Cross-Domain Few-Shot Learning Between Different Imaging Modals for Fine-Grained Target Recognition

Yuan Tai, Yihua Tan, Member, IEEE, Shengzhou Xiong, and Jinwen Tian

Abstract—Fine-grained target recognition in synthetic aperture radar (SAR) or infrared imaging modal is an open problem in some application scenarios where training samples are scarce. Transferring common features from visible optical (VO) samples is effective for the case that SAR (infrared) samples are scarce. However, for the fine-grained target recognition, transferring common features face two issues: first, common features can be divided into fine-grained features and the coarse-grained features. For the fine-grained target recognition task in the few-shot case, how to transfer fine-grained common features needed to be considered. Second, in the SAR (infrared) imaging modal, parts of samples carry much noise because of the limitation of the imaging mechanism, making the subtle difference for the fine-grained target recognition task, making such fine-grained common features not easy to be transferred, especially when training samples are scarce. To handle these issues, corresponding solutions are proposed in this article as follows: first, the common-feature-contrastive loss is proposed to transfer fine-grained common features from VO samples; second, based on the modeling of the heteroscedastic uncertainty, the training strategy of sample quality evaluation is proposed to emphasize the training samples with less noise. Experiments on three datasets, including MSTAR, P-openSARship, and P-V AIS, represent the superiority of the proposed algorithm over baseline and other popular cross-domain few-shot learning algorithms.

Index Terms—Cross-domain few-shot learning (CDFSL), heteroscedastic uncertainty, infrared images, synthetic aperture radar (SAR).

I. INTRODUCTION

BECAUSE of the robustness to complex weather conditions and illumination variation, synthetic aperture radar (SAR) and infrared imaging modals play important roles in many applications, such as military investigation and early warning. Among these applications, distinguishing targets of different subclasses is important for decision-making, which bring the fine-grained target recognition task. However, it is not easy to obtain sufficient SAR (infrared) samples of concerned subclasses in such application scenarios, bringing the few-shot learning problem. Therefore, fine-grained target recognition in the few-shot case is an essential technology in many applications of the SAR (infrared) imaging modal.

Manuscript received 17 August 2022; accepted 4 October 2022. Date of publication 10 October 2022; date of current version 2 November 2022. This work was supported by the National Natural Science Foundation of China under Grant 41371339 and in part by the Fundamental Research Funds for the Central Universities under Grant 2017KFXJ179. (Corresponding author: Yihua Tan.)

The authors are with the School of Artificial Intelligence and Automation, State Key Laboratory of Multispectral Information Processing Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: t_y_@hust.edu.cn; yhtan@hust.edu.cn; xiongsz@hust.edu.cn; jwтан@mail.hust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3212680

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Different from traditional target recognition, fine-grained target recognition aims to distinguish subclasses of targets [1], e.g., the cargo and tug in Fig. 1. The core problem in fine-grained target recognition is to learn fine-grained feature representation that can represent the subtle difference among different subclasses [2]. Further, in the SAR (infrared) imaging modal, compared with the visible optical (VO) targets containing rich discriminative texture features that can represent the fine-grained feature, such features of the SAR (infrared) targets are difficult to extract. For example, the patterns or colors of different subclasses VO targets are various, but for the SAR (infrared) targets, such features are similar among different subclasses. To learn such fine-grained feature, contrastive [3] or triplet loss [4] are used in the existing algorithms of the SAR (infrared) imaging modal [5], [6], [7], [8]. However, these algorithms are prone to be overfitting in the few-shot case. For the few-shot learning in the SAR (infrared) imaging modal, some algorithms are based on data augmentation [9] or using the metalearning training strategy to learn the model [10], [11]. However, such algorithms need sufficient SAR (infrared) samples of related classes, which limits the application scenarios. Recently, Tai et al. [12] proposed to transfer “common features” between SAR and VO imaging modals from a complex source network to a small target network to enhance the target network’s ability to extract common feature so that the requirement for SAR (infrared) samples is reduced. Common features are the features that can be learned by VO
samples and enable to recognize SAR (infrared) samples, which represent the shared features between the VO and SAR (infrared) samples, such as the shape feature of VO and SAR (infrared) ship [12]. However, during transferring the common features, they did not consider the difference of common features among different subclasses, causing the transferred common features to be unable to represent the subtle difference among fine-grained classes, decreasing their performance for the fine-grained target recognition task.

Inspired by the work of Tai et al. [12], which enhanced the target network’s ability to extract common features by transferring them from the source network with the connection-free attention modules, this article deals with the fine-grained target recognition of the SAR (infrared) imaging modal in the few-shot case by considering the following issues.

1) Learning common features that are discriminative among different fine-grained classes are critical to the fine-grained target recognition of the SAR (infrared) imaging modal in the few-shot case. According to the proposed works [5], [8], contrastive loss is effective to learn fine-grained features by pushing the distance of features between different subclasses away from each other. In the normal fine-grained recognition task, fine-grained features from different subclass samples are expected to be as far as possible. So the margin in the contrastive loss, which represent the expected distance between different subclasses is simply set as a relatively large value. However, for the fine-grained target recognition task of the SAR (infrared) imaging modal in the few-shot case, fine-grained common features are expected to be transferred, so that the margin should not be too large or small. The common features learned in [12] locate the middle layers of the network where the transfer weight obtains a higher value. This phenomenon shows that common features exist in a specific area of the embedding space. If the margin is set too large, the distance of fine-grained common features between the different subclasses will be pushed too far to locate in the specific area of the embedding space, losing the ability to recognize SAR (infrared) targets. Conversely, most features cannot provide the loss, making fine-grained common features hard to be transferred. Therefore, obtaining a suitable “margin” is the critical problem for transferring fine-grained common features.

