Contrastive Feature Disentangling for Partial Aspect Angles SAR Non-cooperative Target Recognition

Zaidao Wen, *Member, IEEE*, Jiaxiang Liu, Zhunga Liu, *Member, IEEE*, Sijian Li, and Quan Pan

***Abstract*— Deep learning algorithms have achieved the state of the art progress on synthetic aperture radar (SAR) automatic target recognition (ATR) tasks. They theoretically assume that training and testing samples are independent and identically distributed (i.i.d) for generalization, but it is intractable for practical non-cooperative ATR (NC-ATR) scenarios. In this paper, we propose a novel contrastive feature disentangling framework termed ConFeDent to learn features with improved generalization performance under a condition of a weaker distribution consistency. More specifically, ConFeDent aims to describe the semantic interactions between two arbitrary SAR training samples instead of independently treating them. It can implicitly disentangle features encoding the aspect and identity knowledge from entire samples with a semi-parametric geometric transformation model and a second-order energy model. Except for the identity label, we especially exploit the deductive-based geometry knowledge as supervision to teach the model to learn the concept of aspect angle variation. A progressively amortized inference scheme is constructed for efficient feature learning and recognition in an end-to-end way. Finally, we sequentially stack two parameter-shared ConFeDent to release a strengthened version termed ConFeDent+, which can explicitly utilize and learn more information from cross-category samples. Experimental results on the moving and stationary target acquisition and recognition (MSTAR) benchmark demonstrate the effectiveness of our proposed models in the SAR NC-ATR. In particular, we validate the algorithms in a more challenging scenario where the range of aspect angles for training and testing samples is permitted to be disparate. Our model can achieve much higher recognition accuracy than other SAR ATR algorithms.**

***Index Terms*— automatic target recognition, synthetic aperture radar, feature disentangling, out of distribution, partial aspect angles, contrastive learning, extended operating condition**

I. Introduction

AUTOMATIC target recognition (ATR) is one of the ultimate goals in the field of synthetic aperture radar (SAR), which has attracted much attention in civil and military reconnaissance and surveillance for years [1]–[3]. In contrast to optical sensors, the electromagnetic imaging mechanism of SAR makes the resulting image reconstruction of the specular backscattering of the illuminated target [4]. Therefore, SAR will actually "see" some types of physical structures of a target, and scattering signatures in the resulting SAR image will be highly sensitive to geometric acquisition conditions, e.g., azimuth, depression angle. Crucially, the profile and the topology structure of scattering points of the target will vary a lot as the viewing angle of the SAR platform changes. The issue of geometric acquisition sensitivity will make SAR ATR more challenging than a general image recognition task of optical sensors.

To tackle this issue, conventional scattering center model-based SAR ATR algorithms collect the configurations of strong scattering points/centers of the target in full ($0 \sim 2\pi$) aspect angles [5], [6]. These complete scattering patterns will form a feature template in the hope of collecting all the scattering signatures in different poses uniformly. Then pairwise template matching is used to classify a testing target. However, the scattering features will be influenced by many practical disturbances, such as speckle noise and motion ambiguity, in which case the performance will degrade if the matching metric is inappropriate. To improve the robustness, several hand-crafted visual feature extractors are designed to capture some intrinsic pose-invariant signatures of the target. These symbolic features are normally robust to pose variations, and thus they can usually achieve better performance than scattering ones [7]. A classifier is learned in the pre-computed feature domain. The paradigm of deductive feature extraction followed by classifier learning has dominated for years. The increase in available computation and sufficient samples has enabled the learning process to shift from the hand-crafted feature domain to the raw sensory data domain, which allows to learn features and a classifier simultaneously. Accordingly, many machine learning-based SAR ATR algorithms are developed [8]–[20] in recent years, among which the most notable one will be the convolutional neural networks (CNN). Distinct from the deductive feature extraction algorithms, the learning model addresses the pose sensitivity issue in an inductive way. It aims to search a parametric function from a set of rotating target and their class labels to induce rotation-invariant features in an implicit way. This type of end-to-end learning pipeline can significantly improve the SAR ATR accuracy and efficiency as long as sufficient rotating samples of each class are available.

A. Motivation

Notwithstanding the undeniable success of the above learning models, especially CNN, critics have recently drawn attention to a number of following limitations.

- Weak generalization ability. Compared with the conventional deductive features, the generalization ability of the inductive model relies on a theoretical assumption that the training and testing samples are statistically drawn from an identical distribution. However, due to the sensitivity

This work was supported in part by the National Natural Science Foundation of China under Grant 62176213,61806165, 61790552, U20B2067; the Natural Science Basic Research Plan in ShaanXi Province of China under Grant 2020JQ-196.

The authors are with Key Laboratory of Information Fusion Technology of Ministry of Education, Northwestern Polytechnical University, Xi'an, Shaanxi Province 710072, China.

of the SAR image to the geometric acquisition condition [2], [14], [19], the practical condition for a testing target may be different from the training ones, including variations in depression angle, resolution, target configuration, environmental electromagnetic noise, etc. It is intractable to collect a large volume of targets to cover enough patterns to ensure distribution consistency, yielding the so-called small sample issue [21]. In particular, for a non-cooperative (NC) target, we can only reconnoiter a few samples from limited viewing angles for training, while the testing scenarios are unlimited. In this scenario, it is prone to fail disastrously once exposed to samples outside the training distribution, yielding the complicated out-of-distribution (o.o.d) classification problems [22].

- Low interpretation ability. The current learning model will mostly induce the concept from the samples in the spirit of mathematics rather than the human mind. As a result, it is typically a 'black box inducing' process where the computations are distinct from the humanly comprehensive reasoning and inference. The resulting intermediate features are generally semantic meaningless. As such, the learning-based SAR ATR algorithms may meet some potentially incalculable risks in NC-ATR.

The above two limitations raise an urgent requirement for developing a learning-based algorithm with improved generalization and interpretation ability for practical SAR NC-ATR. A simple and plausible way to alleviate the out-of-distribution (o.o.d) problem in the computer vision area would be a data augmentation trick or novel views generation [23], i.e., manually generating pseudo targets with full aspect angles for distribution completion and alignment. We, however, empirically found that this trick is invalid for SAR images. It might be due to ambiguous artifacts such as over-smoothness caused by nonlinear pixel interpolation that will not correctly reflect the actual physical scattering signature of the real target. Therefore, the pseudo samples do not explicitly provide more discrimination information than the original ones [18], [24], [25]. Alternatively, computer simulation technologies (CST), such as Microwave Studio Asymptotic Solver, can be used to generate the simulated targets for training and transfer the complete aspect angle information to the real data domain via transfer learning algorithm [4], [26]. However, this scheme relies on an accurate 3D computer-aided design (CAD) target model [4], which is intractable for NC targets. The fundamental issue of improving the generalization and interpretation of induced features is not addressed. Wen *et al.* proposed a rotation-aware framework that aims to capture the rotation equivariant and label invariant features of the SAR target [21]. They exploited a self-supervised strategy to enforce the model to understand the rotation variance among several intra-class samples so that it can alleviate the influence of the rotation sensitivity and improve the classification performance under limited training samples. However, the framework does not consider the challenging partial aspect angle (PAA) situation. Additionally, the interpretability of the resulted rotation equivariant features is not guaranteed as well.

B. *Contribution*

This paper proposes a novel contrastive feature disentangling framework termed ConFeDent which aims at alleviating the above issues. Instead of independently learning the pose invariant features for each target in a discriminative way, ConFeDent resorts to disentangling the pose features from their identity ones by exploiting the semantic equivariant relationship between two targets in a contrastive manner. In this way, it can implicitly learn and transfer complete pose knowledge from the cooperative training samples to non-cooperative ones and improve the generalization ability and interpretability of the learned features for SAR ATR. To this end, ConFeDent elaborates a semi-parametric geometric transformation model and a conditional random field (CRF) model as two semantic explainable regularizers to capture the structured and contrastive relation in the feature space. Furthermore, ConFeDent involves a progressively amortized inference for efficient feature extraction and recognition via an encoder-decoder scheme. Finally, we stack two parameter-shared ConFeDent in sequence to produce a strengthened version termed ConFeDent+. Experimental results on the MSTAR benchmark demonstrate the effectiveness of our proposed approach. Compared with the other ATR algorithms, ConFeDent and ConFeDent+ can both achieve higher recognition accuracy in the task of SAR ATR, especially in PAA NC-ATR. The main contributions are summarized as follows.

- We propose a novel contrastive feature disentangling model termed ConFeDent and a strengthened version ConFeDent+ to induce features with improved generalization and interpretation ability. Compared with the other SAR ATR algorithms, the proposed models can achieve much better SAR NC-ATR performance not only under the general standard operating condition (SOC) and extended operating condition (EOC) but also under a more complicated PAA scenario. To the best of our knowledge, it is the first time for a learning based SAR ATR algorithm to evaluate the performance under PAA condition.
- We develop a novel idea of learning to disentangle the pose and identity features by comparing the semantic relations between two arbitrary SAR targets instead of treating them independently. The induced pose feature subspace from all training samples will contain complete aspect angle information. The missing aspect angle information of the NC samples will be sampled and transferred from ones in cooperative/aspect available classes, increasing the combinational generalization ability.
- We incorporate a semi-parametric geometric transformation model and a second-order energy model for contrastive regularization. Except for the identity label, we especially exploit the deduction-based geometry knowledge as supervision to teach the model to learn the concept of aspect angle. By taking the best from both worlds, it inherits the benefit of both deductive and inductive learning analogous to human learning, making the features more interpretable than other deep features for SAR ATR.

The rest of this paper is organized as follows: Section II introduces the preliminary. Section III formally proposes the main framework of ConFeDent and its detailed architecture. Extensive experiments are carried out in Section IV to demonstrate the effectiveness and superiority of the proposed model. Section V summarizes our work and suggests future directions.

II. Problem Formulation and Related Analysis

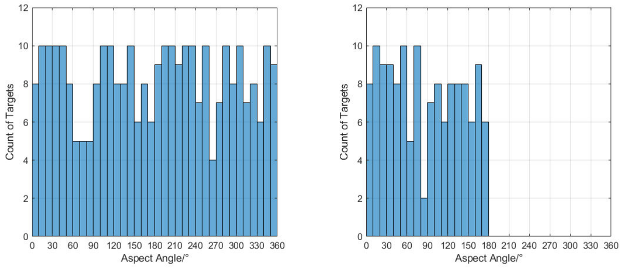
The aims in this section are threefold. First, we formally propose the new type of EOC to formulate the problem of PAA NC-ATR and analyze some insights. Second, we discuss the equivariant property of features to shed light on our solution to PAA NC-ATR. Finally, we introduce the related research on contrastive learning to present our idea.

