Rotation Awareness Based Self-Supervised Learning for SAR Target Recognition With Limited Training Samples

Zaidao Wen, *Member, IEEE,* Zhunga Liu, Shuai Zhang, and Quan Pan

***Abstract*— The scattering signatures of a synthetic aperture radar (SAR) target image will be highly sensitive to different azimuth angles/poses, which aggravates the demand for training samples in learning-based SAR image automatic target recognition (ATR) algorithms, and makes SAR ATR a more challenging task. This paper develops a novel rotation awareness-based learning framework termed RotANet for SAR ATR under the condition of limited training samples. First, we propose an encoding scheme to characterize the rotational pattern of pose variations among intra-class targets. These targets will constitute several ordered sequences with different rotational patterns via permutations. By further exploiting the intrinsic relation constraints among these sequences as the supervision, we develop a novel self-supervised task which makes RotANet learn to predict the rotational pattern of a baseline sequence and then autonomously generalize this ability to the others without external supervision. Therefore, this task essentially contains a learning and self-validation process to achieve human-like rotation awareness, and it serves as a task-induced prior to regularize the learned feature domain of RotANet in conjunction with an individual target recognition task to improve the generalization ability of the features. Extensive experiments on moving and stationary target acquisition and recognition benchmark database demonstrate the effectiveness of our proposed framework. Compared with other state-of-the-art SAR ATR algorithms, RotANet will remarkably improve the recognition accuracy especially in the case of very limited training samples without performing any other data augmentation strategy.**

***Index Terms*— rotation awareness, self-supervised learning, weakly supervised learning, equivariant feature, data augmentation, limited training samples, synthetic aperture radar, automatic target recognition.**

I. Introduction

Automatic target recognition (ATR), which has been investigated as one of the longstanding synthetic aperture radar (SAR) applications, plays an important role in the field of civil and military reconnaissance and surveillance [1], [2]. Due to its electromagnetic mechanism of an active imaging system, a challenging problem arising in this domain is that the scattering signatures of a SAR target will be highly sensitive to different azimuth angles/poses [3]. Therefore, the problem of rotation sensitivity makes SAR ATR a more difficult task in comparison with other remote sensing sensors, such as optical and infrared systems.

In the early years, conventional algorithms based on pattern recognition extracted some features invariant to the target rotation from a SAR image in a handcrafted way. The rotation-invariant features were subsequently fed into an appropriate discriminative classifier for final recognition [4]–[7]. Nevertheless, since the feature extraction phase considers little identity information, discrimination of the resulted features will be generally weak. Consequently, the recognition accuracy of these algorithms is limited. Alternatively, the other type of algorithm based on machine learning has attracted increasingly attention in recent years [8]–[15]. Distinct from the conventional algorithms, the core idea of the learning strategy to address the rotation sensitivity problem is to fit a set of rotating targets as well as their class-labels with a parametric model to characterize the posterior or joint distribution [16]. Through the model, every intra-class rotating targets will be mapped to the same class-label so that their class-specific features can be learned in an adaptive and supervised way ignoring their pose variances. Besides, the model will unify the conventional two phases of feature extraction and classifier construction into an end-to-end learning framework in form. It follows that the discriminative features have made remarkable achievements in the recognition performance in terms of accuracy and efficiency.

A. Motivation and Related Works

Due to the superior performance, numerous supervised learning models such as the sparse model [9], [10], [13], [17], [18], deep neural networks [14], [19]–[22] have been developed and adopted over the latest few years and gradually become mainstream tools for image classification and target recognition. These models handle the rotation sensitivity problem by learning the class-specific features for target discrimination and ignoring the representative intra-class pose variance. Nevertheless, recent theoretical developments have revealed that the features induced from the model are not globally invariant to the target rotation [23], [24]. Although the loss of representative information will not influence the discrimination of features, it will aggravate the overfitting risk and degrade the generalization performance [25]. Consequently, the demand for the amount of rotating SAR targets will increase for model training, which is, however, intractable

Manuscript received Apr. 7, 2020, revised Nov. 19, 2020 and Apr. 25, 2021 accepted Aug. 5. This work was supported in part by the National Natural Science Foundation of China under Grant 61806165, 61790552, U20B2067; the Natural Science Basic Research Plan in ShaanXi Province of China under Grant 2020JQ-196.

Zaidao Wen, Zhunga Liu, Shuai Zhang, and Quan Pan are with Key Laboratory of Information Fusion Technology of Ministry of Education, Northwestern Polytechnical University, Xi'an, Shaanxi Province 710072, China.

in the practical SAR ATR scenario. The issue of deficient rotating SAR targets seems to be a common and central problem in the learning-based SAR ATR models.

In general, two types of techniques can address this problem. According to the learning theory, the overfitting problem is related to the model complexity/capacity and the number of training samples [16]. It follows that the first type of technique aims at reducing the model complexity or augmenting the training samples with some strategies. Chen et al. proposed the so-called all convolutional networks (A-ConvNet) by leaving out the fully connected layers to reduce the amount of model-free parameters [14]. Zhao et al. exploited a convolutional highway unit to train deeper networks with limited SAR data [26]. Besides, some other research imposed the regularization on the latent features or model parameters to constrain the capacity [10], [27]–[29]. Alternatively, the training data can be augmented from other sources with transfer learning [30]–[32], or from original target samples with randomly rotating, shifting, adding noise, generating pseudo target image, and so on [8], [19]. Although these strategies, especially data augmentation, can enhance recognition accuracy, they will not solve the fundamental problem of rotation information loss, but bring in many ambiguous artifacts. It is clear that manually rotating an image by a general angle will require pixel interpolation based on the smoothness assumption. However, the problem of rotation sensitivity will break this assumption so that the actual physical scattering signature of the generated pseudo rotating targets will be highly different from the real ones. Except for recognizing the identity of a target, the practical ATR task will need to further describe its pose, state, action from the scattering signature for interpretation, and decision making. These subsequent high-level descriptive requirements cannot benefit from those less expressive features but will be influenced by the involved ambiguous artifacts. Furthermore, the augmented pseudo targets or other data sources do not explicitly provide new more discriminative, or representative information than the original ones [3], [24]. The improved recognition accuracy is mainly caused by the augmented background clutters [21].

The other type of technique focuses on improving the representation ability of the features for better generalization. Accordingly, instead of ignoring the rotational pattern as the normal discriminative learning models, this type of approach will exploit the azimuth angle to account for the representative rotation information explicitly or implicitly. Song and Xu [33] proposed a zero-shot learning model to discover an expressive feature space spanned by some rotation-invariant features and azimuth angle. This model allows us to recognize a target that has never been trained in advance but still requires sufficient rotating labeled samples of other classes with an explicit azimuth angle to learn the feature space. In addition to exploiting the azimuth angle of each target independently, Pei et al. [3], Zhang et al. [13] and Bai et al. [22] constructed several multi-view learning models to capture the correlations among sequential intra-class targets with different azimuth angles and recognize them collaboratively. However, their resulted features cannot represent the explicit rotational information of the targets, and how to adapt these models to the task of single-view individual recognition has been rarely studied and discussed. In our prior research [34], we developed a self-supervised strategy to generate seven rotational patterns via rotating each SAR target image by several particular angles, namely $\{0^\circ, \pm 90^\circ, \pm 180^\circ, \pm 270^\circ\}$ without need pixel interpolation. Then the model is trained to recognize the rotation actions and the target identity collaboratively. Unfortunately, this strategy will essentially focus on the rotation of the image rather than the target therein so that it still has little ability to characterize the rotational pattern of the target.

In summary, the rotation sensitivity problem as the special challenge in SAR ATR will further aggravate the overfitting risk in the practical condition of limited training targets. Although this problem has been investigated and many aforementioned effective solutions have been proposed in this field, there remains a need for an efficient SAR ATR oriented feature learning approach that can explicitly characterize the rotational patterns of the target under the condition of limited samples.

B. Contribution

This paper develops a novel rotation awareness-based learning framework termed RotANet for SAR ATR under the condition of limited training samples. We declare that the individual azimuth angle cannot explicitly account for the pose variation of a target and thus we propose a new encoding scheme to define and characterize the rotational pattern of an ordered sequence comprising multiple intra-class targets. According to this scheme, a large number of sequences in different orders can be generated from these targets via permutations, which acts as a new data augmentation strategy without bringing in any artifact. By further exploiting the intrinsic relational constraints among these generated sequences as the supervision, we develop a novel self-supervised task which makes RotANet learn to predict the rotational pattern of a baseline sequence and then autonomously generalize to the others without external supervision. It follows that it essentially contains learning and a self-validation process for human-like rotation awareness. This self-supervised task is exploited to regularize the learned feature domain of RotANet in conjunction with an individual target recognition task to improve the generalization ability of the features. It will be applicable for both the single-view and multi-view ATR tasks without the need to change the structure. Extensive experiments are conducted to validate the effectiveness and superiority of the proposed framework. The main achievements, including contributions to the field, are summarized as follows:

- We suggest to pay more attention to learning a function with the labeling and rotation equivariant property for information augmentation, which sheds a new light on addressing the issue of limited rotating targets in SAR scenario. To this end, we propose a new encoding scheme to define and characterize the rotational pattern of an ordered sequence comprising multiple intra-class targets, which implicitly captures their high-order correlations in pose. According to this scheme, a large number of sequences in different orders can be generated via permutations, which acts as a new data augmentation strategy without bringing in any artifact.