2) In the SAR (infrared) imaging modal, some samples carry much noise because of the limitation of the imaging mechanism (see Fig. 1), which may mask the critical parts of the SAR (infrared) images that can be used to distinguish targets of different subclasses, making such fine-grained common features hard to be transferred, especially in the few-shot case. Besides, noise in samples will make the effective features more difficult to be captured and decrease the predictive power of the convolutional neural network (CNN) [13]. Therefore, the noise problem needs to be considered to transfer fine-grained common features.

For the first issue, different from setting a fixed margin based on experience, we define the expected value of margin as common-feature boundary (CF-boundary) to keep the classification ability of common features. Further, based on the CF-boundary, the CF-contrastive loss function is proposed to avoid choosing an unsuitable “margin” manually. Specifically, first, considering that common features between different imaging modal could be more dissimilar from that in the same imaging modal, we obtain the CF-boundary by calculating a mean distance of the corresponding features between VO and SAR (infrared) samples. Then, the “margin” is learned by minimizing its distance to the CF-boundary (see Fig. 2). Besides, additional connection-free attention modules are constructed to connect each feature pair of the source and target networks when the inputs are from different subclasses, namely the push-attention modules. The original connection-free attention modules in [12] are renamed as pull-attention modules for distinguishing. When the inputs of both the source and the target networks are from the same subclass, the pull-attention module is learned to transfer common features. When the inputs are from different subclasses, the push-attention module is learned to make such common features discriminative between different fine-grained classes so that the fine-grained common features are transferred. Generally, the CF-contrastive loss is proposed to learn to transfer fine-grained common features for the fine-grained target recognition in the few-shot case. Note that the CF-boundary that represent the expected distance between the VO and SAR (infrared) samples provides additional supervision information for learning the fine-grained common features.

The second issue can be overcome by decreasing the contribution of samples containing much noise to the network training. To achieve this aim, the training strategy of “sample quality evaluation” (SQE) is proposed, which regards samples with much noise as the low-quality samples and focuses on the high-quality samples during the training process. Specifically, existing works [14], [15], [16] demonstrate that heteroscedastic uncertainty represents the noise inherent in samples, such as sensor noise or motion noise. Therefore, in SQE, the noise of training samples is measured by modeling the heteroscedastic uncertainty, then a lower learning rate is given to the training samples with higher heteroscedastic uncertainty. Generally,
Fig. 3. Overview of the proposed algorithm. Improvements compared with the baseline are represented in the red dotted box. 1) Push-attention modules and the CF-contrastive loss are introduced to learn fine-grained common features. 2) Based on the modeling of heteroscedastic uncertainty, the SQE training strategy focuses on training samples with less noise. VO samples are used to pretrain both networks to obtain the ability to extract common features, and SAR (infrared) samples are input into both networks to extract common features [12]. Then, the fine-grained common features are transferred to the target network’s suitable position to enhance the target network’s fine-grained common feature extraction. Therefore, both the VO and SAR samples are input to both the source and the target networks.

SQE training strategy is proposed to transfer fine-grained common features by mitigating the influence of noise.

In this article, two novel designs are proposed (contents in the red dotted box of Fig. 3): 1) both pull-attention and push-attention modules are established to learn to transfer fine-grained common features with CF-contrastive loss; 2) based on the modeling of heteroscedastic uncertainty, the training strategy of SQE is proposed.

The contributions are as follows.
1) CF-boundary is proposed to represent the expected distance of fine-grained common features between different imaging modalities.
2) CF-contrastive loss is proposed for learning to transfer fine-grained common features, in which the “margin” is learned based on the proposed CF-boundary.
3) Based on the modeling of heteroscedastic uncertainty, the training strategy of SQE is proposed to focus on training samples with less noise, decreasing the difficulty of transferring fine-grained feature representation.

II. RELATED WORK

A. Cross-Domain Few-Shot Learning (CDFSL)

CDFSL is first proposed by FWT [17], which learns representations by using featurewise transform. In CDFSL, auxiliary samples and novel samples obey different distributions. Wang et al. [18] proposed CDSFL-ATA to alleviate the problem of inductive bias for metalearning-based algorithms through task augmentation. Fu et al. [19] proposed a meta-FDMixup network to utilize to guide the model learning. However, these algorithms are limited in visible imaging modal and do not suit for SAR (infrared) imaging modal, such as SAR (infrared) imaging modal.

B. Fine-Grained Recognition for SAR (Infrared) Images

Metric-based algorithms are mainly used to cope with this problem. Lin et al. [20] proposed the task-driven dictionary learning framework based on the SAR-HOG features. Lang et al. [21] proposed a wrapper feature selection framework to learn a joint feature. Wang et al. [22] utilized the geometry and scattering characteristics of the target. Margarit and Tabasco [23] used a fuzzy logic decision rule to combine the scattering features and the geometric features to categorize ships. Xu et al. [24] proposed distribution shift metric learning by adding a distribution regularization term. He et al. [25] designed a task-specific backbone network to extract powerful deep features with triplet loss [26]. To train the network well, they rearrange the OpenSARShip [27] dataset to acquire enough friendly training samples.

