A Novel Category Discovery Method for SAR Images Based on an Improved UNO Framework

Mingyao Chen, Tianpeng Liu, and Li Liu, *Senior Member, IEEE*

***Abstract*—In recent years, synthetic aperture radar automatic target recognition (SAR ATR) has been widely researched for its ability to achieve high-performance target classification through supervised training, facilitating battlefield reconnaissance, and intelligence generation. When dealing with unknown class data for ATR model training, significant time and effort are typically required for manual interpretation and labeling. However, when unknown class data shares the same domain as the known labeled data in the library, leveraging their shared deep semantic knowledge can enable automatic classification and labeling of the unknown class data. In this article, we investigate the novel category discovery (NCD) problem in SAR images, using labeled data to guide the clustering process of new class data. Specifically, we utilize the Unified Objective function training framework to address the training imbalance between labeled and unlabeled data in the NCD process, incorporating various improvements on this foundation. Through a binary segmentation-based strategy, we effectively mitigate the interference of "noise pairs" with significant semantic differences on the model's self-supervised pretraining. In addition, we introduce multicrop consistency loss and equal distance loss to impose constraints on training by leveraging intraclass and interclass relationships in the latent space, thereby obtaining representations with higher interclass separability. Our method achieves state-of-the-art clustering performance in multiple scenarios. Extensive experimental results on the MSTAR benchmark dataset demonstrate the effectiveness of the proposed methods.**

***Index Terms*—Deep embedded clustering, novel category discovery (NCD), synthetic aperture radar (SAR) target recognition.**

I. Introduction

In recent years, the rapid development of synthetic aperture radar (SAR) systems has enabled the collection of a large amount of image data in a short period of time [1], [2], [3], which can be used for training intelligence generation or automatic target recognition (ATR) models based on deep learning [4], [5], [6], [7], [8], [9], [10], [11], [12]. These processes often require sufficient labeled training data. However, in many task scenarios, obtaining target labels requires a significant amount of manpower and time and is difficult to meet real-time requirements. For example, in the target reconnaissance process, SAR may collect various unknown categories of new vehicles or variants of known categories of vehicles, which are difficult to differentiate and label manually. Fortunately, these targets are typically within the same domain or share high-level semantic information with known categories, enabling the differentiation and labeling of unknown class targets with the assistance of information from known categories [13], [14]. Therefore, effectively leveraging the target information of labeled known classes to annotate a new batch of unknown class data has become a challenging task.

The aforementioned scenario can be classified as a novel category discovery (NCD) problem [15], i.e., training a model to utilize the latent knowledge of labeled known class data so that the model can discover new classes and cluster them in unlabeled new class data. In contrast to the semisupervised clustering problem [16], [17], [18], [19], where labeled and unlabeled data share the same label space, NCD involves labeled and unlabeled data categories that do not overlap, adding a higher level of complexity to the task. The existing NCD methods can be roughly categorized into two-stage based methods and single-stage based methods [20], [21]. Early research on NCD typically employed two-stage approaches, such as KCL [22], MCL [23], and DTC [15]. In the first stage, known class data is used to pretrain the model, followed by fine-tuning and clustering with unknown class data in the second stage. The advantage of such methods lies in the separate use of known and unknown class data, catering to application scenarios where both types of data cannot be obtained simultaneously in the same stage [24]. However, this type of approach is highly sensitive to the effectiveness of pretraining, and the final clustering results are greatly impacted by the initial discriminative capacity of unknown class data within the pretrained model. In addition, collapse is prone to occur, i.e., all unknown class samples are grouped into a single category, necessitating a more rigorous training paradigm in the second stage [24], [25].

In view of the above problems, the recently proposed NCD methods are mostly single-stage, where both known and unknown class data are utilized concurrently during training, such as AutoNovel [26], UNO [27], MEDI [28], NCL [29], and NSCL [14]. In comparison to two-stage methods, these approaches can better leverage the similarities and differences between known and unknown class data, leading to improved clustering accuracy for unknown classes. For instance, UNO, as a classic single-stage method, integrates the one-hot labels of known classes with the pseudolabels of unknown classes obtained by manifold regularization to form a unified label representation, enabling joint training through cross-entropy

Received 19 June 2024; revised 30 July 2024; accepted 17 August 2024. Date of publication 21 August 2024; date of current version 5 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61921001, Grant 62022091, and Grant 62201588. *(Corresponding author: Tianpeng Liu.)*

The authors are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: liutianpeng2004@nudt.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3446815

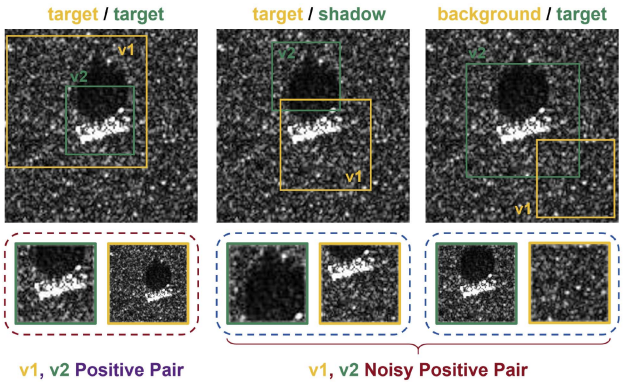


Fig. 1. Noisy pair problem in the pretraining stage caused by Random-Resized-Crop. The semantics of the two cropped views on the left side are both "target." The semantics of the two cropped views in the middle are "shadow" and "target," respectively. And the semantics of the two cropped views on the right side are "target" and "background," respectively. The vast semantic differences in the last two cases will produce noisy pairs.

loss. This effectively addresses the bias caused by the imbalance in supervision strength between known and unknown classes.

Based on this framework, researchers have made numerous improvements in training paradigms and data feature space distributions, leading to enhanced clustering accuracy and generalization. Li et al. [20] enforced consistency between different views via symmetric Kullback–Leibler divergence (sKLD) and introduced the negative sKLD between known and unknown class data within the same batch as loss, thereby elevating the overall discrimination between known class and unknown class samples. In addition, in the scenario of NCD of SAR targets, Huang et al. [30] improved the data augmentation techniques in the pretraining of UNO, enhancing the initial discriminability between known and unknown classes. They also established a class sample library and used prototype replay to alleviate catastrophic forgetting of known classes during subsequent training.

The aforementioned methods have shown promising results in the field of SAR data analysis, but there are still two issues that remain to be addressed. First, in the SAR image augmentation techniques used in the above methods, Random-Resized-Crop has been proven effective during the pretraining stage [20], [24], [27], [30]. It contrasts the local and global characteristics of the targets, forcing the model to focus more on the consistency and difference of the structural features. However, unlike optical natural images, the position and size of SAR targets in image slices have high similarity, and the background accounts for a large proportion. Therefore, employing Random-Resized-Crop may capture parts with significant semantic differences from the SAR target itself [31], such as background or shadows, leading to the generation of "noisy pairs," as illustrated in Fig. 1. When these noisy pairs are used for comparison or imposing consistency constraints, they can detrimentally impact the model training. Second, the abovementioned methods focus on the discriminability between known and unknown classes, promoting separation through sKLD between the two types of data. However, these methods do not address the separability between different classes. Within the feature space obtained through the UNO framework, the separability between unknown classes is significantly lower than that between known classes, leading to low clustering accuracy. In an ideal feature space data distribution, the distances between classes of unknown data should be similar to those between classes of known data. That is, through training, all class representations should possess equal separability.