- RotANet elaborates a novel self-supervised task for human-like rotation awareness, which importantly serves as a task induced regularization on the feature domain to provide a strong learning bias. This task newly exploits the intrinsic relational constraints among self-permuted target sequences as the supervision to make the model learn to predict the rotational pattern of a baseline sequence and then autonomously generalize to the others without external supervision. It follows that it essentially contains a learning and a self-validation process to improve the generalization ability of the features, so that this new learning manner apparently has the advantage over conventional ATR algorithms whose learning process merely relies on the given supervised label without validation.
- RotANet can make better use of the underlying information in the training set, and can simultaneously utilizes the discriminative individual and representative sequential information in a unified model for both single-view and muti-view ATR tasks. On the recognition problem of ten-types of targets in moving and stationary target acquisition and recognition (MSTAR) benchmark, RotANet can improve over 9% higher accuracy than the state-of-the-art baseline approach, when the number of actual training samples is only 5% of the original ones.

The remainder of this paper is organized as follows. Sec. II presents the preliminary to highlight the challenges in SAR ATR. Sec. III formally proposes our rotation awareness based learning framework and its detailed architecture. Extensive experiments are carried out in Sec. IV to demonstrate the effectiveness and superiority of the proposed framework. Sec. V concludes this paper and shows some possible future directions.

II. Preliminary

The aims in this section are twofold. First, we formulate the problem of ATR and analyze the special challenge and some insights into the SAR scenario. Second, we discuss the required property of features to shed light on our solution. Finally, we introduce the related research on self-supervised learning.

A. Problem Formulation and Insight

Theoretically speaking, the purpose of ATR is to create a function $g : \mathcal{X} \rightarrow \mathcal{Y}$ from an input target image domain $\mathcal{X}$ to a predefined identity label domain $\mathcal{Y}$. Compared with the general RGB image recognition tasks, such as face and handwritten digit recognition, SAR ATR will suffer from special challenges due to their different imaging mechanisms. In the visible optical band, light mainly propagates in the form of diffuse reflection so that stable and similar appearance will be presented in different aspect angles as the target rotates [35]. Therefore, it is proved that these optical images of rotating targets are generally residing in a smooth low dimensional manifold. However, in the microwave band such as the X-band for SAR ATR, the propagation of electromagnetic waves is alternatively specular reflection, which makes the strength of scattering in the SAR image sensitive to variations of the target pose therein. Accordingly, it further results in a nonsmooth variation in $\mathcal{X}$ as azimuth angle changing. This issue is illustrated in Fig. 1 in which the difference of azimuth angles between two targets is only 5° while their scattering patterns are much different. The different property of the input image domain $\mathcal{X}$ between optical and SAR will especially yield the problem of rotation sensitivity in SAR ATR scenario, which motivates us to develop a distinct SAR image-oriented processing method for better recognition.

Fig. 1. Illustration of the rotation sensitivity issue of SAR target.

To tackle this problem, the conventional algorithms will design g as a composition of two consecutive sub-functions, namely a feature extractor and a classifier. The feature extractor will be constructed in a handcrafted way to produce some rotation-invariant features according to the SAR image domain knowledge [4], [7], [36]. Alternatively, the discriminative learning model simply addresses this issue by minimizing the following optimization [14], [19], [21], [22]:

$$\min_{g \in \mathcal{G}} \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \ell(g(\mathbf{x}_{i,c}), \mathbf{y}_c), \tag{1}$$

where $\mathbf{x}_{i,c} \sim \mathcal{X}$ and $\mathbf{y}_c \sim \mathcal{Y}$ are the i-th sampled target and its categorical label vector of c-th class, respectively, and $N = \sum_{c=1}^{C} N_c$ counts the total number of the training samples from C classes. Note from the Eq. (1) that this learning process will adaptively tend to a g that can map the rotating samples in class c towards the same label vector. It follows that a preferred optimizer is that g is invariant to the rotation variation given by:

$$g(\mathbf{x}_{i,c}) = g(\mathbf{x}_{j,c}), \ \forall \mathbf{x}_{i,c}, \mathbf{x}_{j,c}. \tag{2}$$

Nevertheless, due to the non-smoothness property of $\mathcal{X}$ for the SAR target image, we have to acquire more targets from different azimuth angles than the normal optical image domain, which is an intractable task in practical SAR applications. Data augmentation via manually image rotation will be a widely used pre-processing method in RGB image classification, which can always improve the performance. However, the manually rotated pseudo-SAR targets, neither reflecting the real scattering property nor bringing in new discriminative information, will empirically decrease the recognition accuracy, and it may mislead the subsequent target description tasks as well. Besides, it is shown that the general convolutional neural networks (CNN), a most widely used discriminative model, will only appear local rotation invariant property. If no extra supervised constrain providing a strong learning bias is imposed on g or $\mathcal{G}$, directly optimizing (1) will not easily obtain the desired rotation invariant function. More importantly, since the intra-class variances are not encoded in the supervised label vector, the representation ability of the intermediate features learned from (1) will be sacrificed. As a consequence, they cannot be further applied in some high-level target interpretation tasks such as pose estimation, action, and motion state description, and they are more unreliable due to adversarial attack [37].

B. From Invariance to Equivariance

According to the above analysis, the discriminative learning based SAR ATR model will facilitate a rotation-invariant function to deal with the problem of rotation sensitivity. In this paper, we alternatively suggest that attention can be newly paid to a more generalized rotation equivariant function to address the issue.

More formally, let $\mathbf{x}_i$ and $\mathbf{x}_j$ be a pair of centering aligned SAR target images drawn from the same class of the input domain. Assuming that there is a latent oracle function $\mathcal{R}_{i\to j}$ transforming $\mathbf{x}_i$ to $\mathbf{x}_j$ given by $\mathcal{R}_{i\to j}(\mathbf{x}_i) = \mathbf{x}_j$. It follows that this oracle will fully capture the entire representative and discriminative information of $\mathbf{x}_i$ and $\mathbf{x}_j$. A function g is said to be equivariant to a group of transformations $\mathcal{R}_{i\to j}, \forall i, j$ if there is a corresponding function $\hat{\mathcal{R}}_{i\to j}$ satisfying the following equality:

$$g(\mathbf{x}_j) = g(\mathcal{R}_{i\to j}(\mathbf{x}_i)) = \hat{\mathcal{R}}_{i\to j}(g(\mathbf{x}_i)),\ \forall i, j. \qquad (3)$$

Note from the comparison between Eq.(3) and Eq.(2) that the invariant property will specially turn to (3) when $\hat{\mathcal{R}}_{i\to j}$ is the identical transformation. In other general cases, it will intuitively encode the variations of the input domain to the function domain of g. It follows that the resulted transformations equivariant features will be more informative than the invariant ones. In this regard, the function obtained from (1) is essentially equivariant to the identity variations while invariant to the other transformations, which makes the resulted features less representative in terms of the intra-class variations. If we turn to the rotation equivariant features, it will shed new light on addressing the issues of rotation sensitivity and limited training samples for SAR ATR simultaneously. In other words, we aim to additionally characterize the rotation variations among intra-class targets and exploit their expressive information for equivariant feature learning instead of discarding them as the general learning paradigm does.

It has been argued that the human visual cognitive system will automatically capture the pose-equivariant features in terms of a self-constructed coordinate frame, which may influence the result of human cognition [38], [39]. Inspired from neuroscience, numerous research in computer vision recently developed some special $\mathcal{G}$ satisfying the rotation equivariant property, such as equivariant transformer network [40], equivariant capsule/convolutional network [39], [41]–[43]. Nevertheless, these models generally suffer from the cost of extra computations. How to efficiently learn the rotation-equivariant features without extra annotation efforts will become a crucial issue.

C. Self-Supervised Learning

For the discriminative learning paradigm (1), a large number of labeled samples are generally required to learn a well-generalized model. Accordingly, it will require extensive costs of collecting and annotating large-scale datasets, which is intractable in many scenarios, especially SAR ATR. In recent years, self-supervised learning algorithms served as a special subset of unsupervised learning methods, have been newly proposed to learn the representative features from unlabeled data themselves without external human-annotations [44]. The core idea of self-supervised learning is firstly to construct a surrogate task to recover some underlying structure or relationship among the training samples. Secondly, a discriminative model will be constructed to address the surrogate task whose supervisory signal is simply generated from the training samples themselves. Finally, the learned model parameters or the intermediate features will be transferred to downstream tasks by fine-tuning. These surrogate tasks generally include image rotation prediction [45], image inpainting [46], temporal order verification or ranking [47], [48], and so on. It ensures that some semantically meaningful features can be captured through the process of accomplishing the tasks. Essentially, this new learning paradigm aims to explore more representative and semantic information among the training samples for supervision instead of merely exploiting the identity labels. The information may not straightforward contribute to the final downstream task, but it will help to improve the generalization ability of the model. Additionally, in this learning paradigm, the training samples will be manually augmented via some semantic transformations to generate the self-supervisory information, which makes it potentially appropriate for learning with limited samples. Therefore, inspired by these observations, this paper will design a self-supervised strategy to learn some semantic features accounting for the target rotational patterns.