C. Uncertainty in Deep Learning

Uncertainty can be captured with the Bayesian network [14]. Specifically, uncertainty consist of two main types: epistemic uncertainty and aleatoric uncertainty [16]. Epistemic uncertainty captures the unknown information from the unseen data, which can be explained with increased training data. On the other hand, aleatoric uncertainty captures uncertainty that cannot be explained by collecting more data. Further, aleatoric uncertainty consist of two subclasses: homoscedastic uncertainty and heteroscedastic uncertainty. Homoscedastic uncertainty is not dependent on the input data. Its value stays constant for inputs in the same task but varies between different tasks. Heteroscedastic
uncertainty is predicted as a model output. Recently, uncertainty has been used in many tasks, such as pixelwise depth regression [28], multitask learning [29], continual learning [30], face recognition [31], 3-D face reconstruction [32], image de-convolution [33], and autonomous driving [34]. In such tasks, uncertainty benefits the network from various aspects. However, the potential of uncertainty remains to be tapped in few-shot learning.

III. REVISITING BASELINE

We review the baseline [12] first for the completely introduction of the proposed algorithm. In general, similar to Fig. 3, the baseline transfer common feature with pull-attention modules. In total, eight ResNet blocks [35] and four Bayesian-CNN with four Bayesian ResNet blocks are introduced into the source and target networks, respectively. We introduce three major contents of the baseline in the following sections.

A. Transfer Common Features With Pull-Attention Modules

In the baseline [12], we regard $S$ and $T$ as the source and target network, respectively. Common features are transferred from $S$ to $T$ by matching each feature between them with learnable pull-attention modules.

Specifically, let $x_t$ be the input image of the SAR (infrared) imaging modal and $y_t$ be the corresponding class label. Then, $x_t$ is fed into $S$ and $T$ to obtain features. Let $S_m^m(x_t)$ be the $m$th layer feature of $S$ and $T_n^n(x_t)$ be the $n$th layer feature of the $T$. The F-norm of between the distance $S_m^m(x_t)$ and $T_n^n(x_t)$ is minimized to transfer features. Meanwhile, the baseline weights the feature connection ($S_m^m, T_n^n$) and its $c$th channel by $c_{m,n}^c$ and $w_{m,n}^c$, respectively. Finally, the objective function is given as follows:

$$
L_t(\theta, \phi_1 | x_t) = \sum_{m,n \in Q} \lambda_{m,n} \sum_c f(F_{m,n}^m(x_t), w_{c,m,n}^c) 
$$

(1)

where $F_{m,n}^m(x_t) = \|S_m^m(x_t) - z_{m,n}^m(T_n^n(x_t))\|_F^2$, $Q$ represent all the feature connections, $z_{m,n}$ is an $1 \times 1$ convolution function to equal channels of the corresponding feature connection ($S_m^m, T_n^n$), and $f(a, b)$ represents pixelwise multiplication is performed between each value in the $c$th channel of $a$ and the value in that of $b$. Note that the dimension of $F_{m,n}^m(x_t)$ is $MN \times C$, where $M$ and $N$ are the width and height of $T_n^n(x_t)$, respectively, $w_{m,n}^c$ is a $1 \times 1$ convolution and $c_{m,n}^c$ is a value, which are obtained by the pull-attention module parameterized by $\phi_1$. $C$ represents the number of channels in $S_m^m(x_t)$. Note that $\theta$ is used to represent the parameters of both $S$ and $T$, denoted as $T_\theta$

B. Accurately Updating Important Parameters

To find important parameters, Bayesian-CNN [36] is constructed as the target network. Specifically, each parameter of the target network is initialized with the Gaussian distribution, and $\theta_t \sim N(\mu_t, \sigma_t^2)$ represent the $t$th parameter. The importance of a parameter is thought to be inversely proportional to its standard deviation. Further, during the training process, the learning rate $\beta_t$ of $\theta_t$ is set based on its standard deviation

$$
\beta_t \leftarrow \beta_t \times \gamma_t
$$

(2)

where $\gamma_t = \frac{1}{\log(1+\epsilon_t)}$ and $\beta$ is the initial learning rate.

C. Objective Function

The Bayesian network is learned by the variational inference. Specifically, the Kullback–Leibler (KL) divergence with the parameters’ posterior and a designed distribution $q(\theta | t)$ parameterized by $\epsilon$ is minimized

$$
\theta^* = \arg \min _\theta KL(q(\theta | t) || P(\theta | x_t, y_t)).
$$

(3)

Further, $N$ Monte Carlo samples [36] is used to approximate (3)

$$
L_{tg}(\theta | x_t, y_t, \epsilon) 
\approx \sum_{j=1}^N - \log \left( P(\theta^j) - \log \left( P\left( y_t | x_t, \theta^j \right) \right) \right) + \log q(\theta^j | t)
$$

(4)

where $j$ is the $j$th operation of sampling.

Finally, the objective function of the baseline is given as follows:

$$
L_{\text{wo}}(\theta, \phi | x_t, y_t, \epsilon) = L_t(\theta, \phi_1 | x_t) + L_{tg}(\theta | x_t, y_t, \epsilon).
$$

(5)

IV. PROPOSED ALGORITHM

A. Overview

Fig. 3 visualizes the overall structure of our algorithm. Compared with the baseline, the proposed algorithm focus on the fine-grained target recognition task by the following operations (contents in the red dotted box):

1) push-attention modules and the CF-contrastive loss are proposed to learn fine-grained common features;
2) an additional fully connected layer is constructed to output the heteroscedastic uncertainty;
3) based on the modeling of heteroscedastic uncertainty, the training strategy SQE is proposed to focus on the training samples with less noise.