Motivated by the need of enhancing interclass separability, and inspired by the recent advancements in deep clustering, in this article we propose an improved SAR target NCD method based on the UNO training framework. We first utilize all SAR images to construct positive pairs through image augmentation for self-supervised pretraining. To address the issue of noisy pairs, we employ a simple and effective threshold-based method to filter out noisy views that do not contain targets. Subsequently, supervised pretraining is conducted using known class data to learn semantic information from labels. The last step is to construct a unified label based on the UNO framework, allowing both types of data to participate in supervised training together, and introduce equal distance loss (EDL) based on the assumption that each class should have equal separability. By setting equidistant points in the feature space, the distance between the class prototype and the equidistant points is penalized to guide each class towards the nearest equidistant point position, which enhances the discriminability between different unknown classes. Furthermore, to address the incompleteness of the augmentation view constraint, we propose multicrop consistency loss to constrain the consistency of multiple enhanced views from the feature space. In conclusion, the main contributions of our work are summarized as follows.

1) We propose a novel method for SAR target NCD based on the improved UNO training framework, which achieves the best results in the SAR field.
2) We enhance the SAR image augmentation techniques in the pretraining stage and implement a threshold-based algorithm to filter out and discard augmented views with large background or shadows, so as to prevent the negative impact of semantic differences between noisy pairs on model training.
3) We use multicrop to generate diversified views and introduce multicrop consistency loss and intraclass loss to jointly constrain the consistency representation of different views in terms of KL divergence and Euclidean distance in the feature space.
4) To address the issue that the discrimination between unknown classes in the feature space is much lower than that of known classes, we propose equal distance loss based on the UNO framework. By forcing each class to move to points with equal distances in the latent manifold, the unknown interclass discrimination is increased.

II. Related Works

A. Novel Category Discovery (NCD)

NCD was first studied in the context of deep embedded clustering (DEC) by Hsu et al. [22]. In [22] and [23], NCD was introduced as a new task, where the goal was to cluster a

new dataset whose classes were similar to an existing labeled dataset but did not overlap with it. Han et al. [15] formally proposed the concept of NCD based on this setting and studied it as a separate problem. This concept has widespread applications in fields such as open-set recognition (OSR) [24], [32], class-incremental learning [30], [33], signal classification [34], and semantic segmentation [35].

The existing NCD methods can be roughly categorized into two-stage methods and single-stage methods, with the main difference lying in whether labeled and unlabeled data are utilized in the same stage. Two-stage methods [15], [22], [23], [36] often acquire semantic knowledge of labeled data through supervised learning in the pretraining stage, and then use transfer learning in the subsequent stage to guide the clustering of unlabeled data. Specifically, KCL [22] and MCL [23] introduce a framework that facilitates transfer learning across domains and tasks, which leverages the pairwise similarity to represent categorical information. DTC [15] conducts supervised training in the initial stage, and then fine-tunes the distribution based on Kullback–Leibler divergence using unlabeled data in the subsequent stage to obtain more refined and discriminative representations. In summary, two-stage methods separate the use of labeled and unlabeled data, aiming to enhance clustering effectiveness through knowledge transfer. However, research indicates that the knowledge provided by supervised training may not all be beneficial for clustering, as it is influenced by factors such as the levels of semantic relevance between labeled and unlabeled data [37] as well as the knowledge relationships among known classes [14], [38]. In addition, two-stage training can result in the final clustering outcome heavily relying on the initial separability of unknown class data in the pretrained model. Therefore, subsequent studies often choose single-stage methods, where labeled and unlabeled data are used simultaneously [14], [20], [26], [27], [28], [29], [38], [39]. AutoNovel [26] first conducts self-supervised pretraining using all data, and then leverages the assumption that the feature vectors of the same class of samples have similar high activation positions to construct pseudolabels. And it finally uses binary cross-entropy loss to discover new classes. NCL [29], NSCL [14], and OpenMix [39] introduce contrastive learning to enhance the discriminability between known and unknown classes by comparing samples and their augmented views, with the key difference being in the construction of the loss function. MDEI [28], on the other hand, reduces the requirement for unknown class data by incorporating metalearning techniques into the NCD task. Instead of using multiple objectives, UNO [27] introduces a unified objective function to transfer knowledge from the labeled data to unlabeled data. IIC [20] improves upon UNO by leveraging sKLD. It enforces constraints on interclass separability and intraview consistency, further enhancing clustering accuracy.

The application of NCD in SAR images is often combined with OSR tasks, clustering unknown class samples identified through the OSR stage. Dai et al. [40] were the first to apply this approach to ship targets, training with both known and unknown class data to address OSR and NCD tasks simultaneously. Chen et al. [24] achieved more accurate estimation of the number of unknown classes and clustering by improving DTC. The above methods fail to address the collapse of the training process, which easily led to all unknown class samples being classified into the same class. CNT [30] improved the pretraining method using VICReg [25], effectively mitigating the aforementioned issue, and focusing on the overall separability of known and unknown classes. In contrast to them, we further consider the separability issue among all classes, which is not achieved in existing SAR NCD methods. In addition, we have made breakthroughs in both pretraining and consistency constraints compared to the above methods.

B. Self-Supervised Pretraining

Self-supervised pretraining refers to training a network by extracting supervised information to construct subtasks without using data labels. The training method is typically selected based on the requirements of downstream tasks to learn representations beneficial for them [26], [30], [41], [42]. Early self-supervised pretraining methods are often based on image-only pretext tasks, utilizing the supervision information within a single image to enhance the model's semantic understanding. One common approach is using context information, where two patches are randomly extracted from a single image, and the model is then tasked with predicting their relative positions to better understand the relationships between scenes and objects [43]. Another approach involves image reconstruction, creating prediction tasks by discarding image colors or adding masks to enhance the model's overall semantic understanding of images [44], [45], [46], [47]. Moreover, Gidaris et al. [48] devised a task for recognizing image rotation angles, requiring the model to predict the angle of the image[26]. This simple method does not leave any explicit low-level clues, so the model must recognize and focus on the main objects in the image and their semantic relationships with the background to complete the task effectively. While these methods have shown significant improvements in representation quality, they lack a standardized training paradigm and may not generalize well to different types of images or downstream tasks.