III. Rotation Awareness Based Self-supervised Learning Framework

In this section, we will formally develop a rotation awareness-based weakly self-supervised learning framework for SAR ATR. First, we propose a new effective scheme to characterize and encode the rotation pattern of sequential intra-class targets. Second, based on the proposed encoding scheme, we further design a surrogate task of sequential targets verification to explore underlying rotational patterns of the multiple intra-class targets. With this task, the equivariant features w.r.t. our designed rotational pattern can be learned in a self-supervised manner without external semantic labeling. Third, the surrogate task will be combined with a target recognition task in an elaborated dual-task learning framework. Finally, the superiority of the framework will be discussed.

A. Rotation Variation Encoding Scheme

As aforementioned, the current learning-based models for SAR ATR will focus on the label equivariant and rotation invariant features. In this paradigm, the only identity labels $\mathbf{y}_c$ modeling the categorical information are exploited as

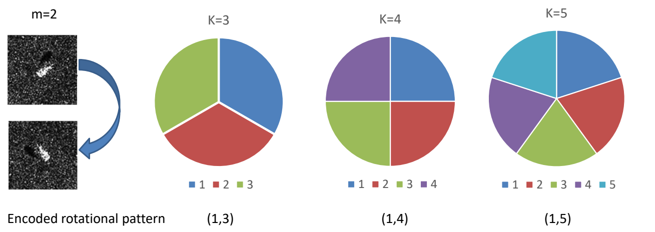


Fig. 2. Illustration of the equidistant encoding scheme of rotational patterns.

the supervisory signal for discriminative feature learning. However, since $\mathbf{y}_c$ ignores the individual variation, much representative information contained in the samples will be wasted during the learning process. To address this issue, the most intuitive strategy will exploit the azimuth angle of each target as the additional supervision [33]. Unfortunately, the rotation essentially describes a pattern of orientation variation of pairwise targets, so that the individual azimuth angle cannot characterize the pattern. Furthermore, it is also intractable to collect sufficient labeled SAR targets with exact full azimuth angles for model learning. Instead, we aim at modeling the pattern of azimuth angles variation of some ordered intra-class target sequence. According to the previous discussion, the rotation action between two target images $\mathbf{x}_i$ and $\mathbf{x}_j$ is encoded in the transformation $\mathcal{R}_{i\to j}$. Conventionally, the $\mathcal{R}_{i\to j}$ is normally parameterized as a matrix in the special orthogonal group of $\mathbb{R}^3$ or Euler angle [49], but it will require the rotation angle between two targets. In this paper, we discretize the polar coordinate into K equidistant indexed cells to form an orientation encoding table shown in Fig. 2, and assign $\mathbf{x}_i$ and $\mathbf{x}_j$ with a corresponding index according to their orientations, respectively. In this regard, the encoding scheme provides a rough measurement of the orientation of a target as weakly supervised information without the requirement of the exact azimuth angle. Then the total rotation actions from $\mathbf{x}_i$ to $\mathbf{x}_j$ will correspond to K^2 types of patterns encoded into a two-dimensional rotational index vector. More generally, let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be $m \geq 2$ ordered target images from the same class and form a sequence. It will implicitly describe a series of rotation actions between two adjacent images to characterize its rotational pattern encoded as an m-dimensional rotational index vector. In this regard, it is possible to exploit a discriminative model to recognize the rotational pattern of any m intra-class samples for rotation equivariant feature learning in which K^m-dimensional one-hot encoding rotation labels $\mathbf{r}$ will be used for supervision. It follows that the number of target sequences will far exceed the target images, and the sequence encoding scheme will provide a data augmentation strategy for training with limited samples.

B. *Self-supervised Task for Rotation Awareness*

Sequential learning has been extensively studied in a variety of fields such as video monitoring, robotic path planning, adaptive control algorithms, and so on, which can provide a more abundant source of information than a single image for model learning [47]. Some recent research on SAR ATR also exploited the idea of sequential learning [21], [22], but they only focused on predicting the shared label of a given input target sequence without considering the relationship among sequences. Besides, their models are not appropriate for the single target recognition task. Instead, we will investigate the underlying structure among different sequences induced from the same targets, which will moreover promote the model to understand the intrinsic representation of the rotating targets via self-supervised strategy.

More specifically, for the above sampled targets $\{\mathbf{x}_i\}_{i=1}^m$, they can actually generate $P_m^m = m!$ different ordered sequences. Although these sequences will correspond to different rotation labels, they share some intrinsic relationship for model autonomous learning. For example, considering a sequence comprising four targets $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ illustrated in Fig. 3, its rotational index vector is given by $[1, 2, 3, 4]^{\mathrm{T}}$. If $\mathbf{X}$ is randomly permuted into a new sequence $\hat{\mathbf{X}} = \mathcal{P}(\mathbf{X}) = \{\mathbf{x}_4, \mathbf{x}_3, \mathbf{x}_2, \mathbf{x}_1\}$ with a permutation operator $\mathcal{P}$, the rotational pattern of $\hat{\mathbf{X}}$ is intrinsically constrained as $\mathcal{P}([1, 2, 3, 4]^{\mathrm{T}})$ without external supervised knowledge. It follows that the rotation label of $\hat{\mathbf{X}}$ can be directly inferred from that of $\mathbf{X}$ in a self-supervised manner, which directly motivates us to develop a new self-supervised learning framework (4) for the rotational pattern awareness:

$$\min_g \mathbb{E}_{\mathbf{X}\sim\mathcal{X}} \ell(g(\hat{\mathbf{X}}), \mathcal{P}(\mathbf{r})) \text{ s.t. } g(\mathbf{X}) = \mathbf{r},\ \hat{\mathbf{X}} = \mathcal{P}(\mathbf{X}) \quad (4)$$

The insight of (4) is that if a well-generalized autonomous model can accurately recognize not only the defined rotational pattern of the sequence $\mathbf{X}$, but also any permutation of $\mathbf{X}$, and vice versa. Predicting the rotational pattern of a permuted sequence can be regarded as a self-validation process to verify whether the model has the ability of rotation awareness. In this regard, this new self-supervised task contains both learning and validating to enhance the generalization ability of the intermediate features. More importantly, we can generate $P_{N_c}^m$ ordered target sequences from each class in this way, which serves as a new type of data augmentation strategy to explore extra information from the available training samples without altering their scattering signatures. The proposed self-supervised task is similar to the research of learning-to-rank in [48] and learning-from-shuffle [47], while we focus on learning the rotation variations in terms of the targets' orientations instead of the temporal correlations. The self-supervised task in [34], [45] aims to predict the rotation action of the image rather than the target therein. Instead, we will capture the intra-class relationship accounting for the rotation variation.

C. *Rotation Awareness Based Learning Framework for ATR*

We have proposed a weakly self-supervised strategy to define a new rotational pattern of multiple target images. Based on this strategy, we develop a novel self-supervised learning task for rotation awareness. This subsection will design an efficient framework for ATR oriented discriminative rotation equivariant feature learning. Observing the sequential rotating SAR target images, our visual cognition system will be aware of not only their identity consistency to associate

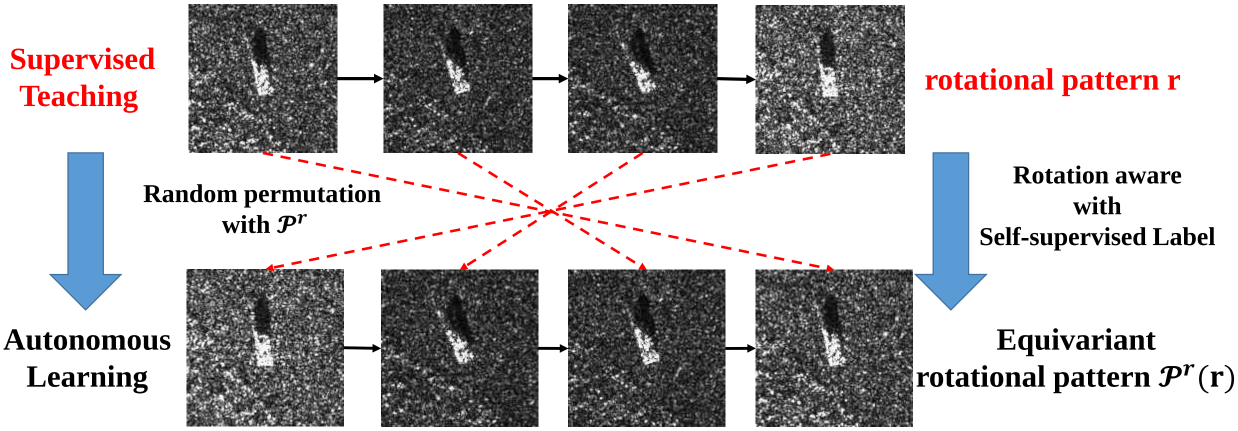

Supervised Teaching
rotational pattern r
Random permutation with $\mathcal{P}^r$
Rotation aware with Self-supervised Label
Autonomous Learning
Equivariant rotational pattern $\mathcal{P}^r(\mathbf{r})$

Fig. 3. Illustration of our designed self-supervised task of equivariant feature learning for rotation awareness.