B. Learn Fine-Grained Common Features

For the fine-grained target recognition task of the SAR (infrared) imaging modal in the few-shot case, common features that discriminative among different subclasses need to be transferred. Therefore, the CF-contrastive loss and push-attention modules are proposed to transfer fine-grained common features.

Specifically, when the inputs of the source and target networks are both from the $p$th subclass, (1) is minimized (the same as the baseline). Further, when the inputs of both networks are from the $p$th and $q$th subclasses, we push the $T_{\theta_p}^p(x_{tp})$ away from the $S_m^m(x_{tp})$ to learn discriminative features by maximizing the following objective function given as follows:

$$
L_{\text{diff}}(\theta, \phi_2 | x_{tp}, x_{tq}) = \lambda_{\text{diff}} \sum_c f(F_{m,n}^m(x_{tp}, x_{tq}), w_{c,m,n}^n)
$$

s.t. $p \neq q$

(6)
where $\lambda_{m,n}^{\text{diff}}$ and $w_{m,n}^{\text{same}}$ have the same form as that in (1) but calculated by push-attention modules. Push-attention modules have the same structure as the pull-attention modules but parameterized by $\phi_2$, $\mathcal{F}^{m,n} = \| (S^{m}(x_{t_p}) - x_{m,n}(T^{m}_\theta(x_{t_q})) )^2_F$, and $f$ is the same as the one in (1).

For a clear description, parts of (1) is renamed as $\mathcal{L}_{\text{same}}^{m,n}(\theta, \phi_1 | x_{t_p}, x_{t_q}')$

$$\mathcal{L}_{\text{same}}^{m,n}(\theta, \phi_1 | x_{t_p}, x_{t_q}') = \lambda_{m,n}^{\text{same}} \sum_c f(\mathcal{F}^{m,n}(x_{t_p}, x_{t_q}'), w_{c,m,n}^{\text{same}})$$

where $x_{t_p}$ and $x_{t_q}'$ are two different samples of the $t$th subclass. Further, the contrastive loss is introduced to minimize $\mathcal{L}_{\text{same}}^{m,n}$ and maximize $\mathcal{L}_{\text{diff}}^{m,n}$ at the same time

$$\mathcal{L}_{\text{contrastive}}^{m,n}(\theta, \phi_1, \phi_2 | x_{t_p}, x_{t_q}) = \text{label} \ast \mathcal{L}_{\text{same}}^{m,n} + (1 - \text{label}) \ast \max(\text{margin} - \mathcal{L}_{\text{diff}}^{m,n}, 0)$$

(8)

where margin is the hyperparameter that control the expected distance between $S^{m}(x_{t_p})$ and $T^{m}_\theta(x_{t_q})$, and label indicates whether the inputs of the source and target network belong to the same subclass. The value of label is 1 when the inputs belong to the same subclass; otherwise, it is 0.

C. CF-Contrastive Loss

In the contrastive loss, the hyperparameters’ margin is difficult to choose manually. This is because that margin with too low value makes little common features provide the loss, and conversely, features of the target network will be pushed too far to keep the ability for classifying SAR (infrared) targets. Therefore, the CF-contrastive loss is proposed to avoid choosing an unsuitable “margin” manually.

Specifically, we give each feature pair a corresponding learnable margin that is regarded as $\upsilon_{m,n}$, where $(m, n)$ represent the feature pair $(S^{m}(x_{t_p}), T^{m}_\theta(x_{t_q}))$. Thus, (8) becomes

$$\mathcal{L}_{\text{contrastive}}^{m,n}(\theta, \phi_1, \phi_2 | x_{t_p}, x_{t_q}) = \text{label} \ast \mathcal{L}_{\text{same}}^{m,n} + (1 - \text{label}) \ast \max(\upsilon_{m,n} - \mathcal{L}_{\text{diff}}^{m,n}, 0).$$

(9)

However, $\upsilon_{m,n}$ will become zero or even negative in the learning process. Thus, the CF-boundary is proposed to constrain the learning of margin, which are designed to represent the expected distance that we would push a common feature away from another one. Considering that common features between different imaging modal could be more dissimilar from that in the same imaging modal, we obtain the CF-boundary by calculating a mean distance of the corresponding features between VO and SAR(infrared) samples. Specifically, the CF-boundary is calculated by the following two steps (see Fig. 4): 1) extracting features of VO and SAR (infrared) samples with the source network; 2) calculating the F-norm between the features of the corresponding layer. For example, $CF^{m} = \sum_c \frac{1}{|M_N|} \| S^{m}(x_s) - S^{m}(x_{t_q}) \|^2_F$ represents one sampling of the CF-boundary of the $m$th layer, where $MN$ is the spatial resolution of the features of $m$th layer and $x_s$ is the corresponding VO samples randomly selected from $X_S = \{x_{s1}, x_{s2}, \ldots \}$. Note that VO and SAR (infrared) samples are not needed to be the same target. Based on the CF-boundary, the following objective function is given to learn the margin

$$\mathcal{L}_{\text{margin}}^{m,n}(x_{s}, | x_{t_q}, x_{t_q}' ) = (\upsilon_{m,n} - CF^{m})^2$$

(10)

where $CF^{m}$ is the mean value of $CF^{m}$ after 200 sampling times.