With the introduction of SimCLR [49] and MoCo [50] frameworks, contrastive-based self-supervised pretraining has been widely studied. This type of method most often uses Siamese architectures, in which the two branches have identical architectures and share weights. By employing techniques such as image augmentation to obtain different views of the same image, positive and negative examples are constructed for supervised training. SimCLR uses augmented views from the same image as positive pairs and views from other images within the same batch as negative pairs with respect to that image. It employs the InfoNCE loss for self-supervised training to minimize the distance between positive pairs and maximize the distance between negative pairs [42], [49], [51]. However, this method faces issues such as high memory usage and its effectiveness being limited by batch size. The series of MoCo [50], [52], [53] decouples the relationship between batch size and the number of contrastive samples, further expanding the applicability of this approach. The above methods maximize the agreement between embedding vectors produced by encoders fed with different views of the same image, but there is still a challenge

to preventing a collapse in which the encoders produce constant or noninformative vectors. Bardes et al. [25] proposed VICReg, which avoids the above two collapse problems by introducing a variety of regularization terms, so it is more suitable for solving the downstream NCD task of SAR images. In this article, we improve this method by filtering out the noisy pairs generated by the Random-Resized-Crop to enhance the quality of the pretrained representation.

III. Proposed Method

This study focuses on addressing the challenge of NCD in SAR images. Following the setting in [20] and [27], given a labeled dataset $\mathcal{D}^l = \{(\boldsymbol{x}_1^l, y_1^l), \ldots, (\boldsymbol{x}_N^l, y_N^l)\}$, the goal is to automatically discover C^u clusters (or classes) in an unlabeled dataset $\mathcal{D}^u = \{\boldsymbol{x}_1^u, \ldots, \boldsymbol{x}_M^u\}$, where each $\boldsymbol{x}_i^l$ in $\mathcal{D}^l$ or $\boldsymbol{x}_i^u$ in $\mathcal{D}^u$ is a SAR image with a single target and $y_i^l \in \mathcal{Y} = \{1, \ldots, C^l\}$ is the corresponding class label of $\boldsymbol{x}_i^l$. Moreover, the classes of labeled data do not overlap with those of unlabeled data.

In the following subsections, we first introduce how to filter out the noisy pairs in pretraining through our proposed algorithm to obtain a better pretrained model for subsequent processes (Section III-A). Next, we illustrate how to use $\mathcal{D}^l$ to estimate the number of unlabeled classes when the number is unknown (Section III-B). Then, we introduce the UNO training framework adopted for the NCD task and the effective improvement techniques applied, such as intraclass loss and interclass loss [20] (Section III-C). Moreover, we introduce the concept of equidistant points and guide the distances between classes to be equal using the equidistant loss, further enhancing the separability of unlabeled classes (Section III-D). Furthermore, we explain how to enhance the representation quality by combining the multicrop technique with the swapped prediction algorithm in UNO, and propose multicrop consistency loss to constrain the consistency of different views (Section III-E). Finally, we summarize the entire training process and overall objective (Section III-F).

A. Pretraining With Noisy Pair Filtering Algorithm

The first step of our work is to use all the data for pretraining to obtain a representation that implicitly includes the semantic understanding of the sample itself. The training initially employs self-supervised learning to avoid biases caused by the strong semantic information contained in labels. According to [30], we utilize VICReg [25] as the model to alleviate potential collapse issues in downstream clustering tasks. Specifically, for a sample $\boldsymbol{x}$, we generate a pair of views through image augmentation, obtain representations Z_1 and Z_2 through an encoder, and proceed with training using the following loss function:

$$\ell(Z_1, Z_2) = \lambda s(Z_1, Z_2) + \mu[\upsilon(Z_1) + \upsilon(Z_2)] + \nu[c(Z_1) + c(Z_2)] \quad (1)$$

where λ, μ, and ν are hyperparameters controlling the importance of each term in the loss. s is to compute the mean-squared Euclidean distance between each pair of features as the invariance criterion. υ computes the variance of the features, and c is used to constrain the expression of the mean of the diagonal elements of the matrix. To obtain a more diverse view, we combine Random-Resized-Crop, ColorJitter, CoarseDropout, ElasticTransform, ShiftScaleRotate, and RandomRotate90 for image augmentation, and each technique is performed at a set probability.

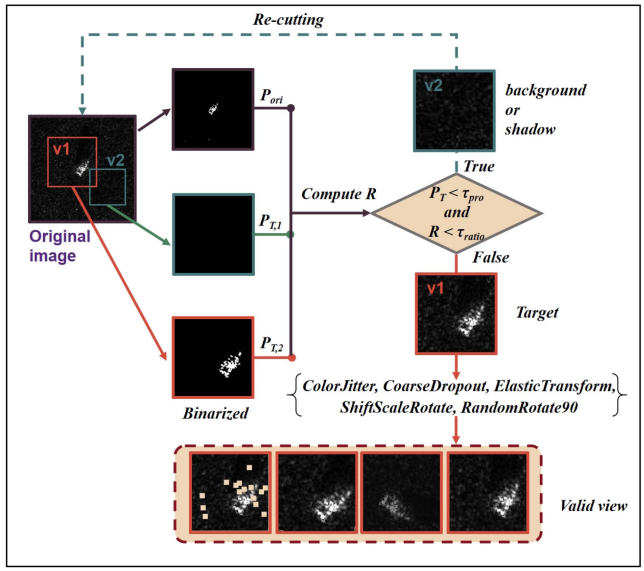


Fig. 2. Augmentation method with noisy pair filtering.

However, augmentation views generated by Random-Resized-Crop may contain only background or shadows. When these views form a noisy pair with the target view, the significant semantic difference between the two will have a negative impact on model training [31]. Fortunately, the pixel values of the background, shadow, and target of the SAR image slice typically exhibit significant differences. Therefore, we can segment the target by binarization and set a threshold to filter out the view without the target, as shown in Fig. 2. Thus, the improved image augmentation algorithm with noisy pair filtering is shown in Algorithm 1.

We first use Random-Resized-Crop to obtain a cropped view $\boldsymbol{x}_T$, and then obtain the binary image of the cropped view and the binary image of the original image according to the set threshold $\tau_{\text{bin}} = 100$ to achieve a rough separation of the target from the background and shadow. The next step is to count the proportion of the target in the cropped view and the original image $\mathcal{P}_T$, $\mathcal{P}_{\text{ori}}$. If $\mathcal{P}_T$ is smaller than the threshold τ_{pro}, it is considered that there is no target in the view. In order to prevent the size of the area taken by the Random-Resized-Crop from being close to the original image, causing $\mathcal{P}_T$ to be small and the view to be mistakenly divided into noisy view, we calculate the ratio $\mathcal{R} = \mathcal{P}_T/\mathcal{P}_{\text{ori}}$, and solve this problem by comparing it with the threshold τ_{ratio}. If the view does not meet the requirements, it is re-cut until a valid view containing the target information is obtained, which will be used for subsequent image augmentation operations.

Following self-supervised pretraining, we can obtain an encoder that has a preliminary comprehension of the data's inherent semantics. Research suggests that supervised training

Algorithm 1: Augmentation Algorithm With Noisy Pair Filtering.