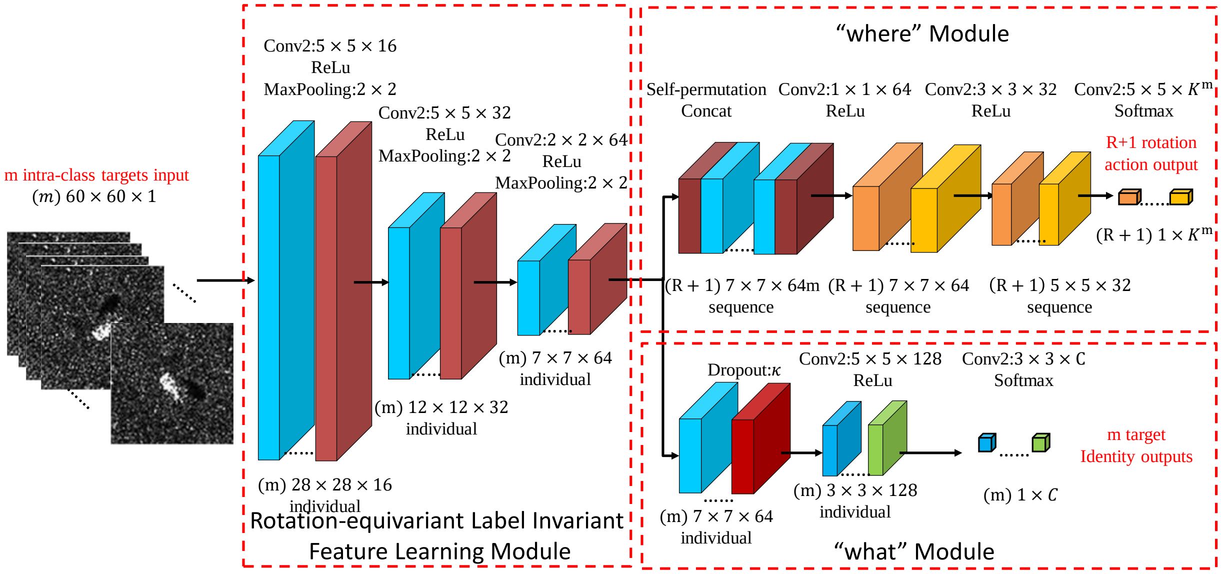


Fig. 4. Overall architecture of rotation awareness based network (RotANet) for SAR ATR.

them in category, but also their sequentially logical variations to distinguish them in orientation. However, the previous research on sequential learning [22] only paid attention to the global property of the target sequence while ignoring the individual character. Consequently, the resulted model played no role in the task of single target recognition. To address this problem, we will place the sequence learning task in a learned equivariant feature domain rather than the input domain. In this domain, features of each target should contain sufficient discriminative information for recognition, and the feature sequence of some intra-class targets should contain enough representative information for rotation awareness. To this end, we first construct a parametric feature extractor to project m targets into a latent feature domain independently as $g(\mathbf{x}_i;\Psi) = \mathbf{z}_i, \forall i = 1,\dots,m$, where $\mathbf{z}_i$ and Ψ contain the features of $\mathbf{x}_i$ and the parameters of the extractor, respectively. To encode the rotational information into the feature domain to enhance the representation ability, we will exploit our designed self-supervised task as the regularization. More specifically, let $\mathbf{Z}$ be a baseline feature sequence constructed from $\{\mathbf{z}_i\}_{i=1}^m$ in a given order, and its rotational label is given by $\mathbf{r}$ that can be accurately predicted by an oracle $\hat{\mathcal{R}}$. If R, $R < m!$ times random permutations are performed on $\mathbf{Z}$ as $\hat{\mathbf{Z}}^r \leftarrow \mathcal{P}^r(\mathbf{Z}), \forall r = 1,\dots,R$, the rotational labels of $\hat{\mathbf{Z}}^r$ will be $\mathcal{P}^r(\mathbf{r})$. To further guarantee the discrimination for target recognition, we construct a classifier $\mathcal{C}$ to associate each $\mathbf{z}_i$ with the identity label $\mathbf{y}_i$. To meet the above requirements, we will propose the following bi-level optimization problem which can be explained as a teacher-student learning framework:

$$\min_{\Psi,\phi,\varphi} \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^m} \sum_{i=1}^m \ell(\mathbf{y}_i, \mathcal{C}(\mathbf{z}_i;\varphi)) + \lambda \sum_{r=1}^R \ell(\mathcal{P}^r(\mathbf{r}), \hat{\mathcal{R}}(\mathcal{P}^r(\mathbf{Z});\phi^*)),$$
$$\text{s.t. } \phi^* = \arg\min_{\phi} \ell(\hat{\mathcal{R}}(\mathbf{Z};\phi),\mathbf{r}),\ \ \mathbf{z}_i = g(\mathbf{x}_i;\Psi) \tag{5}$$

where ϕ stands for the parameters in $\hat{\mathcal{R}}$ to be learned, λ is a hyper-parameter, and the loss functions are cross-entropy between two categorical distributions. In (5), the lower-level problem can be explained as the teacher network to predict the rotation pattern of the baseline sequence correctly [50]. Then the upper-level problem is interpreted as a student network to validate whether the model has the ability of rotation awareness. It can be noticed from (5) that this optimization problem can be also explained as minimizing the potential energy functions in a conditional random fields $\mathbb{P}(\{\mathbf{z}_i\}_{i=1}^m|\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^m, \mathbf{r})$ to regularize the feature domain [16], [24], where two groups of loss functions will respectively characterize the unary and high-order potentials. Nevertheless, it is difficult and time-consuming to solve bi-level optimization due to stochastic sample sampling. Alternatively, we will turn to the following relaxed problem to achieve its lower bound for simplicity.

$$\begin{aligned} &\min_{\Psi,\phi,\varphi} \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^m} \sum_{i=1}^{m} \ell(\mathbf{y}_i, \mathcal{C}(\mathbf{z}_i;\varphi)) + \lambda\ell(\mathbf{r}, \hat{\mathcal{R}}(\mathbf{Z};\phi)) \\ &+\lambda \sum_{r=1}^{R} \ell(\mathcal{P}^r(\mathbf{r}), \hat{\mathcal{R}}(\mathcal{P}^r(\mathbf{Z});\phi)),\ \text{s.t. } \mathbf{z}_i = g(\mathbf{x}_i;\Psi) \end{aligned} \tag{6}$$

Fig. 4 shows the overall architecture of the proposed framework termed rotation awareness-based networks (RotANet) to implement (6), and it that contains three modules to simulate a two-stream structure in the visual cortex of a human being [38]. More specifically, the feature extractor g imitating the primary visual cortex is composed of three 2D-convolution layers with the reflected linear unit (ReLU) activation function and three max-pooling layers. Accordingly, m randomly sampled intra-class target images will independently be fed into this module to produce m feature maps $\{\mathbf{z}_i\}_{i=1}^m$ with the size of $7 \times 7 \times 64$. These feature maps will be fed into "what module" and "where module" for target recognition and self-supervised rotation awareness, respectively. "what module" $\mathcal{C}$ contains a dropout layer and a full convolution layer with the softmax activation function. It is designed to simulate the ventral pathway of the visual cortex for individual target recognition. It follows that it will produce m categorical identity labels for m input targets independently. The dropout rate in the module is denoted by κ. In "where module" module, the feature maps $\{\mathbf{z}_i\}_{i=1}^m$ will be firstly concentrated into the baseline sequence $\mathbf{Z}$ with the size of $7 \times 7 \times 64m$ according to their original sampling order. Then $\mathbf{Z}$ passes through a self-permutation layer to randomly generate R sequences $\{\mathbf{Z}^r\}_{r=1}^R$ for self-learning. According to our relaxation in (6), $R+1$ feature sequences will be simultaneously fed into $\hat{\mathcal{R}}$ to imitate the dorsal pathway of the visual cortex for rotation awareness. $\hat{\mathcal{R}}$ contains three full convolution layers whose activation functions are ReLU, ReLU, and softmax, respectively. Therefore, the overall architecture of RotANet is m inputs, and $m+1+R$ outputs deep networks. To further prevent the model from over-fitting, we exploit the weight decay regularization strategy during training. In the testing phase for SAR ATR, we will remove the "where module" so that both single-view and multi-view query samples can be fed into the framework and obtain the identity label.

D. Framework Discussion

We have proposed an effective weakly self-supervised strategy to encode the rotational patterns of multiple intra-class samples. With this strategy, we further develop a dual-task learning framework for rotation awareness-based SAR ATR. This subsection will present an intuitive discussion to further highlight the superiority.

In contrast to the optical RGB image classification task, the SAR image will especially encounter a challenge of target pose sensitivity due to the different imaging mechanism. Therefore, many state-of-the-art image processing methods or strategies cannot achieve satisfactory performance on the SAR image. It motivates us to develop a new specified image-level algorithm accounting for this issue. According to our analysis in Sec. II, the general idea of the most current algorithms for SAR ATR is to learn rotation-invariant features to alleviate the influence of pose variations. Nevertheless, the discard of rotation information during learning will not only degrade the representation ability but also increase the demand for rotating training samples to obtain a well-generalized model. It will further aggravate the over-fitting issue with limited training samples. In contrast, RotANet will newly pay attention to learning a function with rotation-equivariant property for rotational information augmentation. Distinct from some algorithms leveraging the azimuth angle as the extra supervised information [33], we declare that a single azimuth angle will not essentially model the rotation action that reflects continuous pose variations of a target. Alternatively, we propose a weakly supervised scheme to characterize the orientation variation pattern among targets in a sequence, which will implicitly capture the high-order rotational correlations for equivariant feature learning [51]. As a result, we essentially model the joint distribution of $\{\mathbf{z}_i\}_{i=1}^m$ given m targets without assuming their independence so that the underlying high-order relationship can be captured to regularize the feature domain to make it more structured. More importantly, this scheme will not require the exact azimuth angle of each SAR target so that it will introduce no extra manual labeling efforts. The resulted rotational labels will reflect the knowledge of the correlation and variance among m targets' orientations. In summary, the first contribution of RotANet is an effect weakly-supervised encoding scheme that newly characterizes the joint relationship of m targets for rotation equivariant feature learning instead of treating them independently. The resulted feature domain will be more structured and representative.