Finally, based on (9) and (10), the CF-contrastive loss for learning fine-grained common features is proposed as follows:

$$\mathcal{L}_{\text{CF-contr}} = \sum_{(m,n) \in Q} \mathcal{L}_{\text{margin}}^{m,n} + \mathcal{L}_{\text{contrastive}}^{m,n}$$

(11)

where $Q$ represent all the feature connections.

D. Training Strategy of “SQE”

Existing research [13] shows that noise will make the effective features more difficult to extract. However, some SAR (infrared) samples carry much noise because of the limitation of the imaging mechanism, which is regarded as the “low-quality” samples in this article. Such low-quality samples make the network learn more information about noise, reducing the ability of the network to distinguish targets of different subclasses. Therefore, we aim to reduce the contribution of such samples to network training. To achieve this aim, the training strategy of “SQE” is proposed. In SQE, the heteroscedastic uncertainty is modeled in the target network to measure the quality of training samples because it captures the noise inherent in the samples [14]. Then, a lower learning rate is given to the samples with higher heteroscedastic uncertainty.

Specifically, heteroscedastic uncertainty can be regarded as an output of the network [16], thus an additional fully connected layer of the target network is constructed to obtain the heteroscedastic uncertainty (see Fig. 3): $\hat{y}(\tau, \tau) = T^{\theta}(x_{t_q})$, where $\hat{y}$ is the output class and $\tau$ is the heteroscedastic uncertainty of the input $x_{t_q}$. Besides, to ensure the value of heteroscedastic uncertainty is a small positive number, we map $\tau$ to $[0, 2]$ by: $\hat{\tau} = \left[ \frac{2\arctan(\tau)}{\pi} \right] + 1$, namely constrained heteroscedastic uncertainty (CHU). Further, considering that the parameters is updated with a minibatch of samples, we propose to make the learning rates of parameters inversely proportional to the mean value of
CHU in a minibatch

\[ \beta_i \leftarrow \beta \times \frac{1}{\bar{\tau}} \]  

where \( \bar{\tau} \) is the mean value of CHU \( \tau \) of training samples in a minibatch.

Besides, considering the training strategy of AUIP of the baseline, the learning rate of the target Bayesian-CNN parameters is scaled according to the heteroscedastic uncertainty of a minibatch training samples and parameters’ standard deviation. Specifically, in a minibatch, the learning rate \( \beta_i \) of the \( i \)th parameter \( \theta_i \) is given as follows:

\[ \beta_i \leftarrow \beta \times \frac{1}{\log(1 + e^{\tau_i})} \times \frac{1}{\bar{\tau}} \]  

where \( \beta \) is the learning rate and \( \sigma_i \) is the standard deviation of the \( i \)th parameters.

### E. Objective Function

This section shows the way of modeling the heteroscedastic uncertainty and give the total objective function.

The heteroscedastic uncertainty can be modeled as the regularization term of the objection function [16]. Thus, in this article, the heteroscedasticity uncertainty of input samples is modeled in the target Bayesian-CNN network by the following objective function:

\[ L_{\text{her}}(\theta|x, y) = \frac{1}{2n_2} L_{\text{bayes}}(\theta|x_t, y_t, t) + \frac{1}{2} \log \bar{\tau}^2. \]  

Finally, the total objective function \( L_{\text{total}} \) is given as follows:

\[ L_{\text{total}}(\theta, \phi_1, \phi_2, v | x, x_p, x_t, y_t) = L_{\text{CF-contr}}(\theta, \phi_1, \phi_2, v | x, x_p, x_t, y_t) + L_{\text{her}}(\theta|x_t, y_t, t). \]  

Algorithm 1 shows the training scheme and \( \phi = \{ \phi_1, \phi_2 \} \) represents the parameters of both the pull-attention and the push-attention modules.

### V. EXPERIMENT

#### A. Data Preparation

There are two types of datasets: source domain and target domain datasets. All source domain data are applied for pre-training the both the source and the target networks. The training samples in the target domain dataset are used to train the target network. In this article, the source domain dataset is a VO dataset consisting of two subdatasets, and three target domain datasets include two SAR and one infrared dataset.

1) Source Domain Dataset: The source dataset consists of a VO vehicle and a VO ship subdatasets.

1) VO vehicle subdataset consists of 2639 targets of the vehicle and 2754 samples of background [37].

2) VO ship subdataset consists of 1000 samples of the ship and 3000 samples of the background [38].

2) Target Domain Datasets: MSTAR dataset [39] is the SAR vehicle dataset. Table I expresses the details. Fig. 5 shows several examples of the dataset.

1) P-OpenSARShip is the SAR ship dataset, which comes from the OpenSARShip dataset [27]. In total, five subclass targets are used to construct the P-OpenSARShip dataset, including Cargo, Dredging, Fishing, Tanker, and Tug. The detail is given in Table II. Fig. 6 shows several examples of the dataset.

2) P-VAIS is the infrared ship dataset, which comes from VAIS [40] dataset. In total, seven subclasses infrared targets are used to construct the P-VAIS dataset, including Cargo, Fishing, Passenger, Pleasure, Sailboat, Speedboat, and Tug. The specific number of each category is given in Table III. Fig. 7 shows several examples of the dataset.