Input: Original SAR image $\boldsymbol{x} \in \mathcal{D}^l \cup \mathcal{D}^u$, image-binary threshold τ_{bin}, proportion threshold τ_{pro}, ratio threshold τ_{ratio}.
Output: augmented view $\hat{\boldsymbol{x}}$.
Initialized: Target proportion of augmented view $\mathcal{P}_T = 0$, the ratio of augmented view proportion to original image proportion $\mathcal{R} = 0$.
Compute the proportion $\mathcal{P}_{ori}$ of pixels in image $\boldsymbol{x}$ that are above the threshold τ_{bin};
while $\mathcal{P}_T < \tau_{pro}$ and $\mathcal{R} < \tau_{ratio}$ **do**
  $\boldsymbol{x}_T = RandomResizedCrop(\boldsymbol{x})$
  Compute the proportion $\mathcal{P}_T$ of pixels in view $\boldsymbol{x}_T$ that are above the threshold τ_{bin}
  Update $\mathcal{R} = \frac{\mathcal{P}_T}{\mathcal{P}_{ori}}$.
end
$\hat{\boldsymbol{x}} = OtherAugmentation(\boldsymbol{x}_T)$.
return $\hat{\boldsymbol{x}}$.

using labeled data facilitates the transfer of label semantics in the subsequent NCD phase, leading to the acquisition of an initial discriminative representation that enhances the guidance of unknown class clustering [15], [26], [27], [30]. Therefore, we further train the encoder on the labeled dataset $\mathcal{D}^l$ using cross-entropy loss, as shown below

$$\ell_{\text{ce}}(\boldsymbol{x}, y) = -\sum_{c=1}^{C} y_c \log(p_c) \tag{2}$$

where C represents the number of classes, which equals C^l during the supervised training stage. y_c and p_c are the cth units of the label y and the network prediction $p = \sigma(f(\boldsymbol{x})/\tau)$, respectively. $f(\cdot)$ represents the network formed by connecting the encoder with a linear classifier consisting of C^l output units. $\sigma(\cdot)$ and τ represent a softmax layer and the temperature of the softmax, respectively. Through the above two stages of pretraining, we obtain an encoder and a linear classifier, which will be utilized as the feature extractor and the labeled classification head in the subsequent NCD task, respectively.

B. Unknown Class Number Estimation

To establish a unified training label for subsequent UNO-based training, we need to know the number of unlabeled classes C^l. When studying NCD problems, the number of unlabeled classes is typically set to be known by default [20], [27], [30]. If this number is unknown, we use the method proposed in [24] for estimation, which can provide relatively accurate estimates when the total number of classes is small. The estimation method is illustrated in Algorithm 2. First, we select N_l classes from $\mathcal{D}^l$ (where C^l classes are known) to form a new training set $\mathcal{D}_T^l$. The remaining $C^l - N_l$ classes form the probe set D_{probe}, which is combined with $\mathcal{D}^u$ for estimating the number of unlabeled classes. Then, we independently build a ResNet-18 network and conduct supervised training on $\mathcal{D}_T^l$ using cross-entropy loss to capture semantic information from the labels. The trained network is then used to extract sample features D^*_{probe} and $\mathcal{D}^{u*}$ from D_{probe} and $\mathcal{D}^u$. Next, we divide D^*_{probe} into a validation probe set $D^*_{\text{probe},v}$ containing N_v classes and an anchor probe set $D^*_{\text{probe},a}$ consisting of the remaining $(C^l - N_l - N_v)$ classes. Finally, we use a semisupervised k-means algorithm with $U + (C^l - N_l)$ centers to estimate the number of classes U in $\mathcal{D}^{u*}$. Specifically, we enforce features in the anchor probe set $D^*_{\text{probe},a}$ to be assigned to the corresponding clusters according to their ground-truth labels during clustering, while features in the validation probe set $D^*_{\text{probe},v}$ are treated as unlabeled data. We carry out experiments by varying the U value multiple times and evaluate them using the average clustering accuracy (ACC_v) and the cluster validity index (CVI), which are defined as follows:

$$ACC_v = \max_{\text{perm}\in P} \frac{1}{N_S} \sum_{i=1}^{N_S} \mathbf{1}\{y_i = perm(\hat{y}_i)\}, \tag{3}$$

$$CVI = \sum_{Z\in\mathcal{D}^{u*}} \frac{b(Z) - a(Z)}{\max\{a(Z), b(Z)\}} \tag{4}$$

where y_i and $\hat{y}_i$ represent the ground-truth label and clustering assignment for sample $Z_i \in \mathcal{D}^*_{\text{probe},v}$, respectively. N_S represents the number of samples in $D^*_{\text{probe},v}$, and P represents the set of all permutations of N_v elements (as a clustering algorithm recovers clusters in an arbitrary order). $a(Z)$ represents the average distance between sample Z and all other samples within the same cluster. $b(Z)$ represents the smallest average distance of sample Z to all samples in any other cluster. By recording the ACC_v on $D^*_{\text{probe},v}$ and the CVI on $\mathcal{D}^{u*}$ under each selected value of U, we can obtain the U value that corresponds to the highest ACC_v and the U value that corresponds to the highest CVI. The average of these two values is taken as the final estimate for the number of unlabeled classes. If the average of these two estimates is noninteger, it is rounded down.

C. SAR Image NCD Framework Based on UNO

We choose UNO as the basic framework for the NCD training phase. The UNO model is primarily divided into three components: 1) an encoder E; 2) a labeled classification head h; and 3) an unlabeled classification head g, as illustrated in Fig. 3(a). E and h utilize the feature extractor and linear classifier obtained in the pretraining stage, while g is composed of a multilayer perceptron (MLP) and a linear classifier with C^u output units. During the training process, first, a pair of augmented views $\widehat{\boldsymbol{x}}_1, \widehat{\boldsymbol{x}}_2$ of the SAR image $\boldsymbol{x}$ are obtained through augmentation. The encoder E is then used to extract features from the two views, obtaining feature vectors Z_1 and Z_2, which will be passed through both labeled head h and unlabeled head g to obtain the corresponding logits $\{\widehat{l}_1^L, \widehat{l}_2^L\} \in \mathbb{R}^{C^l}$ and $\{\widehat{l}_1^U, \widehat{l}_2^U\} \in \mathbb{R}^{C^u}$, respectively. After that, the outputs of the same view in different heads are concatenated to obtain $[\widehat{l}_1^L || \widehat{l}_1^U] \in \mathbb{R}^{C^l+C^u}$ and $[\widehat{l}_2^L || \widehat{l}_2^U] \in \mathbb{R}^{C^l+C^u}$, which are then processed through the softmax layer to obtain probability distributions p_1 and p_2. Through the aforementioned operations, we can obtain the probability distribution of the sample's view across the entire class space. Finally, we construct