As the other notable contribution, RotANet develops a novel self-supervised learning module, namely "where module", which importantly serves as a task-induced regularization on the feature domain to provide a strong learning bias. In this module, an elaborated self-supervised task is proposed to predict the rotational patterns of some permuted sequences only supervised by the rotation label of the baseline sequence. Distinct from the general self-supervised tasks that are mostly used for model pre-training, our elaborated task will be learned in conjunction with the downstream task. We show that the self-supervised task will provide a structured regularization on the learned features to enhance the generalization ability.

Moreover, the designed task provides a learning-to-learn manner of guiding the model to have the ability of rotation awareness. It learns to predict the rotational pattern of a baseline sequence and then generalize to some permuted sequences for additional self-validating feedback. On the contrary, the general learning paradigm of self-supervised pre-training and supervised fine-tuning will only rely on the feedback from the downstream supervised label without extra validation. Thanks to this way, the generalization ability of the proposed model will be further improved.

Finally, the developed self-supervised task in conjunction with the rotational pattern encoding scheme will essentially enlarge the training data by a factor of $P^{m}_{N_c-1}$ for each class at most. More formally, for N_c samples in class c, we can sample $C^{m}_{N_c}$ baseline sequences, and then generate $P^{m}_{m} = m!$ permutations on each sequence, which will significantly improve the generalization ability of RotANet. In practical situation, we empirically observe that only $R < m!$ permutations are required to obtain a satisfactory performance. Therefore, it can be also regarded as a new data augmentation method for SAR image processing without bringing in the artifacts of altering the actual scattering signature of the target.

IV. EXPERIMENTS AND ANALYSIS

This section will conduct extensive experiments to evaluate the effectiveness and superiority of RotANet for SAR ATR. First, we will introduce the moving and stationary target acquisition and recognition (MSTAR) database and experimental settings. Second, the influences of different hyper-parameters will be empirically analyzed based on the cross-validation strategy for the model selection. Third, we design some ablation experiments to validate the effectiveness of our framework. Finally, we compare the results of the proposed method with those of the conventional SAR ATR methods to demonstrate its superiority. The overall experiments are performed on a desktop PC with i9 Intel CPU, double GeForce 1080Ti GPUs, and 64GB of memory with TensorFlow 2.3 for 5 times independently, and the average performances are reported[1].

A. MSTAR Dataset and Experimental Settings

MSTAR dataset, collected with the Twin Otter SAR sensor operating at X-band by the Sandia National Laboratory, is the most widely exploited benchmark for SAR ATR [52]. It contains about ten types of publicly released military ground targets obtained at different depression angles and azimuth angles with $0.3m \times 0.3m$ 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. 5 only for illustration. Following the common setting, the central cropped 60×60 magnitude image patches will be exploited in the training and testing phases to avoid the confusion of background cluttering [21]. To validate the model performance in the condition of limited training samples, only a fraction of target images from the training set will be randomly selected for model learning and the sampling rate is η. We do not exploit any other data augmented tricks during model training [14], [19].

(a) BMP2 (b) BTR70 (c) T72 (d) 2S1 (e) BRDM2
(f) BTR60 (g) D7 (h) T62 (i) ZIL131 (j) ZSU234

Fig. 5. Optical sample images of ten types of target in MSTAR database only for illustration.

TABLE I
INFORMATION OF SOC (10 TYPES OF TARGETS)

		Training Set		Testing Set	
Class	Serial No.	Depression	No.	Depression	No.
BMP-2	9563,9566,C21	17°	698	15°	587
BTR-70	c71	17°	233	15°	196
T-72	132,812,S7	17°	691	15°	582
BTR-60	k10yt7532	17°	256	15°	195
2S1	b01	17°	299	15°	274
BRDM-2	E-71	17°	298	15°	274
D7	92v13105	17°	299	15°	274
T-62	A51	17°	299	15°	273
ZIL-131	E12	17°	299	15°	274
ZSU-234	d08	17°	299	15°	274

In the following experiments, two testing conditions, namely the standard operating conditions (SOC) and extended operating conditions (EOC) will be tested to validate the target recognition performance according to the general setting [14]. More formally, the SOC scenario considers the identical serial numbers and configurations with different aspects and depression angles in the training and testing phases, while EOC considers large distinctions including substantial variations in depression angle, target configuration, and version variants. The detail information for SOC and three types of EOC are illustrated in Table I, Table II, Table III and Table IV, respectively.

In our framework, the convolutional kernels in each layer are initialized with the default Xavier uniform, and the bias vectors are zero-initialized. The standard Adam optimizer is exploited for model learning with the mini-batch size of 300. We use an early-stopping strategy to control the training process by monitoring the target recognition accuracy on a validating set. Since only η of target images from the training set are used for model learning, the rest will form the validating set. More specifically, the training process will stop if the monitored value, namely recognition accuracy on the validating set, has no improvement within 30 epochs. Then the model achieving the best performance of the monitored accuracy will be restored for further testing.

B. Hyper-Parameter Analysis

In this part, we will investigate the influence of several important hyper-parameters in our proposed model, including

[1] We also test the code on a single Titan Xp GPU, which can obtain the same performance.

TABLE II
INFORMATION OF EOC-1 (LARGE DEPRESSION VARIATION)

	Class	Training Set		Testing Set	
Class	Serial No.	Depression	No.	Depression	No.
T-72	SN132,A64	17° SN132	232	30° A64	288
2S1	b01	17°	299	30°	288
BRDM-2	E-71	17°	298	30°	287
ZSU-234	d08	17°	299	30°	288

TABLE III
INFORMATION OF EOC-2 (CONFIGURATION VARIANTS)

	Class	Serial No.	Depression	No.
Training Set	T72	132	17°	232
	BMP-2	9563	17°	233
	BRDM-2	E-71	17°	298
	BTR-70	c71	17°	233
Testing Set	T72	S7	17°,15°	419
		A32	17°,15°	572
		A62	17°,15°	573
		A63	17°,15°	573
		A64	17°,15°	573

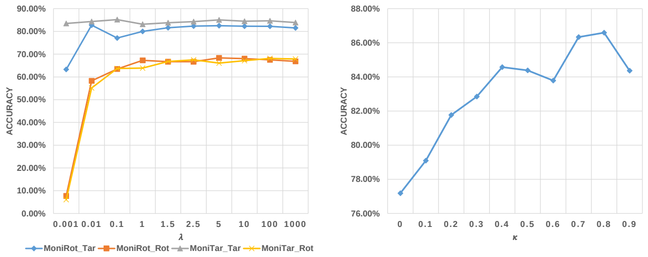
regularization parameter λ and the dropout rate κ. The hyper-parameters in Adam optimizer are default. The weight decay rate is set as 0.004, which is the same with A-ConvNet for fair comparison [14]. K in the rotational encoding table is fixed as 4. All the experiments in this subsection are conducted in SOC with $\eta = 10\%$. Considering the memory burden, the experiments in this subsection will be carried out in the case of $m = 2$ and $R = 1$, and the corresponding framework will be denoted by RotANet-m2R1.

Firstly, λ is considered as a task-induced regularization parameter, and it also balances the performances of two tasks, namely target recognition and rotation awareness. When $\lambda = 0$, the framework will turn to the standard discriminative feature learning model for target recognition without rotation awareness. Its performance will be specially validated in the next subsection. In this part, we only consider the positive value of λ by varying it from $1e^{-3}$ to $1e^{3}$ with $\kappa = 0.5$ fixed. To avoid the influence of restoring the best value of recognition accuracy in the early stopping strategy, we also show the results which are obtained by monitoring the accuracy of rotation awareness. The performances are plotted in Fig. 6(a). From the results, it is clear that the accuracy of rotation awareness is improving as the increase of λ, though it is only treated as an auxiliary task. Special attention should be paid to the increasing accuracy curves of target recognition. The curves of the target recognition accuracy reflect that the rotation awareness task will promote the performance of target recognition. From the figure, it can be noticed that the optimal performance can be achieved when $\lambda = 2.5$, which will be exploited in the following experiments.

Next, the dropout is a typical strategy to prevent the deep model from over-fitting by introducing the multiplicative random binary noise on the activations, which seems like dropping out κ rate of values among the activations during training. Through this trick, some connections in the network will be randomly blocked in each training epoch, which will increase the robustness and generalization of the model if κ is properly selected. In the experiment, we vary the dropout rate κ from 0 to 0.9, and the corresponding performance in terms of recognition accuracy is plotted in Fig. 6(b). Observed from the result that the dropout operation will heavily influence the performance. More specifically, when we omit the dropout layer by setting $\kappa = 0$, the accuracy will only achieve about 77%. When it is set as around 0.7 or 0.8, the accuracy reaches the best value over 86%, almost 9% increment. When the value of κ is further increasing, the performance will be degraded due to the reduced model capacity. In the following experiments, κ will be fixed as 0.7 for balance.