Note that the categories of targets in the VO and SAR (infrared) do not need to be the same but just lie in a similar coarse-grained class. For example, the VO dataset corresponding to the MSTAR dataset is the normal car dataset, which can be easily downloaded through the Internet. Besides, compared with other few-shot learning SAR ATR algorithms, which require samples of similar distribution (dataset) to pretrain the model, the related VO training samples used in our algorithms are much easier to collect in the real scene. For example, algorithms [9], [10], [11] set samples of three classes as the support set and all the samples of the other seven classes are used to pretrain the
TABLE I

DETAILS OF MSTAR DATASET

| Types | 2S1 | BRDM_2 | BTR_60 | D7 | SN_132 | SN_9563 | SN_C71 | T62 | ZIL131 | ZSU_23_4 |
|-------|-----|--------|--------|----|--------|---------|--------|-----|--------|---------|
| Training set | 274 | 274 | 195 | 274 | 232 | 233 | 233 | 273 | 274 | 274 |
| Testing set | 299 | 299 | 237 | 299 | 196 | 196 | 299 | 299 | 299 | 299 |

Fig. 5. Examples in the MSTAR dataset. (a) Examples of 2S1. (b) Examples of T62.

TABLE II

NUMBER OF EACH TYPE OF SHIPS IN SAR IMAGING MODAL

| Types | Cargo | Dredging | Fishing | Tanker | Tug |
|-------|-------|----------|---------|--------|-----|
| Dataset | 170 | 80 | 126 | 170 | 176 |

TABLE III

NUMBER OF EACH TYPE OF SHIP IN INFRARED IMAGING MODAL

| Types | Cargo | Fishing | Passenger | Pleasure | Sailboat | Speedboat | Tug |
|-------|-------|---------|-----------|----------|----------|-----------|-----|
| Dataset | 177 | 28 | 191 | 124 | 289 | 97 | 99 |

Fig. 6. Examples in the P-OpenSARShip dataset. (a) Examples of Cargo. (b) Examples of Dredging.

Fig. 7. Examples in the P-VAIS dataset. (a) Examples of Cargo. (b) Examples of Fishing.

model in the MSTAR dataset, which means they need to collect other military targets. Differently, we just need to collect the normal vehicle targets.

B. Compared Algorithms

For a comprehensive comparison, the compared algorithms consist of two types. The first type is mainstream metalearning-based algorithms, including FWT [17], CDFSL-ATA [18], and meta-FDMixup [19]. The second type is the transfer-based algorithms, including L2T [41], DTLF [42], [43], and NWPU fine-tune and the baseline. These algorithms do not focus on the CDFSL problem but achieve a competitive performance compared with those metalearning-based algorithms. Note that literature [42] and [43] belong to the same algorithm but published in the journal of Remote Sensing and CVPR workshop.

C. Settings

1) Evaluation Setup: All the metalearning-based algorithms are pretrained on the mini-imageNet [45]. For transfer-based algorithms (including the proposed one), network is pretrained with the source domain dataset. The five-way one-shot and five-shot are two popular cases in metalearning-based few-shot learning algorithms [46], [47], in which 15 query samples is set. Therefore, for a fair comparison with metalearning-based algorithms, we adopt scenarios of five-way one-shot and five-shot in our article. Specifically, in the five-way five-shot case for the MSTAR, we randomly sample five classes among the total ten classes at one time, and we conduct such sampling 600 times. Each time, the sampled five class is random and different, so that each class of the total ten classes will be included.

model in the MSTAR dataset, which means they need to collect other military targets. Differently, we just need to collect the normal vehicle targets.
Further, we report the mean value and standard deviation of accuracy of 600 times sampling. Note that, because two kinds of sets exist in the MSTAR dataset (training and testing sets), training samples (support samples) and testing samples (query samples) are sampled from the training and the testing sets, respectively.

2) Training: We train each model for 400 epochs. The optimizer is Adam. For both kinds of attention modules parameters, the learning rate is $10^{-6}$. For the parameters of the target network, the learning rate is 0.1. The initial value and learning rate of each margin are 0.25 and 0.01, respectively. The target network parameters are a mixed Gaussian distribution. Its mean value and variance are 0 and 0.0025. All samples are resized to $32 \times 32$.

### D. Results

1) Comparative Results: Tables IV and V show the comparative results with other SOTA metalearning-based CDFSFL algorithms and transfer-based algorithms. We can see that our algorithm reaches the highest accuracy in all datasets. Specifically, compared with the metalearning-based CDFSFL algorithms, the maximum accuracy improvement reaches 30.09%, 7.32%, and 12.92% for the MSTAR, P-OpenSARship, and P-V AIS datasets, respectively. Compared with transfer-based algorithms, the maximum accuracy improvement reaches 11.47%, 4.98%, and 7.00% for these datasets, respectively.

Besides, to validate the influence of the number of query samples, we sample 35, 55, and 75 query samples in one sampling time of the five-way five-shot case of the MSTAR dataset. The results are shown in Table VI.

Besides, the experiments in the three-way five-shot case are conducted in the MSTAR dataset to compare the proposed algorithm with other few-shot SAR ATR algorithms [9], [10], [11]. Note that such few-shot SAR ATR algorithms set three classes as the support set, and all the samples of the other seven classes are used to pretrain the model. Therefore, for a fair comparison, we also perform the same setting. Table VII tabulates the experimental results.