A. Extended Operating Condition of PAA and Insight

Due to the electromagnetic imaging mechanism of the SAR sensor, the scattering appearance of an illuminated target will be severely sensitive to the acquisition conditions, e.g., relative pose between the target and the sensor. Consequently, a slight pose variation of the target will yield an apparent change in the resulting SAR image where the latent factors accounting for the pose and identity are heavily entangled. The conventional algorithms to alleviate this challenge will exploit some physical-driven or geometric-guided methods to extract some rotation-invariant features in a deductive way. Alternatively, the inductive-based learning models such as CNN will adaptively learn features by designing a network to map each intra-class training sample to the same identity label vector. The induced features, in this way, will account for the identity label and somewhat ignore the intra-class pose variances. Compared with the conventional hand-crafted feature engineering, the inductive learning-based ATR model is possible to discover new data-adaptive solutions which can probably create new recognition knowledge beyond those canonical ones. However, it is unconformable in realistic battlefield deployment scenarios of SAR NC-ATR since the testing target acquisition condition will be mostly different from the training phase. To this end, several extended operating conditions (EOC) summarized in Table I are used to validate the generalization performance of a SAR ATR algorithm [2]. The aspect angles of samples for training and testing are complete in these conventional EOCs, namely uniformly residing in the range of $[0, 2\pi]$. If only a few targets with a partial range of aspect angles are available for training while those in the testing phase are unlimited, it will be a much more complicated condition for SAR ATR. It will be defined as a new EOC termed partial aspect angle (PAA) for SAR NC-ATR, which requires more generalization ability to unseen views.

TABLE I
Summary Of Current Extended Operating Conditions [2]

extended operating conditions	**specification**
acquisition geometry	depression angles squint angles
target state variations	articulation alternative configurations intra-class variability
local target deployment	obscuration: occlusion and layover camouflage deception



(a) cooperative target (b) non-cooperative target

Fig. 1. Illustration of the aspect angles histogram in PAA.

More formally, let $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_c}$ be N_c labeled SAR targets drawn from c-th class for training, where $\mathbf{x}_i$ and $\mathbf{y}_i$ stand for the SAR target image and its identity label, respectively, N_c is the number of the training samples in c-th class. Let C and C_n be the number of overall and NC categories in the training set, respectively. PAA suggests that the aspect angles of the training samples in the NC classes are drawn from a subset $[a, b] \subset [0, 2\pi]$. For example, Figs. 1 illustrate the aspect angles histograms in the case of $a = 0, b = \pi$. The ultimate goal of PAA NC-ATR is to predict the label of a new query sample whose aspect angle is drawn from $[0, 2\pi]$ according to the training data.

B. Equivariant Feature Disentanglement

The deep learning based SAR ATR algorithms classify a query sample $\hat{\mathbf{x}}$ according to the following objective optimization:

$$\arg\max_c \mathcal{E}^*(\hat{\mathbf{x}}), \text{ s.t. } \mathcal{E}^* \in \arg\min_{\mathcal{E}} \sum_c \sum_{i=1}^{N_c} \ell(\mathbf{y}_i, \mathcal{E}(\mathbf{x}_i)) \quad (1)$$

where ℓ is a classification loss function, $\mathcal{E}$ is the designed discriminative model such as a CNN. If $\hat{\mathbf{x}}$ and all training samples $\mathbf{x}_i$ are drawn from the same distribution, Eq.(1) can theoretically learn a well-generalized $\mathcal{E}^*$ that achieves satisfactory classification performance. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be two input SAR target images from the same class. Assuming that there is an oracle function $\mathcal{T}_{i\to j}$ accounting for the transformation from $\mathbf{x}_i$ to $\mathbf{x}_j$, it will entirely capture their similarity and dissimilarity. Then there are a set of transformations $\mathcal{T}_c = \{\mathcal{T}_{i\to j}\}_{i,j=1,\ldots,N_c}$ encoding the empirical intra-class variations among the training samples. Note from Eq.(1) that the discriminative learning will facilitate a function that is invariant to these transformations as:

$$\mathcal{E}^*(\mathbf{x}_i) \approx \mathcal{E}^*(\mathbf{x}_j) = \mathcal{E}^*(\mathcal{T}_{i\to j}(\mathbf{x}_i)), \ \forall i, j \quad (2)$$

Therefore, if the training samples can cover the underlying $\mathcal{T}_c$ implicitly or are augmented by $\mathcal{T}_c$ explicitly, $\mathcal{E}^*$ will be invariant to $\mathcal{T}_c$ conditioned on its hypothesis space. For the aspect angle variation, since canonical CNN involves no explicit spatial rotation operator, it can merely produce local rotation-invariant features by a deep hierarchy of pooling layers. Laptev et al. [27] improved the rotation-invariance of CNN features by proposing a transformation-invariant pooling operation. Jaderberg et al . [28] proposed a spatial transformation network (STN) module to transform the intermediate features in a discriminative manner. Cohen et al. [29] leveraged a smaller symmetric group transformation including a flip and four 90° rotations. The invariant features can be approximated by pooling within the group. Nevertheless, the above approaches are inappropriate for PAA NC-ATR.

Alternatively, in addition to learning invariant features with the discriminative model, the research of discovering equivariant features with generative learning has also attracted increasing attention [30]. It assumes that the input sample is latently controlled and generated by some independent factors. The variations among samples are essentially caused by their change, which is given by:

$$\mathbf{x}_j = \mathcal{G}^*(\mathbf{z}_j) = \mathcal{T}_{i\to j}(\mathbf{x}_i) = \mathcal{T}_{i\to j}(\mathcal{G}^*(\mathbf{z}_i)) = \mathcal{G}^*(\hat{\mathcal{T}}_{i\to j}(\mathbf{z}_i)) \tag{3}$$

where $\mathcal{G}^*$ represents the underlying data generation process entangling the semantically explanatory factors/features $\mathbf{z}$, and $\hat{\mathcal{T}}_{i\to j}$ is a corresponding transformation in the latent space. Compared with the invariant property in Eq. (2), Eq. (3) illustrates an equivariant property that converts the sample transformation $\mathcal{T}$ information into $\hat{\mathcal{T}}$ in the latent domain. By extracting and disentangling the explanatory factors from $\mathbf{x}$, it is easier to recover more information about the input domain. The factors can generally contribute to downstream tasks and make the model transparent [30]. This paper indicates that the latent target pose features are globally shared across all categories. Although the viewing angles of the targets in the NC categories are incomplete in PAA, the missing information can be learned from the samples in the cooperative categories. Therefore, The learned global pose features will be more generalized. To this end, our solution to PAA NC-ATR is to develop a feature disentanglement model to separate the class-shared pose factors from the class-specific identity ones and transfer the complete azimuth angle knowledge from the samples in cooperative classes to those in NC ones.

Nevertheless, the objective function of a generative model for feature disentangling is more expensive to evaluate and harder to design than Eq. (1) since it focuses on the latent domain without explicit ground-truth, yielding an ill-posed inverse problem. Consequently, priors or biases induced constraints or regularizations are necessary for recovering the underlying desired variables [31]–[35]. Totally speaking, two types of strategies have been devoted to feature disentanglement, namely variational-Bayes inference [36] and adversarial learning [37]. InfoGAN exploits an information-theoretic regularization in the generative adversarial network (GAN) for disentanglement [38]. β-VAE [36] over-penalizes the posterior distribution of the factors in variational auto-encoder (VAE) [39] to push it toward a fully factorized prior distribution for disentangling. To further improve the performance of β-VAE, the total correlation has been used as the regularization, yielding Factor VAE [33] and β-TCVAE [40]. Other metrics such as covariance are also used for disentangled representation learning [41]. Nevertheless, the above models treat each training sample independently, and thus they cannot explicitly capture the semantic interactions among multiple samples, e.g., rotation and identity. It is difficult for the above models to disentangle the factors accounting for the desired semantic information without additional supervision.

C. Contrastive Learning

Contrastive learning is appealing for its success in unsupervised representation learning [30], [42]. As a branch of self-supervised learning, it learns representations based on a pre-task of comparing the similarity and dissimilarity between samples. More specifically, instead of treating each training sample individually, it learns to compare two inputs to extract the representations capturing the similar and dissimilar information in terms of the specified semantic criterion, e.g., category. Then, the inductive bias that the representation of positive samples should be close while that of negative ones should be far away will guide and regularize the learning process without extra supervision. It will not only alleviate the labeling limitation encountered in supervised learning but also provide a scalable and accessible mechanism to specify the desired invariant or equivariant properties of the learned function [43]. In the contrastive learning paradigm, an encoder will be first learned to map two input samples to the representations in a latent space. Although the dimension of the latent space is lower than the input domain, it always takes a large computational effort to measure the similarity. To alleviate this issue, a projection head will further project the representations into an embedding space with a much lower dimensional space equipped with a contrastive loss function for measurement. More detailed information can refer to [43] and references therein.

As aforementioned, measuring the similarity or dissimilarity concerning a semantic concept plays a crucial role in contrastive feature learning. It provides an inductive bias to encode the desired transformation equivariant property into the learned features without manual labeling efforts. Inspired by this advantage, this paper will develop a contrastive feature disentanglement framework by designing the metrics in terms of pose and identity similarity between two SAR images.

III. Proposed Framework

In this section, we will formally develop a contrastive feature disentanglement framework for PAA-ATR. First, we will globally propose a new contrastive feature disentangling model to address the PAA-ATR task. Then, we develop two contrastive regularizations on the identity and pose factors, respectively. Finally, we will propose a regularized encoder-decoder architecture for amortized inference and model learning.

A. Contrastive Feature Disentangling Model

Note from the discriminative learning objective (1) that external knowledge used to teach the machine model to classify a target is its image and identity label. However, the image conceals abstract semantics such as the target pose, while the label symbolized as a one-hot vector reflects neither the individual characteristic of intra-class variances and inter-class similarities nor the semantic concept of this class. In this sense, if the pose-identity patterns of an NC category are severely incomplete in the training set, it is intractable for the above inductive learning scheme to generalize the classification ability to an NC target with a never-see view angle. Alternatively, a human can generalize this ability from the perspective of semantic perception and imagination. Moreover, considering two target images from different categories, the human visual cognition system can also measure the pose similarity of the targets in an implicit disentangling manner. These observations demonstrate an inspiration that samples contain more than the identity, and much more representative information is concealed in the entire training set. Therefore, the most notable idea of this paper is learning from populations, i.e., exploiting the semantic relationship among populations as inductive bias, yielding a new feature learning paradigm with improved generalization. It will be a new solution to PAA NC-ATR task by sharing information across all categories. The complete knowledge of the aspect angles will be extracted and exploited by transferring from cooperative classes to NC ones.