TABLE IV
INFORMATION OF EOC-3 (VERSION VARIANTS)

	Class	Serial No.	Depression	No.
Training Set	T72	132	17°	232
	BMP-2	9563	17°	233
	BRDM-2	E-71	17°	298
	BTR-70	C71	17°	233
Testing Set	T72	812	17°,15°	426
		A04	17°,15°	573
		A05	17°,15°	573
		A07	17°,15°	573
		A10	17°,15°	567
	BMP-2	9566	17°,15°	428
		c21	17°,15°	429



(a) (b)

Fig. 6. Performance with different hyper-parameters. (a) λ. (b) κ.

C. Ablation Experiments

In this subsection, we will design some specific experiments via ablation to demonstrate the effectiveness of RotANet.

In the first ablation experiment, two typical architectures should be compared to validate the effectiveness of the proposed (relaxed) self-supervised task. To this end, we will first set $\lambda = 0$ and $m = 1$ to remove the "where module" from the RotANet. In this situation, the rest architecture denoted by RotANet-m1λ0 becomes a conventional CNN for SAR ATR and treats each training sample independently. The other one is to set $m = 1$ but preserve the dual-task learning framework. In this situation, the architecture denoted by RotANet-m1 will turn to estimate the identity and rough orientation of a target image simultaneously. Accordingly, RotANet-m2R1 and the above two variants will be tested on SOC by varying η from 5% to 40%, and the comparison results are illustrated in Fig. 7. Observing from the results, we can notice that RotANet-m2R1 will outperform the other two compared variants at all sampling rates, which demonstrate its effectiveness. For RotANet-m1λ0, the standard CNN based ATR framework will

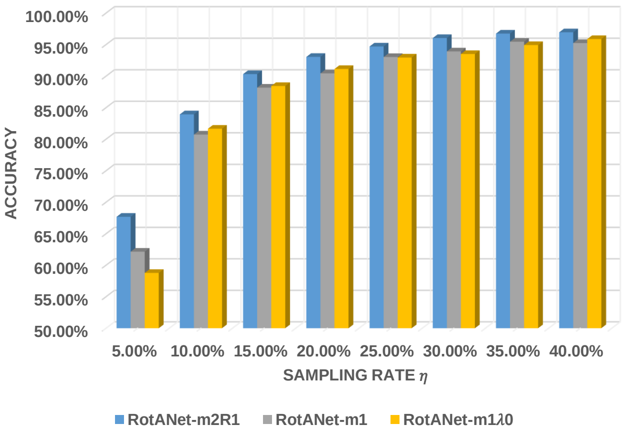


Fig. 7. Comparison results of different variants of RotANet via ablation.

perform the worst when the sampling rate is 5%, which is mainly caused by the over-fitting issue. For the other two architectures, they can achieve better performances due to involving an auxiliary task. As η increases, the accuracy of RotANet-m1λ0 also rises and gradually exceeds or approximates the result of RotANet-m1. We speculate that the rough measurement of the orientation does not provide sufficiently useful information for ATR so that it will mislead the learning process when the training set is enlarged. On the contrary, our proposed RotANet-m2R1 will always be better than RotANet-m1. It demonstrates that the proposed encoding scheme is more expressive and effective. In summary, these planned comparison results reveal the effectiveness of the proposed rotation awareness-based weakly self-supervised strategy.

In the next ablation experiment, we will investigate the impact of sequence length m and permutation amount R. As one of the most notable superiority, sampling, and jointly modeling $m > 1$ intra-class targets can not only capture the correlation among targets to regularize the feature domain more structured but also significantly enlarge the number of training sequences in each class by a factor of $C^{m}_{N_c-1}$. Intuitively speaking, a larger m will yield better performance at the cost of the increasing memory burden, and thus only $m = 2$ and $m = 3$ will be considered in the experiment for simplicity. R, $1 \leq R < m!$ accounts for the influence of our proposed self-validation task. According to our declaration in Sec. III-C, R sequences should be randomly generated via self-permutation during the training phase, which is, however, difficult to realize with the current deep learning library. Instead, we have to use R sequences with predefined permutations for model training. When $m = 2$, $R = 1$ is the only choice. When $m = 3$, we will vary R from 1 to 5 to investigate its influence. Considering the memory limitation, the experiment will be conducted in the condition of $\eta \in \{5\%, 10\%\}$, and the results are illustrated in Fig. 8. From the comparison results of RotANet-m2R1 with other variants in the figure, we can see that increasing m will indeed apparently promote the recognition performance. However, as R increases, the accuracy will slightly rise to the highest value and gradually drop when the sampling rate is 5%. When $\eta = 10\%$, the influence of R becomes weak, and thus R

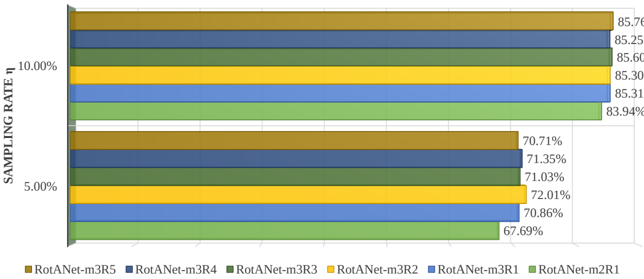


Fig. 8. Comparison results with different m and R.

intuitively plays a less important role than m. We conjecture that this is caused by problem relaxation. During solving (6), we treat the baseline sequence and R permuted ones equally ignoring their intrinsic bi-level optimization relationship for relaxation. When the number of training samples is increasing, $C^{m}_{N_c}$ sequences will be already generated for model learning so that the self-permutated sequences will be redundant and play a less role than those baselines. Moreover, when R increases, the loss of target recognition will contribute less to the total objective function as λ keeps unchanged during experiments. Consequently, the accuracy shown in Fig. 8 may not be the best one achieved via model fine-tuning. Nevertheless, the comparison result will still establish the effectiveness of capturing high order correlations and self-permutation. A similar conclusion can also be inferred from the comparison result of RotANet-m2R1 with RotANet-m1 in Fig. 7.

D. Framework Comparison

Finally, we will compare the proposed RotANet with the other SAR ATR algorithms on different conditions to demonstrate its superiority.

1) Results of SOC: The first comparison experiment will be conducted on SOC, whose training and testing information is illustrated in Table. I. We will compare RotANet-m2R1 with four typical models, namely deep learning model: A-ConvNet [14], feature matching classifier: K-nearest neighbor (K-NN), kernel method: support vector machines with the

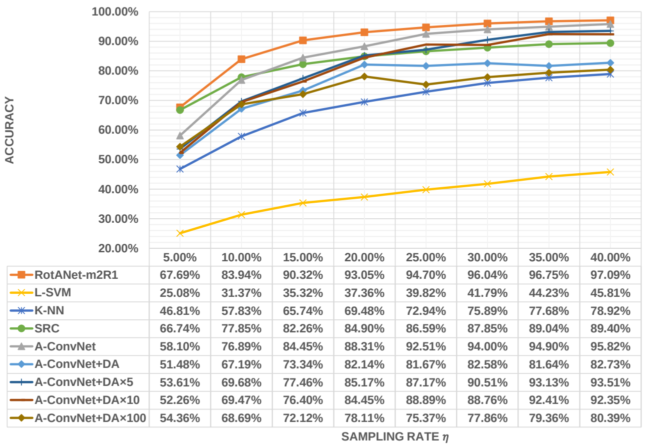


Fig. 9. Comparison results of different SAR ATR frameworks on SOC.

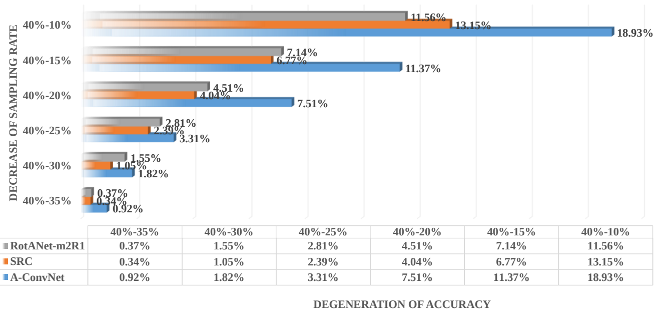


Fig. 10. Accuracy degeneration comparison as sampling rate decrease.