### TABLE VI

| Numbers/algorithm | 15   | 35   | 55   | 75   |
|-------------------|------|------|------|------|
| FWT+protonet      | 55.46±0.55 | 55.81±0.52 | 55.12±0.37 | 56.28±0.59 |
| FWT+matchingnet   | 43.25±0.24 | 43.17±0.21 | 44.02±0.18 | 44.06±0.21 |
| FWT+gmmnet        | 49.96±0.38 | 49.72±0.32 | 49.36±0.40 | 49.58±0.31 |
| CDFSFL-ATA        | 55.91±0.39 | 55.21±0.46 | 56.21±0.58 | 56.34±0.51 |
| FDMixup           | 54.25±0.51 | 54.30±0.62 | 54.12±0.33 | 54.87±0.29 |

| L2T               | 63.21±2.31 | 64.15±2.12 | 63.38±2.32 | 63.55±1.76 |
| DTLF              | 43.13±2.99 | 44.54±2.13 | 44.31±2.56 | 43.65±1.78 |
| NWPU              | 50.21±2.25 | 50.12±2.36 | 51.47±2.14 | 50.95±1.97 |
| Fine-tune         | 55.22±3.50 | 55.31±2.90 | 55.74±1.80 | 55.82±2.01 |
| Baseline          | 74.53±4.15 | 75.34±3.75 | 74.79±3.22 | 75.67±2.15 |

| Ours              | 86.00±2.33 | 85.83±1.77 | 84.82±1.25 | 86.29±1.72 |

### TABLE VII

| Algorithm          | MGA-Net [9] | HIN [10] | CBLPN [11] | Ours |
|--------------------|-------------|----------|------------|------|
| 3-way 5-shot       | 95.99±0.04  | 96.67±0.02 | 95.62±0.01 | 98.17±0.03 |
| 3-way 1-shot       | 91.82±0.04  | 88.85±0.07 | 89.87±0.01 | 91.67±0.06 |

### TABLE VIII

| Dataset/algorithm | MSTAR    | P-OpenSARship | P-VAIS |
|-------------------|----------|--------------|--------|
| Baseline          | 74.53±4.15 | 49.27±1.96   | 62.20±2.80 |
| SQE               | 83.80±2.10  | 50.27±3.20   | 64.00±0.89 |
| CF-contrastive    | 83.87±2.70  | 50.20±4.97   | 64.53±2.07 |
| Ours              | 86.00±2.33  | 54.25±2.51   | 69.20±1.79 |

2) Ablation Experiments: Tables VIII and IX tabulate the results of the ablation experiments for the proposed designs. SQE represents that we use SQE training strategy and without using
Ablation Experimental Results in the Vehicle Dataset (Left-Hand Side) and the Ship Dataset (Right-Hand Side) for the Five-Way One-Shot in Accuracy (in %)

| Dataset/algorithms | MSTAR  | P-OpenSARShip | P-VAIS |
|---------------------|--------|---------------|--------|
| Baseline            | 49.07±2.10 | 42.80±2.76 | 32.13±2.07 |
| SQE                 | 51.20±3.84 | 45.33±3.11 | 33.87±1.36 |
| CF-contrastive      | 52.67±2.36 | 44.07±2.98 | 33.20±0.93 |
| Ours                | 55.73±3.03 | 47.56±2.36 | 35.00±1.58 |

Ablation Experimental Results in Both Datasets for Different Margins in the Five-Way Five-Shot Case (in %)

| Dataset/Margin | MSTAR  | P-OpenSARShip | P-VAIS |
|----------------|--------|---------------|--------|
| 5              | 82.60±2.97 | 51.21±2.36 | 65.21±1.46 |
| 3              | 83.21±2.68 | 52.77±2.62 | 66.63±1.68 |
| 1              | 83.67±2.23 | 51.91±3.15 | 66.96±1.21 |
| 0.5            | 83.14±3.32 | 52.43±2.61 | 66.78±1.59 |
| Ours           | 86.00±2.33 | 54.25±2.31 | 69.20±1.79 |

Ablation Experimental Results in Both Datasets for Different Margins in the Five-Way One-Shot Case (in %)

| Dataset/Margin | MSTAR  | P-OpenSARShip | P-VAIS |
|----------------|--------|---------------|--------|
| 5              | 52.75±2.91 | 44.32±2.56 | 32.42±1.50 |
| 3              | 53.11±2.36 | 44.21±2.63 | 31.74±1.35 |
| 1              | 53.45±2.03 | 45.83±2.17 | 32.23±1.56 |
| 0.5            | 53.93±1.26 | 45.99±2.63 | 31.16±1.05 |
| Ours           | 55.73±3.03 | 47.56±2.36 | 35.00±1.58 |

5) Analysis of Feature-Connection Weight: To explore the knowledge transfer, we record the feature-connection weight $\lambda_{m,n}$. Fig. 9 shows the results. Layer $n$ in each row and column represents the $n$th layer of the source target networks, respectively. In Fig. 9(a), the features of the second and third layers obtaining a higher weight for transferring, which has the same conclusion as the baseline. Further, among all the weights $\lambda_{m,n}$ in Fig. 9(b), $\lambda_{2,4}^{diff} = 1.75$ and $\lambda_{4,2}^{diff} = 1.71$ are significantly higher than other values, which indicates that features in these two feature connections need to be discriminative.

E. Discussion

We analyze the proposed algorithm from aspects of performance and the training process.