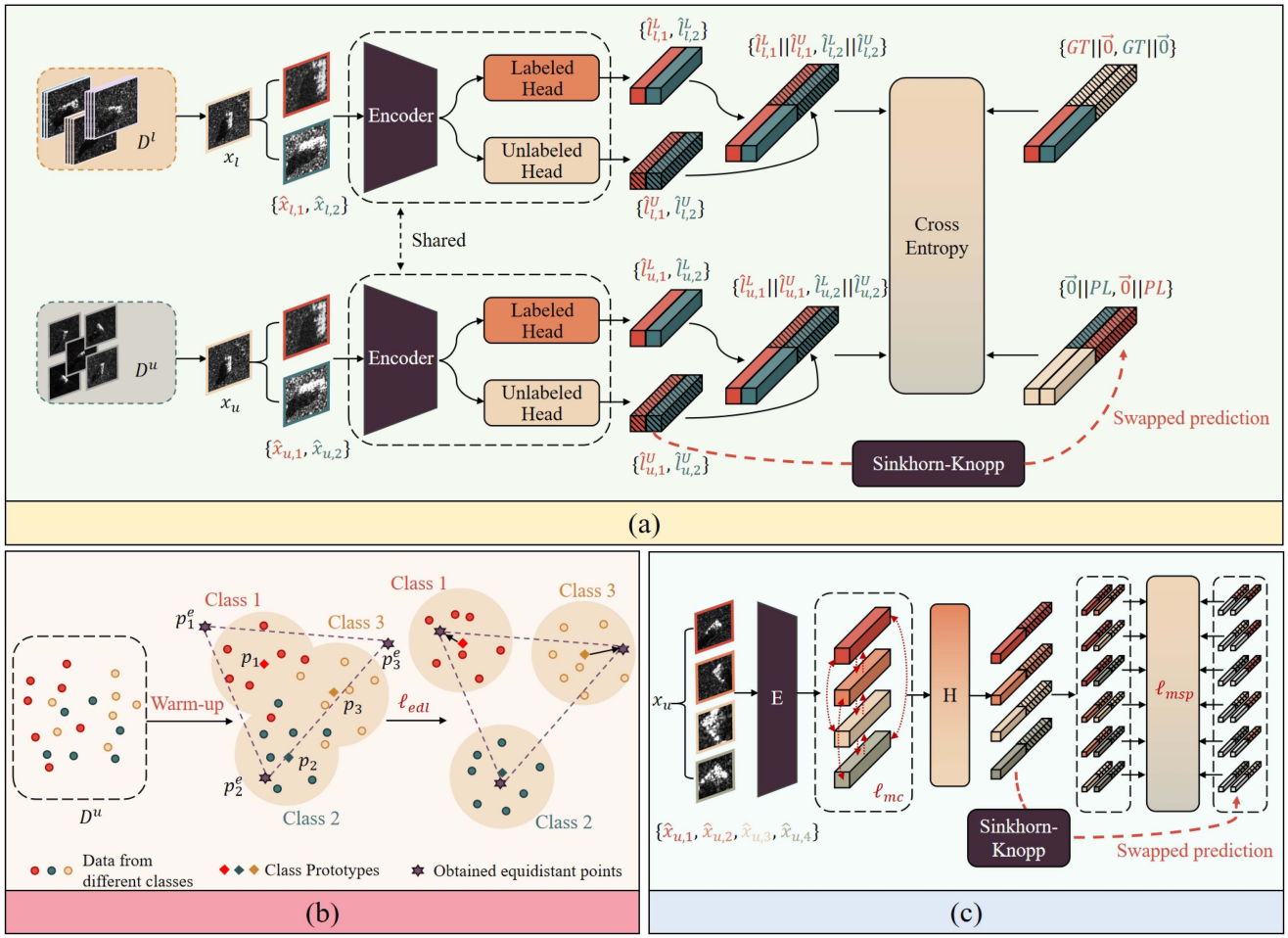


Fig. 3. Framework of our method. (a) Basic framework of UNO. (b) Equal distance loss. (c) Multi-crop swapped prediction and multicrop consistency loss.

the corresponding unified label according to whether the class of the input sample is known. If the class of the SAR image $\boldsymbol{x}$ is known (labeled data), we apply zero-padding to its ground-truth label y^l to obtain the unified labels $y_1 = y_2 = [y^l || \overrightarrow{0}]$. If the class is unknown (unlabeled data), we zero-pad the pseudolabels $\{y_1^u, y_2^u\}$ obtained by using the Sinkhorn–Knopp algorithm [54], which obtains the unified labels $y_1 = [\overrightarrow{0} || y_1^u]$ and $y_2 = [\overrightarrow{0} || y_2^u]$ corresponding to the two views. With the obtained y_1 and y_2, we can utilize the cross-entropy loss from (2) for training, where C equals $C^l + C^u$. This approach offers a unified training paradigm for both labeled and unlabeled data, alleviating the influence of unbalanced training of the two types of data in previous NCD research. In addition, Multihead Clustering and Overclustering techniques have also been applied to improve UNO performance. The utilization of Multihead helps mitigate the issue of individual heads converging to suboptimal clustering configurations, thereby leading to a reduction in such effects and making it beneficial. Overclustering forces the network to produce an alternative partition of the unlabeled data that is more fine-grained, thereby improving the quality of the representations.

Furthermore, UNO employs the method of swapped prediction. When the input sample is unlabeled, by swapping the prediction distributions of the two augmented views, the network can be forced to produce more consistent predictions for different views of the same sample. Therefore, UNO uses the swapped prediction loss to address unlabeled data, as shown below

$$\ell_{sp} = \ell_{\text{ce}}(\widehat{\boldsymbol{x}}_{u,1}, y_2) + \ell_{\text{ce}}(\widehat{\boldsymbol{x}}_{u,2}, y_1). \tag{5}$$

In summary, the overall loss function of UNO is as follows:

$$\ell_{uno} = \begin{cases} \ell_{\text{ce}}(\widehat{\boldsymbol{x}}_1, y) + \ell_{\text{ce}}(\widehat{\boldsymbol{x}}_2, y), & \boldsymbol{x} \in \mathcal{D}^l \\ \ell_{sp}, & \boldsymbol{x} \in \mathcal{D}^u \end{cases}. \tag{6}$$

In addition to the techniques utilized in UNO, we also incorporate two methods proposed in IIC [20]. First, in order to further enhance the overall separability between labeled and unlabeled classes, we have employed interclass loss to explicitly enlarge the distance between each labeled sample and each unlabeled sample using an sKLD distance. For a labeled SAR image $\boldsymbol{x}_l$ and an unlabeled SAR image $\boldsymbol{x}_u$ within the same batch, the loss is as follows:

$$\ell_{\text{inter}} = -\frac{1}{2}(D_{KL}(p^l || p^u) + D_{KL}(p^u || p^l)) \tag{7}$$

Algorithm 2: Unknown Class Number Estimation.