TABLE V

Comparison results on SOC trained with limited samples

Model	30 samples per class	60 samples per class
RotANet-m2R1	**79.66%**	**92.05%**
A-ConvNet	70.37%	85.73%
SRC	75.55%	87.36%
LC-KSVD	72.66%	85.75%
FDDL	74.11%	85.79%
GoogLeNet	78.83%	88.88%
ResNeXt-101	78.17%	87.17%

TABLE VI

Confusion matrix of 5 times results of RotANet-m2R1 trained with 30 samples

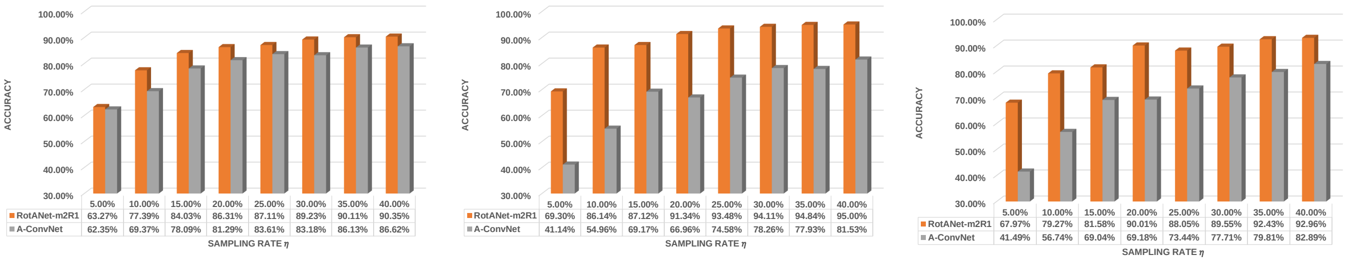
	2S1	BMP-2	BRDM-2	BTR-60	BTR-70	D7	T-62	T-72	ZIL-131	ZSU-234
2S1	973	65	30	13	45	27	95	41	75	6
BMP-2	51	2200	80	31	128	19	109	251	30	36
BRDM-2	26	85	1060	29	75	2	14	16	60	3
BTR-60	5	18	13	785	77	14	29	18	1	15
BTR-70	34	17	4	29	877	2	5	6	6	0
D7	5	2	1	0	5	1284	24	3	2	44
T-62	30	32	8	4	5	25	991	133	61	76
T-72	20	199	13	52	20	6	226	2309	45	20
ZIL-131	65	17	12	17	37	24	54	87	1041	16
ZSU-234	2	3	4	8	0	39	41	14	21	1238

linear kernel (L-SVM) and regression model: sparse representation classifier (SRC). K-NN, a conventional and typical non-parameter classifier will not significantly suffer from the over-fitting risk so that it is suitable for the situation of limited training samples. L-SVM is a standard benchmark classification method for handling limited training samples, and SRC is a parameter-free classifier based on the ridge regression model, which introduces the sparsity induced regularization to address the problem of over-fitting. A-ConvNet is the state-of-the-art deep learning framework for SAR ATR as a baseline algorithm for comparison. Additionally, the widely exploited data augmentation trick of manually rotating the training samples with interpolation is also exploited during A-ConvNet training to show the difference between normal RGB image classification and SAR ATR. We will augment the training set with different numbers including $\mathcal{C}_{N_c}^2$ (A-ConvNet+DA), $5 \times \eta N_c$(A-ConvNet+DA$\times$5), $10 \times \eta N_c$ (A-ConvNet+DA$\times$10), and $100 \times \eta N_c$ (A-ConvNet+DA$\times$100). In the experiment, we vary the sampling rate from 5% to 40% and depict the accuracy curves of different algorithms in Fig. 9. From the results, our proposed model can outperform all comparison methods at all sampling rates. From the figure, RotANet-m2R1 can achieve over 9% higher accuracy than A-ConvNet in the case of 5% training samples. It demonstrates that RotANet-m2R1 is much less prone to over-fitting than A-ConvNet. In this case, SRC also outperforms the A-ConvNet over 8% since it does not rely on the learning process. With the increase in the sampling rate, the accuracy of A-ConvNet gradually exceeds SRC and approximates RotANet-m2R1. Considering the results of A-ConvNet variants trained with the data augmentation trick, we can conclude that the trick plays no role in model learning but will empirically decrease the recognition accuracy. It demonstrates the specificity of the SAR image processing in comparison with the RGB image and further supports our motivation. K-NN and L-SVM are not effective in the scenario, and their accuracies are lowest among all compared algorithms.

In addition to sampling rate, we will directly select 30 and 60 samples per class for model learning in order to verify the performance in the limited training samples. To this end, some other state-of-the-art learning based shallow and deep models are compared, including label-consistency K-SVD (LC-KSVD) [53], Fisher Discriminative Dictionary Learning (FDDL) [54], GoogLeNet [55] and ResNeXt-101 [56]. The comparison results are summarized in Table V, and the confusion matrices of RotANet-m2R1 are summarized in Table VI and VII. We can see from the results that RotANet will outperform all compared algorithms in two situations.

Next, to verify the model robustness to the decrease of training samples, we compute the accuracy degeneration of two algorithms as η drops from 40% to 10% for comparison, including A-ConvNet and SRC. The results are plotted in Fig. 10. According to the figure, we can see that the SRC will be the most robust model when the training samples decrease in most cases. It is because SRC can be regarded as a parameter-free model that is much less sensitive to the decrease of the sampling rate. On the contrary, RotANet and A-ConvNet are two CNN based models which contain more parameters to be learned. If the training samples are insufficient, they are more prone to over-fitting than SRC. From the result, when η drops from 40% to 15%, SRC can acheve a lower degeneration of accuracy than CNN based models. Nevertheless, RotANet can also perform better than SRC if η drops from 40% to 10%, which verifies its robustness in the case of limited training samples.

Finally, to further establish the superiority of RotANet in learning with limited samples, the performances of some other published SAR ATR algorithms with fully sampled training data are also compared, including dictionary learning based-joint dynamic sparse representation (JDSR) [9], autoencoder (AE) with convolutional neural network (AE-CNN), an AE with linear SVM (AE-LSVM) method, a Fisher regularized AE with linear SVM (FAE-LSVM) method [20], and Nonlinear Analysis Cosparse Model (NACM) based two classification frameworks, namely NACM-ML and NACM-MAP [10]. The comparison results are presented in Table IX and the corresponding confusion matrix of RotANet-m2R1 of the total results of 5 times experiments is summarized in Table VIII. It is evident from the average results in the table that RotANet



(a)

(b)

(c)

Fig. 11. Comparison results on different EOCs. (a) EOC-1. (b) EOC-2. (c) EOC-3.

TABLE VII
CONFUSION MATRIX OF 5 TIMES RESULTS OF ROTANET-M2R1 TRAINED WITH 60 SAMPLES

	2S1	BMP-2	BRDM-2	BTR-60	BTR-70	D7	T-62	T-72	ZIL-131	ZSU-234
2S1	1180	25	26	12	25	2	48	14	36	2
BMP-2	34	2631	38	54	24	5	36	76	29	8
BRDM-2	30	37	1238	6	10	4	2	7	36	0
BTR-60	3	3	9	890	37	1	3	19	1	9
BTR-70	8	11	0	5	953	0	0	3	0	0
D7	6	0	0	1	0	1336	2	0	4	21
T-62	6	3	4	9	0	5	1238	44	34	22
T-72	7	95	6	24	1	0	93	2656	11	17
ZIL-131	33	2	2	4	0	10	13	11	1274	21
ZSU-234	0	0	2	1	0	11	4	0	5	1347

TABLE VIII
CONFUSION MATRIX OF 5 TIMES RESULTS OF ROTANET-M2R1 TRAINED WITH 40% SAMPLES

	2S1	BMP-2	BRDM-2	BTR-60	BTR-70	D7	T-62	T-72	ZIL-131	ZSU-234
2S1	1290	20	21	1	4	0	25	3	5	1
BMP-2	1	2903	1	1	1	3	3	15	5	2
BRDM-2	14	30	1273	6	3	0	3	2	39	0
BTR-60	1	2	0	919	20	0	2	25	1	5
BTR-70	6	6	0	4	961	0	0	3	0	0
D7	4	0	0	0	0	1355	1	0	6	4
T-62	8	0	2	2	0	4	1290	40	10	9
T-72	1	18	3	1	0	0	7	2877	1	2
ZIL-131	15	2	0	3	0	10	8	2	1327	3
ZSU-234	0	2	0	0	0	4	1	5	3	1355

can achieve the highest accuracy of 97.09% with only 40% sampled samples while all compared models can merely obtain much lower results even if they exploit 100% sampled data in the training phase. In summary, according to the above experimental results, the superiority of RotANet on SAR ATR can be verified.

2) Results of EOC: The next group of experiments will be tested on three types of EOCs, namely EOC-1 of Table II for large depression variation, EOC-2 of Table III for configuration variants and EOC-3 of Table IV for version variant. The corresponding comparison results with A-ConvNet are plotted in Figs. 11. Note from the three Figures, RotANet-m2R1 clearly shows its robust superiority in comparison with A-ConvNet as it can always achieve a higher recognition accuracy in three conditions with varying sampling rates. In particular, we can observe from Fig. 11(b) and Fig. 11(c) that the accuracy of RotANet-m2R1 is over 10% more than that of A-ConvNet. Overall, our proposed framework can always obtain better results on three conditions of large depression angle, configuration variants and version variant.