Let $\mathbf{r}$ and $\mathbf{f}$ be the latent target pose (relative to the radar) and identity (class-specific) factors, respectively. Based on the generative process (3), the SAR target image $\mathbf{x}$ is latently controlled by $\{\mathbf{f}, \mathbf{r}\}$ through a nonlinear parametric function as $\mathbf{x} \leftarrow \mathcal{G}_\Theta(\mathbf{f}, \mathbf{r})$, where Θ contains the parameters.[1] Therefore, the pose and identity factors of all training samples, namely $\{\mathbf{r}_i\}_{i=1}^{N_c}$ and $\{\mathbf{f}_i\}_{i=1}^{N_c}$ will reside in two disjoint semantic subspaces, respectively. For identity subspace $\mathcal{F}$, it is can be characterized as a normal Euclidean space. For the pose subspace $\mathcal{R}$, since there are always cooperative targets in the training set, the entire $\{\mathbf{r}_i\}_{i=1}^{N_c}$ can cover this subspace as dense as possible. More formally, these points will encode the complete aspect angle information of $[0, 2\pi]$ even if they are not evenly distributed in the PAA condition. Apparently, $\mathcal{R}$ can be modeled as a manifold. It can be noticed that the union of subspaces $\mathcal{F} \bigcup \mathcal{R}$ essentially forms a feature knowledge pool in which the entire disentangled information is shared for combinational generalization [35]. The manner of decomposition-composition is analogous to knowledge processing and generalization of the human brain. To disentangle the above factors, we intend to teach the model to understand two target images by comparing their similarities and differences in identity and pose, respectively. Accordingly, for two targets $\mathbf{x}_i$ and $\mathbf{x}_j$, the contrastive feature disentangling framework can be initially formulated as the following inverse problem:

$$\min_{\{\mathbf{r}_i, \mathbf{f}_i\}_{i=1}^{N_c}, \Theta} \sum_{c=1}^{C} \Omega_f(\{\mathbf{f}_i\}_{i=1}^{N_c}) + \alpha \sum_{c=1}^{C} \Omega_r(\{\mathbf{r}_i\}_{i=1}^{N_c}), \quad \text{s.t. } \forall i,\ \mathbf{x}_i \leftarrow \mathcal{G}_\Theta(\mathbf{f}_i, \mathbf{r}_i),\ \mathcal{F} \perp \mathcal{R} \tag{4}$$

where $\Omega_f(\cdot)$ and $\Omega_r(\cdot)$ are two contrastive regularization functions, and α is a hyper-parameter for balance. Note from Eq. (4) that it originates from the idea of independent component analysis (ICA) for source separation [44] and robust principal component analysis (RPCA) [45]. As aforementioned, the design of $\Omega_f(\cdot)$ and $\Omega_r(\cdot)$ will play an important role in equivariant feature disentanglement, which will be addressed in the following subsections.

B. Contrastive Identity Factor Regularization

The first regularizer $\Omega_f(\cdot)$ measures the similarities of class-specific factors related to the identity. To characterize this metric, we resort to the second-order energy model for pairwise relational learning [46]. More specifically, $\{\mathbf{f}_i\}_{i=1}^{N_c}$ form a pairwise random field conditioned on their identity labels, and $\Omega_f(\cdot)$ is designed to measure its potential energy.

$$\Omega_f(\{\mathbf{f}_i\}_{i=1}^{N_c}) \equiv \sum_i \omega_u(\mathbf{f}_i; \mathbf{y}_i) + \mu \sum_i \sum_j \omega_p(\mathbf{f}_i, \mathbf{f}_j; \mathbf{y}_i, \mathbf{y}_j) \tag{5}$$

where ω_u and ω_p represent the unary and pairwise potential energy functions, respectively, and μ is a balance parameter. Considering first the unary potential ω_u, it measures the compatibility between a supervised label $\mathbf{y}$ and the identity feature vector $\mathbf{f}$. An intuitive way is to learn a projection head $\mathcal{W}(\cdot)$ to embed $\mathbf{f}$ into the label space where the similarity can be derived from cross-entropy loss function $\mathrm{CE}[\cdot, \cdot]$. It is worth noting that $\mathcal{W}(\cdot)$ should be designed as concisely as possible to encode enough class-specific information in $\mathbf{f}$. An over-parameterized $\mathcal{W}(\cdot)$ can warp a simple non-informative distribution to a complex one, in which case the regularization becomes invalid. Accordingly, it is designed as a one-layer perception with a softmax activator in our model.

$$\omega_u(\mathbf{f}; \mathbf{y}) = \mathrm{CE}[\mathbf{y}, \mathcal{W}(\mathbf{f})] \tag{6}$$

For pairwise potential ω_p, it straightforwardly measures the similarity between $\mathbf{f}_i$ and $\mathbf{f}_j$. The features of positive pairs drawn from the same class $\mathcal{C}^+$ should be similar while otherwise separated. If the dimension of $\mathcal{F}$ is high, a projection head $\mathcal{P}$ is needed to embed the features into a lower dimensional space equipped with a contrastive metric. Three forms of loss functions have been used in current contrast learning models [43], including triplet loss, probabilistic noise contrastive estimation-based loss, and distance functions. We empirically found that the Euclidean or Manhatten distance can both achieve improved performance on the validation set. As a result, we consider the following function.

$$\omega_p(\mathbf{f}_i, \mathbf{f}_j; \mathbf{y}_i, \mathbf{y}_j) = \begin{cases} \|\mathcal{P}(\mathbf{f}_i), \mathcal{P}(\mathbf{f}_j)\|_q^q, & if\ \mathbf{y}_i = \mathbf{y}_j \\ \mathrm{const}, & \mathrm{otherwise} \end{cases} \tag{7}$$

where $q = 1, 2$ for Manhatten and Euclidean distance, respectively.

So far, we have almost finished designing the regularization function for discriminative factors. The unary and pairwise potentials will impose different regularizations on $\mathbf{f}$. More formally, ω_u enables the model to capture identity information, while ω_p encourages the consistency between two intra-class vectors. Nevertheless, we find that ω_p will turn to over-penalization in the early training phase, yielding a catastrophic

[1] It is worth noting that since the model does not consider other informative factors, the mentioned factors cannot be used to accurately reconstruct the image, which will be explained in Section III-B.

collapse. In this case, ω_p rapidly achieves its lower bound by forcing two vectors identical for any inputs. It is reasonable to introduce a delayed mechanism before the model has extracted discriminative factors. In our framework, we use a progressive regularization scheme by dynamically varying the value of μ increasing from 0 to 1 according to the following curriculum schedule:

$$\mu = \frac{1-\exp(-10\rho)}{1+\exp(-10\rho)} \tag{8}$$

where ρ is the ratio of the current epoch to the maximum steps of iterations to reflect the training progress.

C. Pose Contrastive Transformer Based Regularization

For the second regularizer $\Omega_r(\cdot)$, it characterizes the pose factors and measures their similarity on a manifold $\mathcal{R}$. Compared with Ω_f, new challenges occur from the following two perspectives. First, because it is intractable to annotate the entire continuous manifold explicitly and exactly, we cannot construct a supervised unary function to encode the pose information into $\mathbf{r}$ as ω_u does. Second, as $\mathcal{R}$ is a continuous manifold, the general Euclidean metric such as ℓ_1 of ℓ_2 norm is not suitable to measure the similarity of points therein.

To address the first challenge, we especially exploit the deductive-based geometry knowledge as supervision to teach the model to learn the concept of "what is the pose of a target". Considering two pose factors $\mathbf{r}_i$ and $\mathbf{r}_j$, if they have indeed captured the entire pose information, there will be an explicit operator $\mathcal{T}_{\theta_{i\to j}}$ warping $\mathbf{r}_i$ to $\mathbf{r}_j$ as $\mathcal{T}_{\theta_{i\to j}}\mathbf{r}_i = \mathbf{r}_j$, and vice versa. If $\theta_{i\to j}$ continuously varies, $\mathbf{r}_i$ will smoothly travel on the manifold $\mathcal{R}$ via $\mathcal{T}_{\theta_{i\to j}}\mathbf{r}_i$. According to the geometry knowledge, $\mathcal{T}_{\theta_{i\to j}}$ corresponds to an affine (a matrix) transform of the 2D coordinates of $\mathbf{r}_i$[2], followed by sampling and interpolation [28], [47]. The parameters $\theta_{i\to j}$ in the operator will have a clear geometric explanation to measure the pose discrepancy. If $\theta_{i\to j}$ is known as a prior in advance or computed from pose estimation, it can be regarded as a deductive based supervised information to learn a pose feature extractor $\mathcal{E}$ via:

$$\mathcal{E}^* \in \arg\min_{\mathcal{E}} \ell_{\mathcal{R}}(\mathcal{T}_{\theta_{i\to j}} \circ \mathcal{E}(\mathbf{x}_i), \mathcal{E}(\mathbf{x}_j)) \tag{9}$$

where $\ell_{\mathcal{R}}$ is a pairwise metric in $\mathcal{R}$. $\mathbf{r}_i = \mathcal{E}^*(\mathbf{x}_i)$ should be a conceptually pose representation of $\mathbf{x}_i$ following the principle of deductive reasoning. If $\theta_{i\to j}$ is not known, we assume that it can be aware from $\mathbf{r}_i$ and $\mathbf{r}_j$. To this end, the geometric transformation model based $\mathcal{T}_{\theta_{i\to j}}$ is attached with a pose discrepancy aware network to estimate $\theta_{i\to j}$ as $\mathcal{D}_\varphi(\mathbf{r}_i, \mathbf{r}_j) = \theta_{i\to j}$, and φ contains the inductive parameters of the network. As the geometric component does not need to learn, we call the module semi-parametric [23], which inherits the benefit of both deductive and inductive learning analogous to human learning.

Considering the second challenge, instead of explicitly selecting an appropriate $\ell_{\mathcal{R}}$ on the manifold, we will alternatively measure their similarity in another semantically plausible space called imaginary target subspace. It is motivated by the following consideration. If $\mathcal{R}$ and $\mathcal{F}$ have been fully disentangled, we can randomly sample factors from them and create an imaginary target image with the help of $\mathcal{G}_\Theta$. More specifically, let $\{\mathbf{r}_i, \mathbf{f}_i\}$ and $\{\mathbf{r}_j, \mathbf{f}_j\}$ be the pose and identity features of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. As aforementioned, $\mathcal{G}_\Theta$ acting as a contextualisation head [43] will generate a blurry image $\mathcal{G}_\Theta(\mathbf{r}_j, \mathbf{f}_j)$ which is defined as the imaginary counterpart of $\mathbf{x}_j$. If it holds true, $\mathcal{G}_\Theta(\mathcal{T}_{\theta_{i\to j}}\mathbf{r}_i, \mathbf{f}_j)$ should also generate an analogous result. Therefore, we can measure the similarity of $\mathbf{r}_i$ and $\mathbf{r}_j$ in the imaginary target subspace. Nevertheless, a risk may come across once $\mathbf{f}_j$ is not completely disentangled from $\mathbf{r}_j$ and already encodes the pose information. To alleviate the risk of information leaking, we use $\mathbf{f}_i$ instead of $\mathbf{f}_j$ so that the entire information of $\mathbf{x}_j$ is blind to $\mathcal{G}_\Theta$ during the imagination phase. This strategy consolidates the combinational generation ability and further encourages the identity consistency between $\mathbf{f}_i$ and $\mathbf{f}_j$. Theoretically speaking, any identity vector of the same class can be sampled from $\mathcal{F}$ to replace $\mathbf{f}_j$, but this strategy will not be considered in this paper. In summary, Ω_r called Pose Contrastive Transformer Regularizer will be designed as:

$$\Omega_r(\{\mathbf{r}_j\}_{j=1}^{N_c}) \equiv \begin{cases} \sum_i \sum_j \ell\left(\mathcal{G}_\Theta(\mathcal{T}_{\theta_{i\to j}}\mathbf{r}_i, \mathbf{f}_i), \mathbf{x}_j\right), & if\ \mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}^+ \\ \text{const}, & \text{otherwise} \end{cases}$$
$$\text{s.t. } \forall i, j, \theta_{i\to j} = \mathcal{D}_\varphi(\mathbf{r}_i, \mathbf{r}_j),\ \mathbf{x}_j \leftarrow \mathcal{G}_\Theta(\mathbf{r}_j, \mathbf{f}_j) \tag{10}$$

where ℓ is a perception loss function measuring the similarity between an imaginary and a true target. According to the previous analysis, the imaginary target will be blurry with rough orientation and identity rather than details. The general pixel-wise metric for image reconstruction, such as mean square error (MSE) or mean average error (MAE), will lead to over-regularization. It will encode some extra undesired information in $\mathbf{r}$ and $\mathbf{f}$ for accurate reconstruction. Therefore, a relaxed metric tailored for measuring the similarity between an imaginary target and its real counterpart is required.