Input: Unlabeled class feature set $\mathcal{D}^{u*}$, anchor probe set $\mathcal{D}^{*}_{probe,a}$, validation probe set $\mathcal{D}^{*}_{probe,v}$.
Output: Final estimation result of the number of unlabeled classes $\hat{U}$.

Initialized: the maximum value of average clustering accuracy $ACC_{max} = 0$, the maximum value of cluster validity index $CVI_{max} = 0$, $\hat{U}_{acc}$, $\hat{U}_{cvi}$.
for $U \leftarrow 0, 1, \ldots, N$ **do**
  Perform semi-supervised k-means on $\mathcal{D}^{*}_{probe,a} \cup \mathcal{D}^{*}_{probe,v} \cup \mathcal{D}^{u*}$ with $U + (C^l - N_l)$ centers, forcing samples in $\mathcal{D}^{*}_{probe,a}$ to assign to the ground-truth class labels;
  Compute ACC_v on $\mathcal{D}^{*}_{probe,v}$ and CVI on $\mathcal{D}^{u*}$;
  if $ACC > ACC_{max}$ **then**
    $\hat{U}_{acc} = U$,
    $ACC_{max} = ACC$.
  end
  if $CVI > CVI_{max}$ **then**
    $\hat{U}_{cvi} = U$,
    $CVI_{max} = CVI$.
  end
end
return $\hat{U} = (\hat{U}_{acc} + \hat{U}_{cvi})/2$.

where p^l and p^u are the probability distributions generated for the $\boldsymbol{x}_l$ and $\boldsymbol{x}_u$, respectively. D_{KL} presents the Kullback–Leibler (KL) divergence. Second, to address the issue of insufficient consistency constraints on the outputs of different augmented views of the same sample in UNO, we utilize the intraclass loss to minimize the distance between the output probability distributions of the two views. For two augmented views $\widehat{\boldsymbol{x}}_1$ and $\widehat{\boldsymbol{x}}_2$ of a sample $\boldsymbol{x}$, the loss is as follows:

$$\ell_{\text{intra}} = \frac{1}{2}(D_{KL}(p_1||p_2) + D_{KL}(p_2||p_1)) \tag{8}$$

where p_1 and p_2 are the probability distributions of the augmented views $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ after passing through the corresponding classification heads and the softmax layer. If $\boldsymbol{x}$ is a labeled class sample, then $p_1 \in \mathbb{R}^{C^l}$ and $p_2 \in \mathbb{R}^{C^l}$. Otherwise, $p_1 \in \mathbb{R}^{C^u}$ and $p_2 \in \mathbb{R}^{C^u}$.

D. Equal Distance Loss (EDL)

Although the introduction of the interclass loss function has enhanced the overall separability between labeled and unlabeled classes, this method does not address the separability between any two classes. Ideally, the distributions of different classes in the latent space should tend to be equidistant. Inspired by [21], we introduce the concept of equidistant points and propose an algorithm for computation. In addition, we design an algorithm to guide samples from each class towards equidistant points during the training process. The entire process is shown in Fig. 3(b). Specifically, we start by utilizing the feature extractor E to acquire the feature vectors of all samples, followed by the application of K-means [55] to generate m centroids as the class prototypes $\boldsymbol{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m\}$, where $m = C^l + C^u$. Then, we postulate that the desired m equidistant points are denoted as $\boldsymbol{P}^e = \{\boldsymbol{p}^e_1, \ldots, \boldsymbol{p}^e_m\}$, with each pair of points having equal and sufficiently large distances d_{eq} between them. Let $d_{\max}$ be the largest pair-wise distance between the prototypes. d_{eq} is set to $\alpha \times d_{\max}$, where α is a hyperparameter greater than 1. Next, let $d_{ij}(\boldsymbol{P}^e)$ be the Euclidean distance between $\boldsymbol{p}^e_i$ and $\boldsymbol{p}^e_j$ in the feature space, as shown below

$$d^2_{ij}(\boldsymbol{P}^e) = ||\boldsymbol{P}^e_i - \boldsymbol{P}^e_j||^2_2. \tag{9}$$

Therefore, our objective can be defined as minimizing the following function:

$$\sigma(\boldsymbol{P}^e) = \sum_{1 \le i < j \le m} \sum w_{ij}(d_{eq} - d_{ij}(\boldsymbol{P}^e))^2 \tag{10}$$

where w_{ij} indicates the relative importance of measurement d_{eq}. Due to the fixed value of d_{eq}, all w_{ij} are set to 1 in the experiment. Hence, (10) can be further written as

$$\sigma(\boldsymbol{P}^e) = \sum_{i<j} d^2_{ij}(\boldsymbol{P}^e) + \sum_{i<j} d^2_{eq} - 2\sum_{i<j} d_{eq} d_{ij}(\boldsymbol{P}^e). \tag{11}$$

According to [56], we define the matrix valued function $\boldsymbol{B}(\boldsymbol{X})$ to simplify the solving process, as shown below

$$\begin{aligned} b_{ij}(\boldsymbol{X}) &= -d_{eq} s_{ij}(\boldsymbol{X}) \quad \text{if } i \neq j, \\ b_{ii}(\boldsymbol{X}) &= \sum d_{eq} s_{ij}(\boldsymbol{X}). \end{aligned} \tag{12}$$

Here

$$s_{ij}(\boldsymbol{X}) = \begin{cases} d^{-1}_{ij}(\boldsymbol{X}), & \text{if } d_{ij}(\boldsymbol{X}) \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

Moreover, we define the matrix $\boldsymbol{V}$ as follows:

$$\begin{aligned} v_{ij} &= -1 \quad \text{if } i \neq j, \\ v_{ii} &= -\sum v_{ij}. \end{aligned} \tag{14}$$

Both $\boldsymbol{B}(\boldsymbol{X})$ and $\boldsymbol{V}$ are real symmetric matrices with nonpositive off-diagonal and nonnegative diagonal elements, whose rows and columns sum to 0. Then, (11) can be written as

$$\sigma(\boldsymbol{P}^e) = \mathbf{tr}\boldsymbol{P}^{eT}\boldsymbol{V}\boldsymbol{P}^e + \sum_{i<j} d^2_{eq} - 2\,\mathbf{tr}\boldsymbol{P}^{eT}\boldsymbol{B}(\boldsymbol{P}^e)\boldsymbol{P}^e \tag{15}$$

where the third term can be bounded as follows [56]:

$$\mathbf{tr}\boldsymbol{P}^{eT}\boldsymbol{B}(\boldsymbol{P}^e)\boldsymbol{P}^e \geq \mathbf{tr}\boldsymbol{P}^{eT}\boldsymbol{B}(\boldsymbol{X})\boldsymbol{X}. \tag{16}$$

The second term of (15) is a constant, so the surrogate function that majorizes $\sigma(\boldsymbol{P}^e)$ can be expressed as follows:

$$\xi(\boldsymbol{P}^e, \boldsymbol{X}) = \mathbf{tr}\boldsymbol{P}^{eT}\boldsymbol{V}\boldsymbol{P}^e - 2\,\mathbf{tr}\boldsymbol{P}^{eT}\boldsymbol{B}(\boldsymbol{X})\boldsymbol{X}. \tag{17}$$

Finally, we propose an algorithm to solve the optimization process of $\boldsymbol{P}^e$, as shown in Algorithm 3.