TABLE IX
COMPARISON RESULTS ON SOC

Model	Training Rate η	Overall Accuracy
RotANet-m2R1	**40%**	**97.09%**
DL-JDSR	100%	91.48%
AE-CNN	100%	84.70%
AE-LSVM	100%	87.04%
FAE-LSVM	100%	91.29%
NACM-ML	100%	91.57%
NACM-MAP	100%	94.22%
SRC	100%	95.04%
LC-KSVD	100%	95.13%
FDDL	100%	94.07%
GoogLeNet	100%	96.07%
ResNeXt-101	100%	96.78%

V. CONCLUSION

In this paper, we present an efficient learning framework termed RotANet for SAR ATR based on rotation awareness. Instead of extracting the rotation-invariant features as the conventional model does, RotANet newly focuses on the rotation equivariant features. We newly design a weakly encoding scheme to explicitly characterize the pose variations among multiple intra-class samples. According to this encoding scheme, we further develop a self-supervised task to make the model learn to predict the rotational pattern of a baseline target sequence and then autonomously generalize to the other self-permuted ones for validation. It follows that this task essentially contains a learning and a self-validation process for human-like rotation awareness by exploiting the intrinsic relational constraints among sequences. This self-supervised task is exploited to regularize the learned feature domain of RotANet in conjunction with an individual target recognition task to improve the generalization ability of the features. Extensive experiments on the MSTAR benchmark database demonstrate the effectiveness of our proposed framework. Thanks to the self-supervised module, RotANet will be more appropriate for the case of deficient rotating targets in the training phase. From the experimental results, we can see that when the sampling rate is 5%, RotANet can still achieve over 67% accuracy on SOC.

In our future work, we will directly handle the bi-level optimization (5) to further enhance the recognition performance. Additionally, the proposed self-supervised strategy will be also devoted to the target recognition task of other image sources. Finally, the azimuth angle of these 5% samples is empirically uniformly sampled from 0 to 2π in this paper. Nevertheless, we may only exploit a few training targets with partial azimuth angles for practical consideration. In this case, SAR scattering signatures will be also considered to address this challenge task.

Acknowledgment

The authors would like to thank the anonymous reviewers for their careful review and helpful suggestions to improve the manuscript. We would also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] 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,” vol. 4, pp. 6014–6058, 2016.

[2] Leslie M Novak, Gregory J Owirka, William S Brower, and Alison L Weaver, “The automatic target-recognition system in SAIP,” *Lincoln Laboratory Journal*, vol. 10, no. 2, 1997.

[3] 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.

[4] Katsushi Ikeuchi, Takeshi Shakunaga, Mark D Wheeler, and Taku Yamazaki, “Invariant histograms and deformable template matching for SAR target recognition,” in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1996, pp. 100–105.

[5] 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.

[6] Jianxiong Zhou, Shi Zhiguang, Cheng Xiao, and Qiang 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] Lee C Potter and Randolph L Moses, “Attributed scattering centers for SAR ATR,” *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 79–91, 1997.

[8] Kangning Du, Yunkai Deng, Robert Wang, Tuan Zhao, and Ning Li, “SAR ATR based on displacement-and rotation-insensitive CNN,” *Remote Sens. Lett.*, vol. 7, no. 9, pp. 895–904, 2016.

[9] Yongguang Sun, Lan Du, Yan Wang, Yinghua Wang, and Jing Hu, “SAR automatic target recognition based on dictionary learning and joint dynamic sparse representation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1777–1781, 2016.

[10] 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.

[11] B. Ding, G. Wen, C. Ma, and X. Yang, “An efficient and robust framework for sar target recognition by hierarchically fusing global and local features,” *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5983–5995, 2018.

[12] G. Dong and G. Kuang, “Classification on the monogenic scale space: Application to target recognition in sar image,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2527–2539, 2015.

[13] Haichao 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.

[14] Sizhe Chen, Haipeng Wang, Feng Xu, and Ya-Qiu 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] M. Liu, S. Chen, J. Wu, F. Lu, X. Wang, and M. Xing, “SAR target configuration recognition via two-stage sparse structure representation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2220–2232, April 2018.

[16] Christopher M Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., 2006.

[17] Ganggang Dong, Gangyao Kuang, Na Wang, Lingjun Zhao, and Jun 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.

[18] Ganggang Dong and Gangyao 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.

[19] Jun Ding, Bo Chen, Hongwei Liu, and Mengyuan Huang, “Convolutional neural network with data augmentation for SAR target recognition,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, 2016.

[20] 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.

[21] 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.

[22] 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.

[23] K. Kavukcuoglu, M. A. Ranzato, R. Fergus, and Yann Le-Cun, “Learning invariant features through topographic filter maps,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[24] 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.

[25] Z. Wen, B. Hou, and L. Jiao, “Discriminative dictionary learning with two-level low rank and group sparse decomposition for image classification,” *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3758–3771, Nov 2017.

[26] Lin Zhao, Kefeng Ji, Kang Miao, Xiangguang Leng, and Huanxin Zou, “Deep convolutional highway unit network for SAR target classification with limited labeled training data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1091–1095, 2017.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *J. Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] G. Dong, G. Kuang, N. Wang, and W. Wang, “Classification via sparse representation of steerable wavelet frames on grassmann manifold: Application to target recognition in sar image,” *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2892–2904, 2017.

[29] Z. Wen, B. Hou, and L. Jiao, “Discriminative nonlinear analysis operator learning: When cosparse model meets image classification,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3449–3462, 2017.

[30] Zhongling Huang, Zongxu Pan, and Bin Lei, “Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data,” *Remote Sensing*, vol. 9, no. 9, pp. 907, 2017.

[31] R. Shang, J. Wang, L. Jiao, R. Stolkin, B. Hou, and Y. Li, “SAR targets classification based on deep memory convolution neural networks and transfer parameters,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2834–2846, Aug 2018.

[32] C. Zhong, X. Mu, X. He, J. Wang, and M. Zhu, “SAR target image classification based on transfer learning and model compression,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 412–416, March 2019.

[33] 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.

[34] S. Zhang, Z. Wen, Z. Liu, and Q. Pan, “Rotation awareness based self-supervised learning for SAR target recognition,” in *IGARSS 2019 - 2019 IEEE Int. Geosci. and Remote Sens. Symposium*, July 2019, pp. 1378–1381.

[35] 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.

[36] Huan Ruohong and Yang Ruliang, “SAR target recognition based on MRF and gabor wavelet feature extraction.,” 2008.

[37] Y. Zhang, X. Tian, Y. Li, X. Wang, and D. Tao, “Principal component adversarial example,” *IEEE Trans. Image Process.*, vol. 29, pp. 4804–4815, 2020.

[38] Gee-wah Ng, *Brain-Mind Machinery: Brain-inspired Computing and Mind Opening*, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2009.

[39] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton, “Dynamic routing between capsules,” in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017.

[40] Kai Sheng Tai, Peter Bailis, and Gregory Valiant, "Equivariant transformer networks," in *Proceedings of the IEEE conf. Machine Learning (ICML)*, 2019.

[41] Jan Eric Lenssen, Matthias Fey, and Pascal Libuschewski, "Group equivariant capsule networks," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2018.

[42] Taco S. Cohen and Max Welling, "Group equivariant convolutional networks," in *Proceedings of the IEEE conf. Machine Learning (ICML)*, 2016.

[43] Junying Li, Zichen Yang, Haifeng Liu, and Cai Deng, "Deep rotation equivariant network," *Neurocomputing*, vol. 290, pp. 26–33, 2017.

[44] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[45] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[46] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[47] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," in *European Conference on Computer Vision*, 2016.

[48] X. Liu, J. v. d. Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in CNNs by self-supervised learning to rank," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1862–1878, Aug 2019.

[49] Yi Ma, Stefano Soatto, Jana Kosecka, and S. Shankar Sastry, *An Invitation to 3-D Vision-From Images to Geometric Models*, Springer-Verlag New York, 2004.

[50] J. H. Bae, D. Yeo, J. Yim, N. S. Kim, C. S. Pyo, and J. Kim, "Densely distilled flow-based knowledge transfer in teacher-student framework for image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 5698–5710, 2020.

[51] Y. Huang, W. Wang, L. Wang, and T. Tan, "Conditional high-order boltzmann machines for supervised relation learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4297–4310, Sep. 2017.

[52] Eric R. Keydel, Shung W. Lee, and John T. Moore, "MSTAR extended operating conditions - a tutorial," *Proc Spie*, 1996.

[53] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651–2664, 2013.

[54] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *2011 International Conference on Computer Vision*, 2011, pp. 543–550.

[55] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[56] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zaidao Wen (M'18) received the B.S. and Ph.D in electronic engineering from Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China in Xidian University, Xi'an, China, in 2010 and 2017, respectively. Between 2014-2015, he was an Exchange Ph.D. student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Spain. His doctoral dissertation was granted the Excellent Doctoral Dissertation of Shaanxi Province in 2020. He is currently an associate professor in the School of Automation, Northwestern Polytechnical University (NPU). His current interests include compressed sensing and sparse model, cognitive machine learning, Synthetic Aperture Radar image interpretation, and multisource automatic target recognition.

Zhun-Ga Liu Zhun-Ga Liu was born in China. He received the bachelors, masters, and Ph.D. degrees from Northwestern Polytechnical University (NPU), Xian, China, in 2007, 2010, and 2013, respectively. He has been a Professor with the School of Automation, NPU, since 2017. His current research interests mainly focus on pattern recognition, information fusion and belief functions.

Shuai Zhang received the Bachelor's and Master's degrees from Zhengzhou University and Northwestern Polytechnical University, China, in 2017 and 2020, respectively. His current interests include deep learning and automatic target recognition.

Quan Pan received the Bachelor's degree in automatic control from Huazhong University of Science and Technology, Wuhan, China, in 1982, and the Master's and Doctor's degrees in control science and engineering from NPU, Xi'an, China, in 1991 and 1997, respectively.

Since 1998, he has been a Professor in NPU. His main research interests include information fusion and pattern recognition.