On one hand, instead of measuring the pixel-wise similarity, we alternatively resort to the relaxed statistic similarity. For the ture SAR target, the magnitude of pixels will follow a non-Gaussian distribution [48], e.g., Gamma law in (11)

$$p(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1}\exp(-bx), \tag{11}$$

where a and b are the distribution parameters, and Γ is the gamma function. We expect the imaginary target to keep distribution consistency with its real counterpart. Accordingly, we choose symmetric Kullback-Leibler (SKL) distance to measure the distribution discrepancy [49]. More formally, let $p_x(x)$ and $p_{\hat{x}}(x)$ be the magnitude distribution of the pixel in the true and imaginary target, respectively. SKL distance is defined as:

$$\begin{aligned} J_D\left(p_x(x), p_{\hat{x}}(x)\right) = \mathrm{KL}[p_x(x)||p_{\hat{x}}(x)] + \mathrm{KL}[p_{\hat{x}}(x)||p_x(x)] \\ = a_{\hat{x}}\left(b_x/b_{\hat{x}} - 1\right) + a_x\left(b_{\hat{x}}/b_x - 1\right) \\ + \left(a_x - a_{\hat{x}}\right)\left[\ln\left(b_{\hat{x}}/b_x\right) + \Psi\left(a_x\right) - \Psi\left(a_{\hat{x}}\right)\right] \end{aligned} \tag{12}$$

where Ψ is the digamma function.Note from Eq.(12) that the SKL distance essentially measures the difference between two

[2] $\mathbf{r}_i$ will be a set of 2D feature maps in our situation.

pairs of distribution parameters. As a result, they should be estimated from $\hat{\mathbf{x}}$ and $\mathbf{x}$, respectively. In our model, we prefer the following moment estimation in Eq. (13) because it allows us to compute the derivation of SKL (12) with respect to $\hat{\mathbf{x}}$. The derivation can be further backpropagated to model parameters in an end-to-end learning paradigm.

$$b \leftarrow \mathrm{mean}(\mathbf{x})/\mathrm{var}(\mathbf{x}), \;\; a \leftarrow b \cdot \mathrm{mean}(\mathbf{x}) \tag{13}$$

where $\mathrm{mean}(\mathbf{x}), \mathrm{var}(\mathbf{x})$ stand for the mean value and variance of $\mathbf{x}$, respectively. More specifically, Eq. (13) can be implemented as a module/layer stacked on top of $\mathcal{G}_\Theta(\mathbf{r}, \mathbf{f})$. It will be a supplementary part of the contextualisation head to transform $\mathcal{G}_\Theta(\mathbf{r}, \mathbf{f})$ to the estimated parameters a and b. Then these two values will be used to evaluate the SKL (12).

On the other hand, SKL accounting for the distribution similarity reflect no spatial structure information. Any a repermutation of the pixels will yield the same SKL. To avoid this ambiguity, we additionally involve a second-order statistic metric, namely the Pearson correlation coefficient measuring the orientation similarity as:

$$r(\mathbf{x}, \hat{\mathbf{x}}) = \frac{R\mathbf{x}^T\hat{\mathbf{x}} - \sum_{i=1}^{R} x_i \sum_{i=1}^{R} \hat{x}_i}{\sqrt{R\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{R\sum_i \hat{x}_i^2 - (\sum_i \hat{x}_i)^2}} \tag{14}$$

where R is the dimension of $\mathbf{x}$ or $\hat{\mathbf{x}}$. We finally combine them to form the loss ℓ in (10) as:

$$\ell\left(\mathbf{x}_j, \hat{\mathbf{x}}_j\right) = J_D\left(p_{\mathbf{x}_j}(x), p_{\hat{\mathbf{x}}_j}(x)\right) - r(\mathbf{x}_j, \hat{\mathbf{x}}_j) \tag{15}$$

D. Amortized Inference and Overall Architecture

So far, we have developed a contrastive feature disentangling model (ConFeDent) (4), where we elaborate two regularization functions (5) and (10) to inject the identity and pose information into the factors $\mathbf{f}$ and $\mathbf{r}$, respectively. More specifically, Ω_f imposes a discriminative and structural regularization on the pairwise intra-class identity factors. It will not only encode the class-specific information in $\mathbf{f}$ but also encourage the pairwise similarity. Alternatively, Ω_r involves a semi-parametric geometric transformation model to characterize the relationship between two points on the manifold $\mathcal{R}$ with deductive geometric knowledge. More importantly, we construct the imaginary target subspace combinational generalization. Instead of designing a metric on the manifold, we measure the similarity between two pose factors in the imaginary target subspace with an elaborated loss (15), which can further promote feature disentanglement and generalization ability.

Thanks to these strategies, ConFeDent will be more semantic explainable than the other deep learning-based SAR ATR algorithms. Nevertheless, it is difficult to solve the problem (4) with a general algorithm because it is essentially a bi-level optimization, and the variables are coupled. Moreover, feature disentanglement is an inverse problem which generally requires iterative optimization process [44], [45]. To address these issues, inspired by the recently prevalent amortized inference technique [44], we resort to learning an optimizer that directly produces the approximated solutions, yielding an encoder-decoder feature disentanglement architecture. To this end, concerning first the latent features $\mathbf{f}$ and $\mathbf{r}$, we declare that their optimal values can be obtained by a parameterized encoder $\mathcal{E}_\Psi$ as $\mathbf{f}, \mathbf{r} = \mathcal{E}_\Psi(\mathbf{x})$, where Ψ contains the encoder parameters. $\mathcal{E}_\Psi$ can be also regarded as a feature extractor that factorizes the image into the regularized identity and pose factors, respectively. Therefore, the final optimization problem for ConFeDent learning will be summarized as:

$$\begin{aligned} \min_{\Psi,\Theta,\varphi} \sum_{c=1}^{C} \Omega_f(\{\mathbf{f}_i\}_{i=1}^{N_c}) + \alpha \sum_{c=1}^{C} \Omega_r(\{\mathbf{r}_i\}_{i=1}^{N_c}) \\ +\beta \sum_{i}\sum_{j \neq i} \ell(\mathbf{x}_{i,j}, \mathcal{G}_\Theta(\mathbf{f}_{i,j}, \mathbf{r}_{i,j})), \\ \text{s.t. } \theta_{i\to j} = \mathcal{D}_\varphi(\mathbf{r}_i, \mathbf{r}_j), \;\; \mathbf{f}_i, \mathbf{r}_i = \mathcal{E}_\Psi(\mathbf{x}_i), \;\; \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}^+, \end{aligned} \tag{16}$$

where the bi-level optimization (4) is relaxed as an unconstrained one, and β is a hyper-parameter. This relaxed optimization can be efficiently solved by a standard gradient-based method. Several remarks should be highlighted for the training phase. First, we can sample at most $\mathcal{A}_{N_c}^2$ target pairs from each class to train the model, where $\mathcal{A}$ is the permutation operator. In this regard, ConFeDent will be more appropriate for learning with limited training samples because of its improved generalization ability. We encourage the model to utilize data pairs from cooperative categories at the beginning of the training phase to cover the latent pose manifold. Second, as $\theta_{i\to j}$ and $\theta_{j\to i}$ are reciprocal, it is also suggested to reverse $\mathcal{T}_{\theta_{i\to j}}$ to obtain $\mathcal{G}_\Theta(\mathcal{T}_{\theta_{j\to i}}\mathbf{r}_j, \mathbf{f}_j)$ for supplemental regularization.

In the deployed testing phase, we remove the projection head $\mathcal{P}$ in Ω_f and the pose contrastive transformer Ω_r. Then ConFeDent becomes a standard fully convolution network (FCN) architecture that can be universally deployed in a general SAR ATR task. For a query sample $\tilde{\mathbf{x}}$, its label vector $\tilde{\mathbf{y}}$ can be efficiently obtained as:

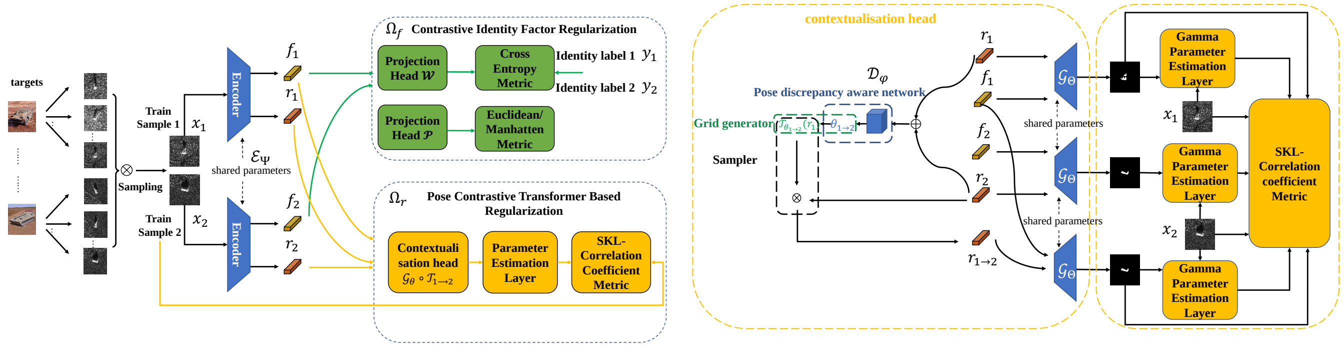
$$\tilde{\mathbf{y}} = \mathcal{W}^*(\tilde{\mathbf{f}}), \;\; \text{s.t. } \tilde{\mathbf{f}} \leftarrow \mathcal{E}_{\Psi^*}(\tilde{\mathbf{x}}) \tag{17}$$

The overall architecture of ConFeDent for reproduction is illustrated in Fig. 2, and the detailed network information of each module is summarized in Table II.