1) Performance: First, according to the abovementioned experimental results, it can be concluded that the proposed algorithm obtains the highest accuracy among all the algorithms on all datasets. Also, ablation experiments represent that the proposed designs are effective. We explain the following reasons.

1) We distinguish samples of different fine-grained based on the proposed CF-contrastive loss so that the common features are discriminative enough, increasing the network performance for the fine-trained target recognition task.

2) In the SAR (infrared) imaging modal, some training samples carry much noise, which is harmful to the network training, especially in the few-shot case, but existing algorithms did not deal with this problem. This article decreases the contribution of low-quality samples to the network training and strength the learning of targets features. Thus, the performance of the network is increased.
Fig. 9. Value of weight $\lambda_{m,n}$ of MSTAR dataset for five-way five-shot in the pull-attention modules and push-attention modules. (a) $\lambda_{m,n}^{\text{same}}$ of the pull-attention modules. (b) $\lambda_{m,n}^{\text{diff}}$ of the push-attention modules.

Besides, we can see a large difference in the recognition accuracy between MSTAR and P-OpenSARship. This is because the P-OpenSARship dataset is very challenging. Specifically, first, the sizes of samples are different, which distorts the samples feed into the network; second, for some samples, more than two ships exist in one sample. Differently, MSTAR is a well-arranged dataset whose samples are the same size, and each sample only contains one target. Therefore, the recognition accuracy between MSTAR and P-OpenSARship is very different.

2) Training Process: Fig. 10 compares the training process of the baseline and ours on the P-V AIS dataset. From Fig. 10, we can see that the proposed algorithm converge faster than that of the baseline, but fluctuate more. This is because the low-quality training samples are fitted to a lower degree, which impacts the loss value.

3) Complexity: Although the Bayesian network contains more parameters than the original CNN, the target network containing 9.8 M parameters also has the advantage of scale to some popular small networks, such as ResNet18 (11.2 M) [35], VGG19 (20.0 M) [50], and Xception (20.8 M) [51]. Besides, on the GeForce 2080Ti, the inference time for each testing sample only needs 0.00078 s, which meets the real-time requirements.

VI. CONCLUSION

A CDFSL-DIM algorithm for fine-grained target recognition is proposed in this article. The CF-contrastive loss is proposed to learn distinguishing common features. Besides, we propose a novel training strategy, namely SQE, which adjusts the learning rates of the minibatch according to the heteroscedastic uncertainty of training samples. Compared with other metalearning-based CDFSL algorithms, the maximum accuracy improvement reaches 30.09%, 7.32%, and 12.92% for the MSTAR, P-OpenSARship, and P-V AIS datasets, respectively. Compared with transfer-based algorithms, the corresponding accuracy improvement reaches 11.47%, 4.98%, and 7.00% for these datasets, respectively. For the future work, algorithms for reducing the number of parameters of the target network are worth studying.

REFERENCES

[1] D. Lin, X. Shen, C. Lu, and J. Jia, “Deep LAC: Deep localization, alignment and classification for fine-grained recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1666–1674, doi: 10.1109/CVPR.2015.7298775.

[2] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, “Embedding label structures for fine-grained feature representation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1114–1123.

[3] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2005, pp. 539–546, doi: 10.1109/CVPR.2005.202.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 815–823.

[5] Y. Xu and H. Lang, “Ship classification in SAR images with geometric transfer metric learning,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 8, pp. 6799–6813, Aug. 2021.

[6] J. He, Y. Wang, and H. Liu, “Ship classification in medium-resolution sar images via densely connected triplet CNNs integrating fisher discrimination regularized metric learning,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 4, pp. 3022–3039, Apr. 2021.

[7] Y. Xu, H. Lang, X. Chai, and L. Ma, “Distance metric learning for ship classification in SAR images,” in Proc. Image Signal Process. Remote Sens. XXIV, vol. 10789, 2018, pp. 441–451.

[8] Y. Xu and H. Lang, “Distribution shift metric learning for fine-grained ship classification in SAR images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 2276–2285, 2020.

[9] M. Yang, X. Bai, L. Wang, and F. Zhou, “Mixed loss graph attention network for few-shot SAR target classification,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5216613, doi: 10.1109/TGRS.2021.3124336.

[10] L. Wang, X. Bai, C. Gong, and F. Zhou, “Hybrid inference network for few-shot SAR automatic target recognition,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 11, pp. 9257–9269, Nov. 2021, doi: 10.1109/TGRS.2021.3051024.
Yuan Tai received the B.E. degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 2017, where he has been working toward the Ph.D. degree in artificial intelligence with the National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Artificial Intelligence and Automation since 2017.

Yihua Tan (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2004. From 2005 to 2006, he was a Postdoctoral Staff with the Department of Electronics and Information, HUST. Since 2005, he has been with the School of Artificial Intelligence and Automation, HUST, where he is currently a Professor. During 2010–2011, he was a Visiting Scholar with Purdue University, focusing research on remote sensing image analysis. He has authored more than 80 papers in journals and conferences. His current research interests include digital image/video processing and analysis, object detection and recognition, and machine learning.

Shengzhou Xiong received the M.S. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2018, where he has been working toward the Ph.D in artificial intelligence with the National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Artificial Intelligence and Automation.

Jinwen Tian received the Ph.D. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 1998. He is currently a Professor with the School of Artificial Intelligence and Automation, HUST. His current research interests include remote sensing image analysis, image compression, computer vision, and fractal geometry.