, USA: Curran Associates, Inc., 2017.

[12] Q. Yang, X. Wei, B. Wang, X.-S. Hua, and L. Zhang, "Interactive self-training with mean teachers for semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5941–5950.

[13] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," 2020, *arXiv:2005.04757*.

[14] Y.-C. Liu et al., "Unbiased teacher for semi-supervised object detection," 2021, *arXiv:2102.09480*.

[15] M. Xu et al., "End-to-end semi-supervised object detection with soft teacher," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3040–3049.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 28, pp. 91–99.

[17] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[19] J. Redmon and A. Farhadi, "YOLOV3: An incremental improvement," 2018, *arXiv:1804.02767*.

[20] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of Yolo algorithm developments," *Procedia Comput. Sci.*, vol. 199, pp. 1066–1073, 2022.

[21] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020, *arXiv:2005.12872*.

[23] Z. Cai, S. Liu, G. Wang, Z. Ge, X. Zhang, and D. Huang, "Align-DETR: Improving DETR with simple IoU-aware BCE loss," 2023, *arXiv:2304.07527*.

[24] S. Zhang et al., "Dense distinct query for end-to-end object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7329–7338.

[25] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 6748–6758.

[26] A. Vaswani et al., "Attention is All You Need," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Redhook, NY, USA: Curran Associates, Inc., 2017.

[27] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Redhook, NY, USA: Curran Associates, Inc., 2020, pp. 6256–6268.

[28] X. Hu, Y. Zeng, X. Xu, S. Zhou, and L. Liu, "Robust semi-supervised classification based on data augmented online ELMs with deep features," *Knowl.-Based Syst.*, vol. 229, 2021, Art. no. 107307.

[29] C. Gong, D. Wang, and Q. Liu, "Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13683–13692.

[30] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[31] A. R. Sajun and I. Zualkernan, "Survey on implementations of generative adversarial networks for semi-supervised learning," *Appl. Sci.*, vol. 12, no. 3, 2022, Art. no. 1718.

[32] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Redhook, NY, USA: Curran Associates, Inc., 2016.

[33] D. Berthelot et al., "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, H. Wallach et al., Eds., vol. 32, Redhook, NY, USA: Curran Associates, Inc., 2019.

[34] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Redhook, NY, USA: Curran Associates, Inc., 2020, pp. 596–608.

[35] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Redhook, NY, USA: Curran Associates, Inc., 2022, pp. 32424–32437.

[36] Y.-C. Liu, C.-Y. Ma, and Z. Kira, "Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9819–9828.

[37] B. Chen et al., "Label matching semi-supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14381–14 390.

[38] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[39] V. S. Frost, J. A. Stiles, K. S. Shanmugan, and J. C. Holtzman, "A model for radar images and its application to adaptive digital filtering of multiplicative noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-4, no. 2, pp. 157–166, Mar. 1982.

[40] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

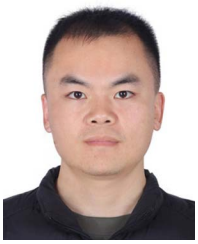
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

Zhuangzhuang Tian received the B.S. degree in communication engineering from Hunan University, Changsha, China, in 2014, and the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, in 2020.

He is currently a Research Associate with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang, China. His research interests include image processing, computer vision, and machine learning in remote sensing images.

Wei Wang received the M.S. degree in communication and information processing from the National University of Defense Technology, Changsha, China, in 2013, and the Ph.D. degree in geoinformatics from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2018.

He is currently an Associate Professor with the College of Electronic Science, National University of Defense Technology. His research interests include SAR/PolSAR image processing, machine learning, and automation target recognition.



Kai Zhou received the B.S. degree in communication engineering from Hohai University, Nanjing, China, in 2015, and the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2022.

He is currently a Research Associate with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang, China. His research interests include waveform design, electronic countermeasures, SAR signal processing, and machine learning.

Xiaoxiang Song received the B.S. degree in communication engineering from the School of Computer and Electronic Information, Guangxi University, Guangxi, China, in 2016, and the Ph.D. degree in communication engineering from the Institute of Communications Engineering, Army Engineering University, Nanjing, China, in 2021.

He is currently a Research Associate with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang, China. His research interests include intelligent unmanned system, edge computing, and remote sensing.

Yilong Shen received the B.S. degree in information security and countermeasure from the Beijing Institute of Technology, Beijing, China, in 2011.

He is currently a Research Associate with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang, China. His research interests include SAR imaging and SAR signal processing.

Shengqi Liu received the B.S. degree in communication engineering, and the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2009 and 2016, respectively.

He is currently an Associate Professor with the College of Electronic Science, National University of Defense Technology. His research interests include radar signal processing, feature extraction, and automatic target recognition.