In Line 3, as $\boldsymbol{V}$ is a singular matrix, we compute its Moore–Penrose inverse matrix $\boldsymbol{V}^+$ for subsequent calculations. Subsequently, we optimize through iterative alternation. In each iteration, we first use the current $\boldsymbol{P}^e$ as $\boldsymbol{X}$ to compute the matrix function $\boldsymbol{B}(\boldsymbol{X})$. In the second step, we substitute $\boldsymbol{B}(\boldsymbol{X})$ and $\boldsymbol{X}$ into $\xi(\boldsymbol{P}^e, \boldsymbol{X})$, setting the derivative of $\xi(\boldsymbol{P}^e, \boldsymbol{X})$ with respect to $\boldsymbol{P}^e$ to zero, and derive the updated $\boldsymbol{P}^e$ with the minimum

Algorithm 3: Calculating Equidistant Points.

Input: Expected distance between any two points in $\boldsymbol{P}^e$: d_{eq}.
Output: Equidistant points in latent space: $\boldsymbol{P}^e$.
1: **Initialized:** Randomly initialized $\boldsymbol{P}^e$.
2: Construct matrix $\boldsymbol{V}$;
3: Calculate the Moore–Penrose inverse matrix $\boldsymbol{V}^+$ of $\boldsymbol{V}$;
4: **for** $i \leftarrow 0, 1, \ldots, 10$ **do**
5: $\quad \boldsymbol{X} = \boldsymbol{P}^e$;
6: $\quad$ Calculate $\boldsymbol{B}(\boldsymbol{X})$ using $\boldsymbol{X}$ and d_{eq};
7: $\quad \boldsymbol{P}^e = \boldsymbol{V}^+\boldsymbol{B}(\boldsymbol{X})\boldsymbol{X}$.
8: **end**
9: return $\boldsymbol{P}^e$

Algorithm 4: Training With Equal Distance Loss.

Input: All training data $\mathcal{D}^l \cup \mathcal{D}^u$, Feature extractor E.
1: **Initialized:** E after 10 epochs of warm-up training, assignment frequency $\boldsymbol{\beta} = 0$; $|\boldsymbol{\beta}| = m$.
2: Use K-means to obtain class prototypes $\boldsymbol{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m\}$;
3: Calculate the equidistant points $\boldsymbol{P}^e = \{\boldsymbol{p}_1^e, \ldots, \boldsymbol{p}_m^e\}$;
4: **for** each epoch e **do**
5: $\quad$ **for** each batch $\boldsymbol{X} \in \mathcal{D}^l \cup \mathcal{D}^u$ **do**
6: $\quad\quad \boldsymbol{Z} = E(\boldsymbol{X})$;
7: $\quad\quad$ Assign the nearest prototype from $\boldsymbol{P}$ for each $\boldsymbol{Z}$ to obtain $\boldsymbol{Z}_p$;
8: $\quad\quad$ Update E with equal distance loss $\ell_{edl}(\boldsymbol{Z}, \boldsymbol{Z}_p)$;
9: $\quad\quad$ Recompute $\boldsymbol{Z}$ with updated E: $\boldsymbol{Z} = E(\boldsymbol{X})$;
10: $\quad\quad$ Recompute prototype assign $\boldsymbol{Z}_p$ for each new $\boldsymbol{Z}$;
11: $\quad\quad$ **for** $\boldsymbol{z}_i$ in $\boldsymbol{Z}$ **do**
12: $\quad\quad\quad$ Retrieve assignment index $c_{\boldsymbol{z}_i}$ of $\boldsymbol{z}_i$ from $\boldsymbol{Z}_P$;
13: $\quad\quad\quad \boldsymbol{\beta}[c_{\boldsymbol{z}_i}] = \boldsymbol{\beta}[c_{\boldsymbol{z}_i}] + 1$;
14: $\quad\quad\quad \boldsymbol{p}_{c_{\boldsymbol{z}_i}} = (1 - \frac{1}{\boldsymbol{\beta}[c_{\boldsymbol{z}_i}]})\boldsymbol{p}_{c_{\boldsymbol{z}_i}} + \frac{1}{\boldsymbol{\beta}[c_{\boldsymbol{z}_i}]}(\boldsymbol{z}_i + \boldsymbol{p}^e_{c_{\boldsymbol{z}_i}})$;
15: $\quad\quad$ **end**
16: $\quad$ **end**
17: **end**
18: return $\boldsymbol{P}^e$

$\xi(\boldsymbol{P}^e, \boldsymbol{X})$ value. By solving this iteratively, $\boldsymbol{P}^e$ is obtained where the distance between any two points is equal.

Next, we use the obtained $\boldsymbol{P}^e$ to guide the clustering process to achieve a more distinctive distribution of classes. Since the initial separability of unlabeled classes is poor at the beginning of training, the initialized class prototypes $\boldsymbol{P}$ obtained using K-means may be inaccurate. Therefore, different from [21], we set the first ten epochs as the warm-up stage. After the warm-up training, all classes have a certain degree of separability. At this point, we introduce equidistant points and involve them in the training process in the form of equal distance loss, as shown in Algorithm 4.

In Line 8 of the Algorithm 4, the definition of equal distance loss is as follows:

$$\ell_{\text{edl}}(\boldsymbol{Z}, \boldsymbol{Z}_p) = \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} ||\boldsymbol{z}_i - \boldsymbol{z}_{p,i}||_2^2 \tag{18}$$

where N_{batch} is the number of samples in the current batch, and $\boldsymbol{z}_i$ and $\boldsymbol{z}_{p,i}$ are, respectively, the sample in the current batch and its corresponding closest class prototype. This method alternates between learning from pseudolabels derived from class prototypes and modifying the class prototypes themselves, thereby guiding each sample to move to the designated equidistant position in the latent space, ultimately enhancing the separability between all classes.