E. A Strengthened Version: ConFeDent+

In ConFeDent, the most notable component is the design of pose contrastive transformer regularizer which aims to warp the pose features $\mathbf{r}_i$ of $\mathbf{x}_i$ to $\mathbf{r}_j$ of $\mathbf{x}_j$ via $\mathcal{T}_{\theta_{i\to j}}$. Instead of directly measuring the similarity in the pose manifold $\mathcal{R}$, we alternatively consider implementing this task in the imaginary target image domain with the help of $\mathcal{G}_\Theta$ as we have the real counterpart of the imaginary target. It follows that the training samples $(\mathbf{x}_i, \mathbf{x}_j)$ has to be restricted as an intra-class pair for self-supervision, which, however, reduces the sufficient utilization of these scarce training data. Additionally, knowledge sharing among different categories proceeds implicitly.

The crucial point of the above limitation is the lack of the observed real NC target corresponding to the imaginary one in the training set. It is not possible to supervise the imaginary


targets
Sampling
Train Sample 1
Train Sample 2
x_1
x_2
Encoder
Encoder
$\mathcal{E}_\Psi$
shared parameters
f_1
r_1
f_2
r_2
Ω_f Contrastive Identity Factor Regularization
Projection Head $\mathcal{W}$
Cross Entropy Metric
Identity label 1 y_1
Identity label 2 y_2
Projection Head $\mathcal{P}$
Euclidean/ Manhatten Metric
Ω_r Pose Contrastive Transformer Based Regularization
Contextualisation head $\mathcal{G}_\theta \circ \mathcal{T}_{1\rightarrow 2}$
Parameter Estimation Layer
SKL-Correlation Coefficient Metric
contextualisation head
$\mathcal{D}_\varphi$
Pose discrepancy aware network
Grid generator
$\mathcal{T}_{\theta_{1\rightarrow 2}}(r_1)$
$\theta_{1\rightarrow 2}$
Sampler
$r_{1\rightarrow 2}$
$\mathcal{G}_\Theta$
shared parameters
Gamma Parameter Estimation Layer
SKL-Correlation coefficient Metric

(a) Overall framework of ConFeDent

(b) Pose Constrastive Transformer

Fig. 2. The framework of the proposed ConFeDent.

TABLE II
DETAILED INFORMATION OF EACH MODULE OF CONFEDENT

Module	Architecture	Input	Output
$\mathcal{E}_\Psi$	Conv:K5-S1-C16-ReLU MaxPool:P2-S2 Conv:K5-S1-C32-ReLU MaxPool:P2-S2 Conv:K5-S1-C64-ReLU MaxPool:P2-S2 Dropout Conv:K5-S1-C128-ReLU	$\mathbf{x}$	$\mathbf{f}$
	Conv:K5-S1-C16-ReLU MaxPool:P2-S2 Conv:K5-S2-C64-ReLU MaxPool:P2-S2 Dropout Conv:K5-S2-C128-Tanh	$\mathbf{x}$	$\mathbf{r}$
$\mathcal{D}_\varphi$	FC:C60-ReLU FC:C30-ReLU FC:C6	$\mathbf{r}_1$,$\mathbf{r}_2$	$\theta_{1\rightarrow 2}/\theta_{2\rightarrow 1}$
$\mathcal{G}_\Theta$	ConvT:K5-S1-C64-valid-ReLU UpSampling:P2-S2 ConvT:K6-S1-C32-valid-ReLU UpSampling:P2-S2 ConvT:K5-S1-C16-valid-ReLU UpSampling:P2-S2 ConvT:K5-S1-C1-valid-ReLU	$\mathbf{f}$,$\mathbf{r}$	$\mathcal{G}_\Theta(\mathbf{f},\mathbf{r})$
$\mathcal{P}$	Identity Mapping	$\mathbf{f}$	$\mathcal{P}(\mathbf{f})=\mathbf{f}$
$\mathcal{W}$	FC:CC-Softmax	$\mathbf{f}$	$\mathbf{y}$

NC target image. To address this limitation, we propose a strengthened version called ConFeDent+ which can effectively exploit arbitrary pair of samples for self-training. The motivation of ConFeDent+ is that a robust target imagination should be bidirectional. Specifically, ConFeDent can forward imaging a target object and thus can also backtrace to the original real target from the imaginary counterparts. Let $\mathbf{x}_i^1,\mathbf{x}_j^2$ be a sample pair from class 1 and 2, respectively. Their pose and identity features can be obtained by the well-trained ConFeDent as $\mathbf{r}_i^1,\mathbf{f}_i^1 \leftarrow \mathcal{E}_{\Psi^*}(\mathbf{x}_i^1)$, and $\mathbf{r}_j^2,\mathbf{f}_j^2 \leftarrow \mathcal{E}_{\Psi^*}(\mathbf{x}_j^2)$, respectively. Then ConFeDent can generate two imaginary targets $\hat{\mathbf{x}}_j^1$ and $\hat{\mathbf{x}}_i^2$ which stand for the target in class 1,2 with the approximate pose of $\mathbf{r}_j$, $\mathbf{r}_i$, respectively. Based on the above motivation, ConFeDent should also imagine $\mathbf{x}_i^1,\mathbf{x}_j^2$ from $\hat{\mathbf{x}}_j^1$ and $\hat{\mathbf{x}}_i^2$ in a backtrack way. To this end, $\hat{\mathbf{x}}_j^1$ and $\hat{\mathbf{x}}_i^2$ will be treated as two latent variables which are fed into a next ConFeDent model with shared parameters in the hope to output the reconstruction of the original $\mathbf{x}_i^1,\mathbf{x}_j^2$. Therefore, ConFeDent+ comprises of two parameter-shared ConFeDent in series which is shown in Fig. 3. In the second ConFeDent, we remove the semi-parametric geometric transformation module and pairwise combine the features of $\hat{\mathbf{x}}_j^1$ and $\hat{\mathbf{x}}_i^2$ with $\mathbf{r}_i^1,\mathbf{f}_i^1$, and $\mathbf{r}_j^2,\mathbf{f}_j^2$ to feed into $\mathcal{G}_\Theta$, yielding four outputs. In ConFeDent+, we compute the MAE loss between $\mathbf{x}_i^1,\mathbf{x}_j^2$ and four outputs instead of imaginary loss (15). Compared with the original ConFeDent in Table II, we slightly modify several building blocks to construct ConFeDent+. Specifically, as ConFeDent+ is a deeper architecture, batch normalization layers are inserted in the branch of pose factor extractor in $\mathcal{E}_\Psi$ after each convolution layer for stable training. Additionally, we insert a “ConvT:K1-S1-C128-valid-ReLU” layer at the beginning of $\mathcal{G}_\Theta$. The testing phase of ConFeDent+ is the same with ConFeDent.

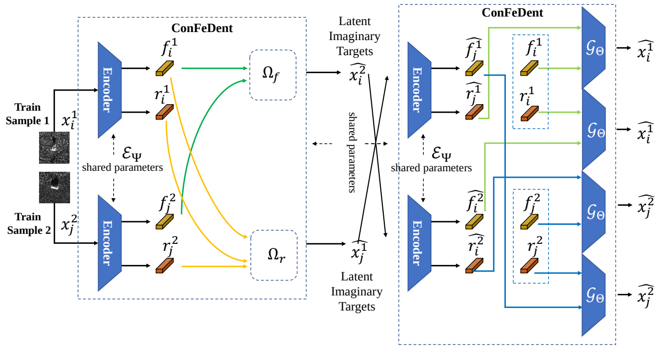


Fig. 3. The framework of the proposed ConFeDent+.

IV. EXPERIMENTS AND ANALYSIS

This section will conduct extensive experiments to evaluate the effectiveness and superiority of ConFeDent and ConFeDent+ for SAR ATR, especially PAA NC-ATR. First, we will introduce the experimental dataset of moving and stationary target acquisition and recognition (MSTAR) and comparison algorithms. Second, we design some ablation experiments to validate the effectiveness of the proposed strategies in our framework. Finally, we compare the proposed method with other SAR ATR models to demonstrate its superiority. The overall experiments are performed several times and report the average results.

A. Experimental Settings

1) Dataset Setting: MSTAR dataset, collected with the SAR sensor operating at X-band by the Sandia National Laboratory, is the most widely exploited benchmark for SAR ATR [2]. It contains about ten types of publicly released military ground targets obtained at multiple depression angles and $0 \sim 360°$ aspect angles with approximately $5°$ interval, and $0.3m \times 0.3m$ image resolution, including armored personnel carrier: BMP-2, BRDM-2, BTR-60, and BTR-70; tank: T-62, T-72; rocket launcher: 2S1; air defense unit: ZSU-234; truck: ZIL-131; bulldozer: D7. Their optical images are shown in Figs. 4 only for illustration. Following the common setting [14], the central cropped 88×88 magnitude image patches will be exploited for training, validating, and testing.

The following experiments will validate the performance of each SAR algorithm under two typical SAR geometric acquisition conditions, i.e., SOC and EOC, each of which will incorporate the PAA training setting to increase the difficulty. More specifically, the canonical SOC and EOC-D illustrated in Table III and Table IV consider identical serial numbers and configurations with slight and large depression angle variation between training and testing targets, respectively. Considering the PAA NC-ATR setting, only a part of the samples with $17°$ depression angle will be selected in the training set according to the following operations. First, we randomly choose m and n categories as the cooperative and NC ones, respectively, denoted by CmNn. Next, NC targets with $17°$ depression angle will be divided into two groups according to the aspect angle, e.g., $[0, \pi)$ and $[\pi, 2\pi]$ in our experiments. We randomly select only one group of samples termed aspect available NC targets (ANT) into the training set for each NC category, leaving one sample in the validating set according to the leave-one-out strategy. The rest group termed aspect unavailable NC targets (UNT) will be discarded. As a result, about 50% NC targets are used for training. Finally, for cooperative categories, we uniformly select 50% samples with full aspect angle into the training set for class balance consideration, and the others will be added to the validating set. Throughout the experiments, we do not conduct any data augmentation procedure for all algorithms unless specially declared. The targets with 15% depression angle will be used for testing without any change. Nevertheless, to demonstrate the NC-ATR performance clearly, the testing set will be correspondingly divided into the following three groups, i.e., targets in the cooperative classes Test_{CT}, NC targets with seen aspect Test_{ANT} and NC targets with unseen aspect Test_{UNT}. It is worth noting that we should pay more attention to the recognition performance on Test_{UNT} since it requires more generalization ability. Fig. 5 illustrates the above dataset splitting process.

2) The Experimental Settings of ConFeDent: The parameters in ConFeDent are initially in a default way without pretraining. The standard momentum-based stochastic gradient descent is exploited for model learning, and the learning rate and the momentum rate are 0.001 and 0.9, respectively. The batch size is 100. The weight decay regularization with a rate of 0.004 is imposed on model parameters except φ. We use the early stopping strategy to control the training process by monitoring the recognition accuracy of the validating set. The training process will stop if the recognition accuracy has no improvement within 30 epochs. α and β are 1.5 and 1 according to the grid search.

(a) BMP-2 (b) BTR-70 (c) T-72 (d) 2S1 (e) BRDM-2
(f) BTR-60 (g) D7 (h) T-62 (i) ZIL-131 (j) ZSU-234

Fig. 4. Optical sample images of ten types of target in MSTAR dataset only for illustration.

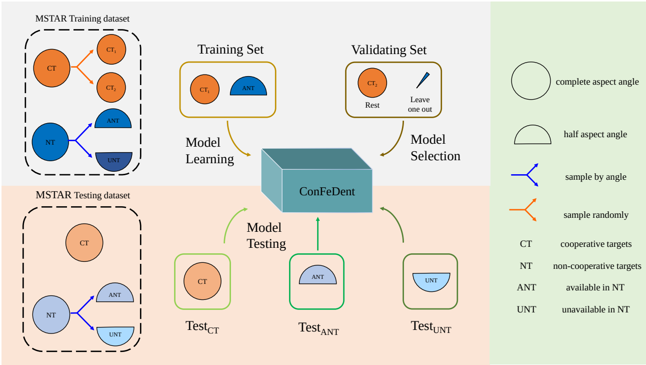


Fig. 5. Illustration the procedure of dataset splitting.

B. Ablation Experiments

In this subsection, we will design some specific experiments via ablation to demonstrate the effectiveness of our proposed strategies in ConFeDent. The experiments in this subsection will be carried out under SOC-C5N5 without loss of generality.

1) Validation of Pose Contrastive Transformer: The most crucial strategy in ConFeDent to improve the generalization ability is the design of Ω which will warp the pose features of a target $\mathbf{x}_i$ to that of the other intra-class target $\mathbf{x}_j$ with the semi-parametric geometric transformation model. To evaluate its ability, we randomly select 4 pairs of intra-class targets and illustrate the visual illustration of the imaginary target images $\mathcal{G}_\Theta(\mathbf{r}_{1,2}, \mathbf{f}_{1,2})$, $\mathcal{G}_\Theta(\mathcal{T}_{\theta_{2\to1}}\mathbf{r}_2, \mathbf{f}_2)$, and $\mathcal{G}_\Theta(\mathcal{T}_{\theta_{1\to2}}\mathbf{r}_1, \mathbf{f}_1)$ in Figs. 6. We can see that all imaginary images are the blurry version of the original inputs with the same orientation. The aspect angle of $\mathcal{G}_\Theta(\mathcal{T}_{\theta_{2\to1}}\mathbf{r}_2, \mathbf{f}_2)$ and $\mathcal{G}_\Theta(\mathcal{T}_{\theta_{1\to2}}\mathbf{r}_1, \mathbf{f}_1)$ are identical to that of $\mathcal{G}_\Theta(\mathbf{r}_1, \mathbf{f}_1)$ and $\mathcal{G}_\Theta(\mathbf{r}_2, \mathbf{f}_2)$, respectively, though the aspect angle discrepancy between $\mathbf{x}_1$ and $\mathbf{x}_2$ are large. Therefore, these results can demonstrate the effectiveness of the proposed model for disentangling the pose factors among intra-class samples and cross-transformation.

2) Evaluating the Strategies in Ω_f and Ω_r: We have introduced a dynamic hyper-parameter μ in Ω_f to avoid over-

TABLE III
TARGET INFORMATION OF SOC

Class	Train(Depression=17°)		Test(Depression=15°)	
	Aspect		Aspect	
	$[0, \pi)$	$[\pi, 2\pi]$	$[0, \pi)$	$[\pi, 2\pi]$
2S1	145	154	137	137
BMP-2	354	344	310	277
BRDM-2	145	153	136	138
BTR-70	108	125	107	89
BTR-60	136	120	103	92
D7	152	147	138	136
T-62	147	152	134	139
T-72	343	348	307	275
ZIL-131	146	153	136	138
ZSU-23-4	145	154	136	138

TABLE IV
TARGET INFORMATION OF EOC-D

Class	Train(Depression=17°)		Test(Depression=30°)	
	Aspect		Aspect	
	$[0, \pi)$	$[\pi, 2\pi]$	$[0, \pi)$	$[\pi, 2\pi]$
2S1	145	154	140	148
BRDM-2	145	153	138	149
T-72	118	114	138	150
ZSU-23-4	145	154	139	149

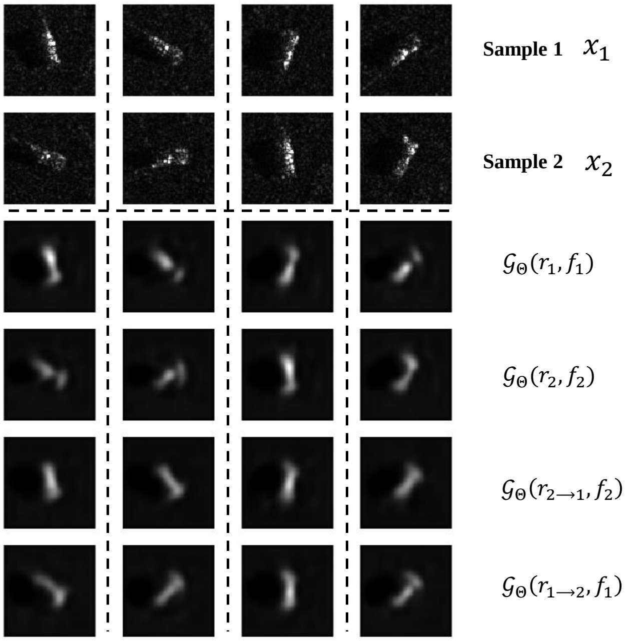

Sample 1 x_1
Sample 2 x_2
$\mathcal{G}_\Theta(r_1, f_1)$
$\mathcal{G}_\Theta(r_2, f_2)$
$\mathcal{G}_\Theta(r_{2\to1}, f_2)$
$\mathcal{G}_\Theta(r_{1\to2}, f_1)$

Fig. 6. Illustration of intra-class cross imagination ability of ConFeDent.

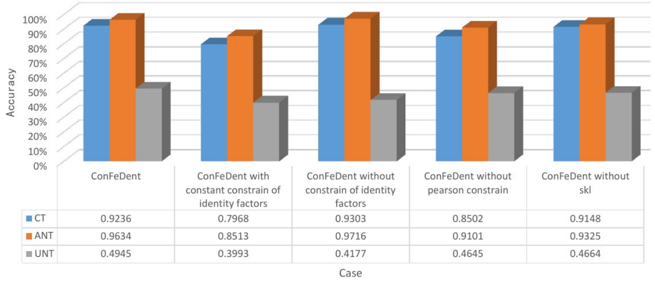


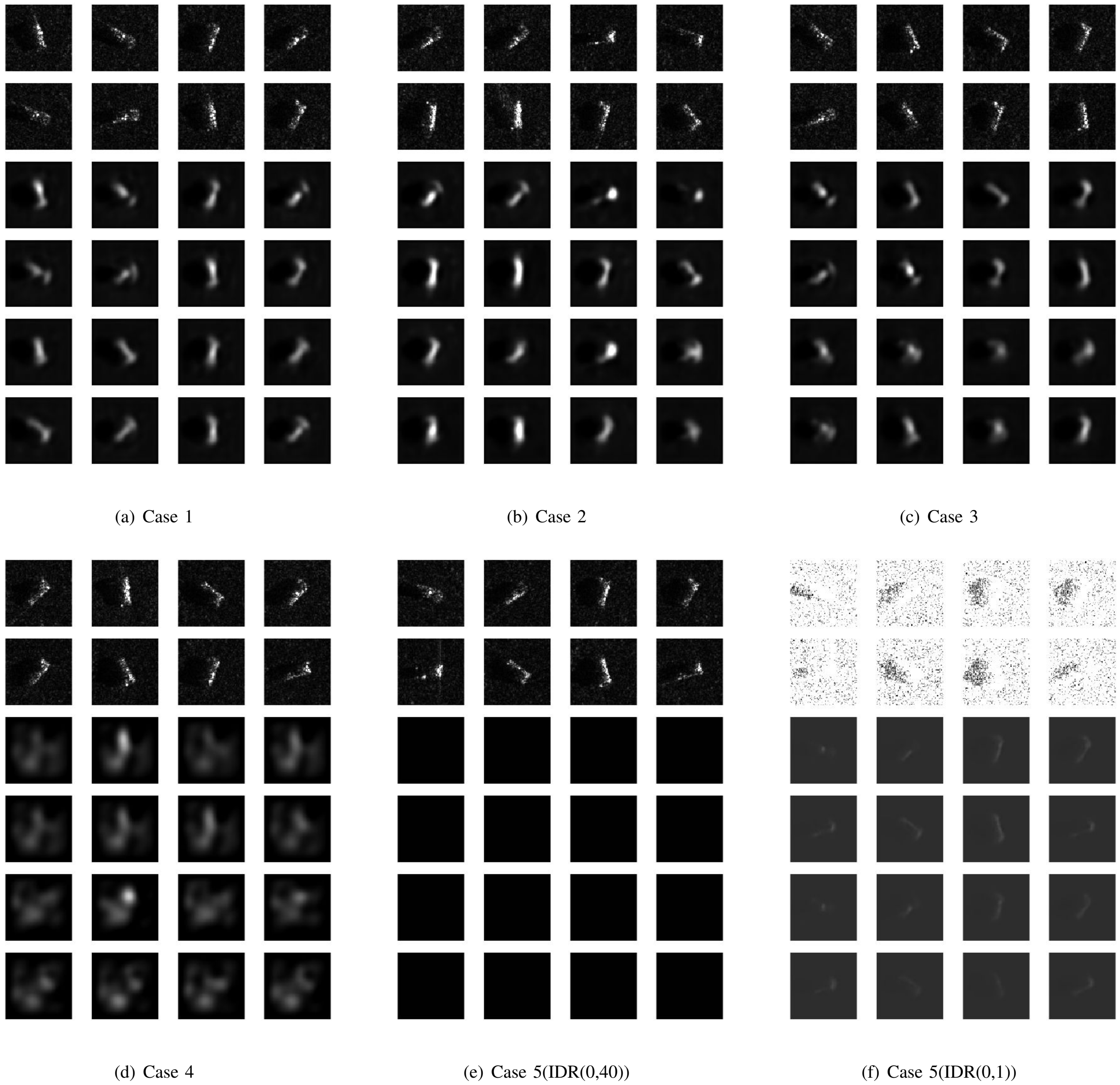
Fig. 7. The classification accuracy of ablation experiments in SOC-C5N5.

penalization and utilized a curriculum schedule to progressively increase μ according to (8). In Ω_r, we elaborate a new similarity metric (15) which combines the first-order SKL and second-order Pearson correlation coefficients functions. It measures the pose similarity between a blurry imaginary SAR target and its real counterpart. To verify the effectiveness of the above strategies, we conduct several experiments to evaluate the impacts on classification accuracy and visual results under following conditions: 1) the standard ConFeDent, 2) specified $\mu \neq 0$ via cross-validation, 3) $\mu = 0$, 4) removing Pearson correlation from (15) and 5) removing SKL from (15). The corresponding results are summarized in Fig. 7 and Fig. 8. Let us observe the classification accuracies in Fig. 7. The accuracy of Case 2 will always be lower than Case 1 and Case 3, which is caused by over-penalization of the pairwise potential function. Comparing Case 1 with Case 3, although Case 3 can achieve slightly higher accuracy on CT and ANT, it performs much worse on UNT. Comparing the results of Case 1, Case 4, and Case 5, we find that the Pearson correlation coefficient and the SKL distance will both help to improve the classification accuracy. These results demonstrate that the ATR will benefit from feature disentanglement because of the improved generalization ability. Let us observe the visual results shown in Figs. 8. Comparing Fig. 8(a) with Figs. 8(b) and 8(c), there is no obvious variation of the resulting images. Nevertheless, the cross-transformations of the first, second the fourth target pairs in Fig. 8(c) by removing the pairwise term are failed ones, which implies its influence on pose factor disentanglement. Alternatively, in Fig. 8(d), removing the Pearson correlation coefficient will generate more blurry images which cannot recognize any orientation information. Finally, in Fig.8(e) and Fig.8(f), removing the SKL distance will generate darker images which cannot obtain any useful information. For better illustration, we normalize the pixel values to the range [0,1] for contrast enhancement. We can see that the imaginary targets are more compact and complete, but the corresponding pixel values are low. It means that the statistical distribution characteristics are different from the original image. In summary, SKL and Pearson correlation will play distinct roles in feature regularization.

C. Algorithms Comparison on SAR ATR

In this subsection, we will compare the performance of ConFeDent and other algorithms on the SAR NC-ATR task under different conditions. Firstly, comparison experiments will investigate the general SAR ATR performance under the SOC with reduced training samples. Next, we combine the PAA setting with SOC and EOC-D to evaluate the ATR performance. Finally, algorithms will be compared on much more complicated PAA conditions where the available range of aspect angles of NC are only $[0, 2\pi/3]$ and $[0, \pi/2]$, respectively.

1) Comparison Results Under SOC: The first comparison experiment will be conducted on SOC, whose training and testing samples are the original ones in Table. III without PAA consideration. To demonstrate the effectiveness and superiority



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

(e) Case 5(IDR(0,40))

(f) Case 5(IDR(0,1))

Fig. 8. Ablation experiments on different the strategies in Ω_f and Ω_r. IDR is an abbreviation for Image Display Range.

of ConFeDent and ConFeDent+ for general SAR-ATR, we compare the results with those of the following typical models: A-ConvNet [14], RotANet [21], K-nearest neighbor (K-NN), and sparse representation classifier (SRC) [9]. A-ConvNet achieves state-of-the-art accuracy with a standard FCN architecture, and thus it is treated as a baseline SAR ATR algorithm. We also extend A-ConvNet with several strategies to improve its generalization ability. First, we exploit the data augmentation strategy to manually generate the pseudo targets with full aspect angles for NC categories. To this end, the standard data augmentation strategy in the Tensorflow library will rotate each sample to augment the pseudo targets with unseen aspect angles. They will be exploited to train A-ConvNet, yielding a variant denoted by A-ConvNet*. Additionally, several prevalent deep network models for nature image classification will be also compared via fine-tuning the pre-trained models, including ResNet50 [50], VGG [51]. In the experiment, we vary the training sampling rate from 10% to 50% with a 10% interval considering a small sample situation. The comparison recognition accuracy will be plotted in Fig. 9. As can be seen from the figure that ConFeDent and ConFeDent+ can outperform other compared algorithms in most cases. More importantly, the accuracy curves of ConFeDent+ is more flat than others, which implies its robustness to reduced amount of training samples.

2) Recognition Results Under PAA-SOC: The following experiments will validate several situations accounting for training with different numbers of NC categories, including C5N5, C1N9, and C0N10 according to the setting in Sec.

TABLE V

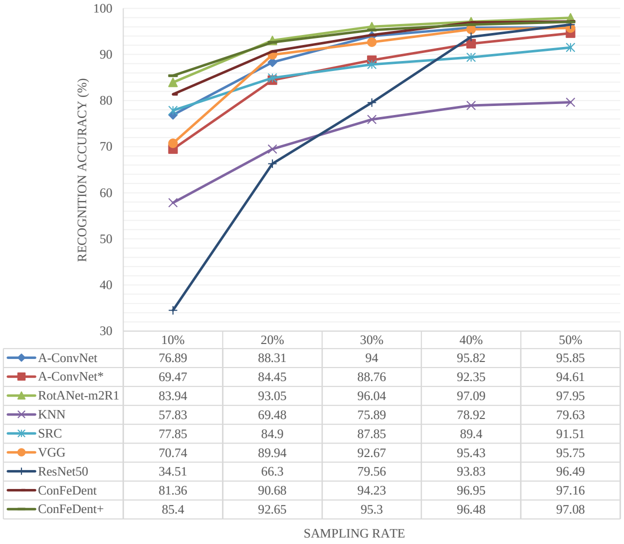
RECOGNITION AVERAGE ACCURACY OF DIFFERENT ALGORITHMS UNDER PAA-SOC

Methods	C5N5			C1N9			C0N10	
	CT	ANT	UNT	CT	ANT	UNT	ANT	UNT
SRC	91.06%	95.99%	31.88%	94.83%	96.31%	33.43%	96.50%	36.11%
A-ConvNet	89.77%	92.83%	23.67%	***96.13%***	***97.15%***	31.07%	***97.29%***	33.98%
A-ConvNet*	91.49%	96.05%	28.29%	90.95%	94.72%	27.44%	96.04%	29.43%
A-ConvNet+STN	92.21%	***97.39%***	38.92%	93.46%	95.79%	32.76%	95.70%	33.77%
InceptionV3	77.97%	83.85%	38.62%	80.08%	82.23%	41.02%	84.90%	39.21%
ResNet50	64.38%	71.85%	32.33%	65.85%	75.40%	36.51%	76.34%	37.49%
VGG16	87.94%	93.24%	39.39%	82.10%	87.52%	38.49%	95.11%	37.80%
ConFeDent	***92.36 %***	96.34%	***49.45%***	90.92%	94.00%	***43.78%***	96.96%	***43.86%***
ConFeDent+	**93.84 %**	**97.86%**	**56.73%**	**97.27%**	**97.19%**	**49.10%**	**97.86%**	**55.87%**

TABLE VI

RECOGNITION AVERAGE ACCURACY OF DIFFERENT ALGORITHMS UNDER PAA-EOC-D

Methods	C3N1			C2N2			C1N3			C0N4	
	CT	ANT	UNT	CT	ANT	UNT	CT	ANT	UNT	ANT	UNT
A-ConvNet	***79.58%***	78.59%	29.22%	***80.54%***	78.24%	35.29%	82.20%	***84.75%***	32.14%	84.94%	40.42%
A-ConvNet*	76.32%	***89.05%***	55.39%	75.71%	***88.13%***	46.31%	76.96%	83.17%	42.04%	81.57%	38.88%
A-ConvNet+STN	76.71%	83.56%	51.18%	**80.94%**	78.67%	44.98%	73.81%	71.31%	44.37%	**85.77%**	***53.97%***
InceptionV3	59.16%	58.70%	47.21%	58.38%	58.86%	44.51%	54.09%	60.72%	49.20%	60.99%	48.73%
ResNet50	41.96%	38.93%	29.91%	36.37%	27.79%	25.87%	37.75%	32.28%	29.66%	28.07%	25.53%
VGG16	69.25%	80.44%	48.21%	64.10%	73.07%	50.91%	66.22%	79.22%	49.22%	61.51%	46.87%
ConFeDent	75.42%	88.28%	***73.28%***	74.98%	84.10%	***58.66%***	***82.58%***	76.83%	***49.65%***	68.08%	53.36%
ConFeDent+	***76.83%***	**94.44%**	**79.45%**	74.83%	**91.40%**	**68.34%**	**87.93%**	**88.68%**	**54.88%**	***85.16%***	**54.87%**



	10%	20%	30%	40%	50%
A-ConvNet	76.89	88.31	94	95.82	95.85
A-ConvNet*	69.47	84.45	88.76	92.35	94.61
RotANet-m2R1	83.94	93.05	96.04	97.09	97.95
KNN	57.83	69.48	75.89	78.92	79.63
SRC	77.85	84.9	87.85	89.4	91.51
VGG	70.74	89.94	92.67	95.43	95.75
ResNet50	34.51	66.3	79.56	93.83	96.49
ConFeDent	81.36	90.68	94.23	96.95	97.16
ConFeDent+	85.4	92.65	95.3	96.48	97.08

Fig. 9. Comparison recognition accuracy under SOC.

IV-A. In addition to the aforementioned algorithm, STN is a plug-and-play module to produce the spatial transformation-invariant features [28]. We insert an STN module into A-ConvNet, yielding A-ConvNet+STN for comparison. The detailed classification results of all algorithms are summarized in Table V, where the top 2 results in each case is respectively denoted with **bold** and ***italic*** fonts. Checking the results, most algorithms can obtain satisfactory accuracy on CT and ANT in three situations. In particular, ConFeDent+ obtains top-1 accuracy in all cases. For UNT in three cases, ConFeDent+ can improve the accuracy of the rank-2 ConFeDent by $5\% \sim 12\%$, and ConFeDent will further improve the accuracy in UNT by 10% on average than others. This phenomenon firmly demonstrates the superiority of ConFeDent and ConFeDent+ in addressing PAA NC-ATR. Comparing the accuracy of ConFeDent and A-ConvNet on three UNT, we can see that ConFeDent can achieve 25.78%, 12.71%, and 9.88% improvement. It can somewhat demonstrate that ConFeDent can indeed benefit from the increased amount of cooperative samples. The result of A-ConvNet* trained with the manually generating pseudo targets is lower than A-ConvNet. It verifies that the data augmentation trick used in nature image classification may be unsuitable for SAR ATR anymore. We can conclude from the accuracy of A-ConvNet+STN that the STN module can somewhat improve the performance of A-ConvNet in most cases. However, due to different mechanisms and motivations, the generalization ability of STN on UNT is still weaker than our proposed ConFeDent, and it is less efficient for STN to address the o.o.d classification task. For three prevalent pre-trained deep learning architectures, VGG can achieve better performance than the other two. In the extreme case of C0N10 where there is no cooperative category, we assume that the aspect angles of entire NC categories can fill the range of $[0, 2\pi]$. According to the result, ConFeDent and ConFeDent+ can also outperform the other algorithms, which further validates the superiority of knowledge sharing.

3) Comparison Results of PAA-EOC-D: Next, we will conduct the comparison experiments PAA-EOC-D. In this combined condition, the training and testing samples will appear larger depression angle (17° versus 30°), yielding a more challenging o.o.d classification task than PAA-SOC. Since the number of all target categories in EOC-D is less

TABLE VII

COMPARISON ACCURACY IN SMALLER RANGE OF ASPECT ANGLES ON PAA-SOC C5N5

Methods	$[0, 2\pi/3]$			$[0, \pi/2]$		
	CT	ANT	UNT	CT	ANT	UNT
AconvNet	82.19%	97.21%	12.74%	78.82%	94.34%	22.85%
AconvNet*	87.86%	95.27%	16.09%	86.59%	96.69%	14.04%
AconvNet+STN	89.99%	97.53%	***22.07%***	87.93%	95.69%	18.10%
InceptionV3	61.73%	78.79%	18.05%	78.96%	84.96%	28.02%
ResNet50	34.68%	48.99%	18.66%	64.45%	78.56%	24.58%
VGG16	86.69%	93.73%	14.53%	90.18%	95.11%	26.52%
ConFeDent	***91.01%***	***97.63%***	17.67%	***92.50%***	**98.06%**	***29.41%***
ConFeDent+	**92.10%**	**97.84%**	**25.63%**	**93.63%**	***97.67%***	**36.97%**

than PAA-SOC, we can exploit $\sum_{n=1}^{4} C_4^n$ combinations of the NC class from C0N4 to C3N1 in total. The comparison results are summarized in Table VI. During experiments, we specially observed a new challenging phenomenon that the performance of different NC combinations varies significantly. The recognition accuracy for some combinations is surprisingly unstable for all algorithms, especially ConFeDent and ConFeDent+. We conjecture the reason is that the four types of targets in EOC-D differ a lot. More specifically, 2S1 is a carriage motor howitzer, BRDM-2 is an amphibious armored reconnaissance vehicle, T72 is a main battle tank and ZSU-23-4 is an anti-aircraft vehicle. According to their appearances shown in Figs. 4, T72 and 2S1 are much dissimilar to BRDM-2 and ZSU-23-4. Intuitively their information will be more difficult to be shared, in which case the performance of ConFeDent and ConFeDent+ will be degraded. Then the overall average accuracy will be pulled down. Nevertheless, it can be also observed from the results in Table VI that the ConFeDent+ achieves the best performance in most cases, and ConFeDent can also outperform other compared algorithms under the UNT condition. Under the extreme condition C0N4, our proposed methods can also achieve better performance. Similar to the results of previous experiments in the PAA-SOC condition, the STN module can improve the recognition performance of A-ConvNet under the UNT condition. It proves that our proposed method can effectively utilize the additional information transferred from cooperative targets.

4) Comparison Results of Smaller Range PAA: Finally, we will conduct comparative experiments under more severe conditions by considering smaller ranges of aspect angles. To this end, we test two situations where the range of aspect angles of NC for training will be $\pi/2$ and $2\pi/3$. Accordingly, we can divide the original training samples in Table III into 4 and 3 parts corresponding to the situation of $\pi/2$ and $2\pi/3$, respectively. We randomly pick one part of the samples for actual training for each NC. Through this operation, the corresponding training samples will be approximately 25% and $1/3$ of the original ones. As a typical example, the algorithms will be tested on the SOC-C5N5 dataset once more, and Table VII summarizes the comparison results. Although in the $\pi/2$ case, the accuracy of ConFeDent under UNT is not the highest, it is undoubtedly the best performing method under all conditions combined. In the $2\pi/3$ case, ConFeDent achieves the best classification results under all conditions. A-ConvNet* is worse than its original model. It can be seen from the results that manual rotation in the CT condition does improve the classification effect by increasing the diversity of samples to enhance the generalization of the method in this condition. However, this operation is not equipped to deal with the recognition problem under o.o.d distribution. The STN module improves the classification performance under the UNT condition while reducing the classification performance under the other two conditions, and although this has some application value, it is a compromise. While both ResNet50 and InceptionV3 perform well under the UNT condition, similar to the STN module, they lose out under other conditions, especially ResNet50.

V. CONCLUSION

This paper proposed a novel contrastive feature disentangling model termed ConFeDent and a strengthened version ConFeDent+ to induce features with improved generalization and interpretation ability for SAR ATR. The proposed models develop a novel idea of learning to disentangle the pose and identity features by comparing the semantic relations between two arbitrary SAR targets instead of treating them independently. A semi-parametric geometric transformation model and a second-order energy model are designed for contrastive regularization where we especially exploit the deduction-based geometry knowledge as supervision to teach the model to learn the semantic concept of aspect angle. Then the induced pose and identity feature subspaces from all training samples will contain complete aspect angle and class-specific information, respectively, increasing the combinational generalization ability. By taking the best from both worlds, it inherits the benefit of both deductive and inductive learning analogous to human learning, making the features more interpretable. Compared with the other SAR ATR algorithms, experimental results show that the proposed models can achieve better performance under the general SOC. In particular, under more complicated PAA-SOC and PAA-EOC-D scenarios, our models can outperform the compared algorithms by a large margin.

In the training process of ConFeDent and ConFeDent+, the cooperative samples are equivalently treated with the NC ones. Future research will consider the notable contribution of those cooperative training samples via meta-learning paradigm.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their careful review and helpful suggestions to improve the manuscript.

REFERENCES

[1] L. M. Novak, G. J. Owirka, W. S. Brower, and A. L. Weaver, "The automatic target-recognition system in SAIP," *Lincoln Laboratory Journal*, vol. 10, no. 2, 1997.

[2] E. R. Keydel, S. W. Lee, and J. T. Moore, "MSTAR extended operating conditions - a tutorial," *Proc Spie*, 1996.

[3] K. El-Darymli, E. W. Gill, P. Mcguire, D. Power, and C. Moloney, "Automatic target recognition in synthetic aperture radar imagery: A state-of-the-art review," *IEEE Access*, vol. 4, pp. 6014–6058, 2016.

[4] S. Niu, X. Qiu, B. Lei, C. Ding, and K. Fu, "Parameter extraction based on deep neural network for SAR target simulation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4901–4914, 2020.

[5] L. C. Potter and R. L. Moses, "Attributed scattering centers for SAR ATR," *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 79–91, 1997.

[6] J. Zhou, S. Zhiguang, C. Xiao, and Q. Fu, "Automatic target recognition of SAR images based on global scattering center model," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3713–3729, 2011.

[7] Y. Wang, P. Han, X. Lu, R. Wu, and J. Huang, "The performance comparison of adaboost and SVM applied to SAR ATR," in *CIE Int. Conf. Radar*, Oct 2006, pp. 1–4.

[8] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Multi-view automatic target recognition using joint sparse representation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 48, no. 3, pp. 2481–2497, 2012.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[10] G. Dong and G. Kuang, "SAR target recognition via sparse representation of monogenic signal on grassmann manifolds," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 3, pp. 1308–1319, 2016.

[11] G. Dong, G. Kuang, N. Wang, L. Zhao, and J. Lu, "SAR target recognition via joint sparse representation of monogenic signal," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 7, pp. 3316–3328, 2015.

[12] Z. Wen, B. Hou, Q. Wu, and L. Jiao, "Discriminative feature learning for real-time SAR automatic target recognition with the nonlinear analysis cosparse model," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 7, pp. 1045–1049, July 2018.

[13] X. Bai, R. Xue, L. Wang, and F. Zhou, "Sequence SAR image classification based on bidirectional convolution-recurrent network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9223–9235, Nov 2019.

[14] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, 2016.

[15] S. Deng, L. Du, C. Li, J. Ding, and H. Liu, "SAR automatic target recognition based on euclidean distance restricted autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3323–3333, July 2017.

[16] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, 2016.

[17] Q. Song and F. Xu, "Zero-shot learning of SAR target feature space with deep generative neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2245–2249, Dec 2017.

[18] F. Zhou, L. Wang, X. Bai, and Y. Hui, "SAR ATR of ground vehicles based on LM-BN-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7282–7293, Dec 2018.

[19] J. Zhang, M. Xing, and Y. Xie, "Fec: A feature fusion framework for sar target recognition based on electromagnetic scattering features and deep cnn features," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2174–2187, 2021.

[20] Z. Liu, L. Wang, Z. Wen, K. Li, and Q. Pan, "Multilevel scattering center and deep feature fusion learning framework for sar target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[21] Z. Wen, Z. Liu, S. Zhang, and Q. Pan, "Rotation awareness based self-supervised learning for sar target recognition with limited training samples," *IEEE Trans. Image Process.*, vol. 30, pp. 7266–7279, 2021.

[22] R. Geirhos, J. Jacobsen, C. Michaelis, and et al., "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, no. 2, pp. 665–673, 2020.

[23] A. Palazzi, L. Bergamini, S. Calderara, and R. Cucchiara, "Warp and learn: Novel views generation for vehicles and other objects," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 44, no. 4, pp. 2216–2227, 2022.

[24] J. Pei, Y. Huang, W. Huo, Y. Zhang, J. Yang, and T. Yeo, "SAR automatic target recognition based on multiview deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2196–2210, April 2018.

[25] U. Schmidt and S. Roth, "Learning rotation-aware features: From invariant priors to equivariant descriptors," in *IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp. 2050–2057.

[26] D. Malmgren-Hansen, A. Kusk, J. Dall, A. A. Nielsen, R. Engholm, and H. Skriver, "Improving sar automatic target recognition models with transfer learning from simulated data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1484–1488, 2017.

[27] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys, "Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 289–297.

[28] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015, pp. 2017–2025.

[29] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*. PMLR, 2016, pp. 2990–2999.

[30] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[31] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 4114–4124. [Online]. Available: http://proceedings.mlr.press/v97/locatello19a.html

[32] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *ICCV*, 2017.

[33] H. Kim and A. Mnih, "Disentangling by factorising," in *Int. Conf. Machine Learning (CVPR)*. PMLR, 2018, pp. 2649–2658.

[34] J. Lin, Z. Chen, Y. Xia, S. Liu, T. Qin, and J. Luo, "Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation," *IEEE Trans Pattern Anal Machine Intell*, 2019.

[35] Z. Wen, J. Wang, X. Wang, and Q. Pan, "A review of disentangled representation learning," *Acta Automatica Sinica*, vol. 48, no. 2, pp. 351–374, 2022.

[36] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[37] M. Mathieu, J. Zhao, P. Sprechmann, A. Ramesh, and Y. LeCun, "Disentangling factors of variation in deep representations using adversarial training," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 5047?5055.

[38] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 2180–2188.

[39] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[40] R. T. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *arXiv preprint arXiv:1802.04942*, 2018.

[41] C. Michael, A. Faruk, R. B. Girshick, Z. Larry, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," in *ICLR*, 2016.

[42] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "a simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn*, 2 2020, pp. 1–20.

[43] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193 907–193 934, 2020.

[44] A. J. Bell, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, 1995.

[45] R. Vidal, Y. Ma, and S. Sastry, "Robust principal component analysis," *Generalized Principal Component Analysis. Interdisciplinary Applied Mathematics*, vol. 40, 2016.

[46] C. Sutton and A. Mccallum, *An Introduction to Conditional Random Fields for Relational Learning*. Now Publishers Inc, 2007.

[47] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An Invitation to 3-D Vision-From Images to Geometric Models*. Springer-Verlag New York, 2004.

[48] B. Hou, Z. Wen, L. Jiao, and Q. Wu, "Target-oriented high-resolution sar image formation via semantic information guided regularizations," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1922–1939, 2018.

[49] X. Qin, H. Zou, S. Zhou, and K. Ji, "Region-based classification of sar images using kullback–leibler distance between generalized gamma distributions," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 8, pp. 1655–1659, 2015.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[53] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, no. 2, 2012.