E. Multicrop Consistency Loss

As described in Section A, UNO [27] leverages a views pair strategy built on image augmentation and uses the swapped prediction mechanism through the Sinkhorn–Knopp algorithm [54] to obtain pseudolabels for augmented unlabeled images. This method implicitly compares two views of the same sample by correlating the network's output predictions. According to [57], comparing more views can effectively improve model performance. Hence, we increase the number of augmented views to N_v, employ Random-Resized-Crop, ColorJitter, and Random-Rotation for augmentation, and utilize the approach illustrated in Section III-A to filter out noisy pairs. Due to the proportion of the cropped view to the original image being random, a variety of views depicting different parts and sizes of the target will simultaneously appear. Therefore, a new multicrop swapped prediction loss for a sample is defined as follows:

$$\ell_{msp} = \sum_{i=1,j=1,i\neq j}^{N_v} \ell_{\text{ce}}(\widehat{\boldsymbol{x}}_{u,i}, y_j) \tag{19}$$

where $\widehat{\boldsymbol{x}}_{u,i}$ represents the augmented view of the sample $\boldsymbol{x}_u$. y_j denotes the unified label of the augmented view $\widehat{\boldsymbol{x}}_{u,j}$ (concatenating the zero vector and the corresponding pseudolabel obtained from the Sinkhorn–Knopp algorithm). Therefore, the training loss of UNO is changed to the following function:

$$\ell_{uno-msp} = \begin{cases} \frac{1}{N_v} \sum_{i=1}^{N_v} \ell_{\text{ce}}(\widehat{\boldsymbol{x}}_i, y), & \boldsymbol{x} \in \mathcal{D}^l \\ \frac{1}{C_{N_v}^2} \ell_{msp}, & \boldsymbol{x} \in \mathcal{D}^u \end{cases}. \tag{20}$$

In order to ensure consistency among the outputs of different views of the same sample, IIC [20] employs sKLD to impose consistency constraints on the predictions produced by the classification heads. However, this approach does not directly enforce constraints at the feature space level, which may not suffice to enhance the representation quality in NCD tasks. The modified swapped prediction process is shown in Fig. 3(c). Therefore, we propose the multicrop consistency loss for multicrop swapped prediction, which enforces consistency based on Euclidean distance in manifold, as shown below

$$\ell_{\text{mc}} = \frac{1}{C_{N_v}^2} \sum_{i=1,j=1,i\neq j}^{N_v} ||\widehat{\boldsymbol{z}}_i - \widehat{\boldsymbol{z}}_j||_2^2 \tag{21}$$

where $\widehat{\boldsymbol{z}}_i$ and $\widehat{\boldsymbol{z}}_j$ represent the output feature vectors of two different augmented views of the same sample. Compared with the intraclass loss used in IIC, this loss directly constrains the generation of high-quality representations within the manifold,

participates in the training process together with the intraclass loss and plays a complementary role.

F. Overall Training Process and Objective

The overall training process is summarized as follows: First, when we obtain a dataset of images $\mathcal{D}^u$ containing only unknown classes, we perform self-supervised pretraining using both the known class dataset $\mathcal{D}^l$ and $\mathcal{D}^u$, using (1) as the loss function. Second, we estimate the number of classes contained in $\mathcal{D}^u$ using the method presented in Section III-B. Third, based on the UNO training framework, we add output heads to the model, adjusting the output units of the heads according to the estimated number of unknown classes from the previous step. Fourth, we perform NCD training on the model using $\ell_{uno-msp}$, ℓ_{inter}, ℓ_{intra}, ℓ_{mc}, ℓ_{edl}, simultaneously leveraging both $\mathcal{D}^l$ and $\mathcal{D}^u$. During the initial warm-up phase, only the first four loss functions are used. Once the warm-up training is completed, we determine equidistant points in the feature space based on the Euclidean distances between class centers, introducing ℓ_{edl} to assist the subsequent training process.

In summary, the overall loss function of the NCD stage is expressed as follows:

$$\ell_{\text{ncd}} = \ell_{uno-msp} + \ell_{\text{inter}} + \ell_{\text{intra}} + \ell_{\text{mc}} + \ell_{\text{edl}}. \tag{22}$$

Through the above process, clustering of $\mathcal{D}^u$ can be achieved based on the estimated number of unknown classes.

IV. Experiments

In this section, we conducted extensive experiments on the Moving and Stationary Target Automatic Recognition (MSTAR) dataset to demonstrate the effectiveness of our method. Firstly, we provide a detailed introduction to the MSTAR dataset. Second, we compare our proposed method with current state-of-the-art methods to validate the NCD performance on SAR targets. Finally, we conduct ablation studies and discuss the impact of some hyperparameters and proposed techniques on the experimental results. The laptop used in our experiments has an Intel Core i9-13900HX CPU, an NVIDIA GeForce RTX 4060 Laptop GPU, and 16 GB of RAM on the Windows 11 system.

A. Experimental Setup

Datasets: In this article, we utilize the MSTAR dataset for our experiments. The MSTAR is a standardized database for SAR target recognition that encompasses ten classes of ground targets, including BMP2, BTR70, T72, BTR60, 2S1, BRDM2, D7, T62, ZIL, and ZSU. Within these classes, BMP2 and T72 are distinguished by three variants, namely BMP2-c9563, BMP2-c9566, BMP2-c21, T72-132, T72-812, and T72-s7. The radar operates in the X-band and employs spotlight mode imaging with an image resolution of 0.3× 0.3 m. Each image spans an elevation angle of approximately 3°, with data acquisition depression angles set at 15° and 17°, respectively. Similar to the reference [24], [30], we designate 17° as the training set and 15° as the testing set. Fig. 4 illustrates various target objects and their corresponding optical images. It can be seen that different types of SAR images exhibit high similarity and are more difficult to distinguish than optical images. The numbers of the ten-class targets are shown in Table I.

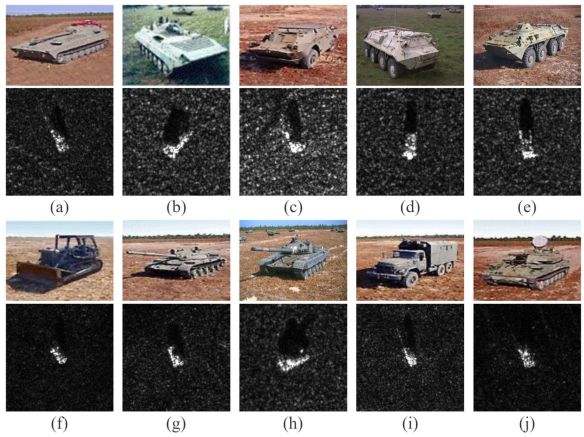


Fig. 4. SAR images and corresponding optical images of ten classes of targets in MSTAR dataset. (a) 2S1. (b) BMP2. (c) BRDM2. (d) BTR60. (e) BTR70. (f) D7. (g) T62. (h) T72. (i) ZIL131. (j) ZSU234.

TABLE I
Number of Ten-Class Targets

Target	Number (17°)	Number (15°)
2S1	299	274
BMP2	233	196
BRDM2	298	274
BTR60	256	195
BTR70	233	196
D7	299	274
T62	299	273
T72	232	196
ZIL131	299	274
ZSU234	299	274

Metrics: In the following experiments, we use accuracy as the metric to evaluate the classification of labeled classes (ACC_l) and average clustering accuracy as the metric to measure the clustering performance of unlabeled classes (ACC_u). The average clustering accuracy is presented as follows:

$$\text{ACC}_\text{u} = \max_{\text{perm}\in\text{P}} \frac{1}{\text{N}_\text{t}} \sum_{\text{i}=1}^{\text{N}_\text{t}} \mathbb{1}\{\text{y}_\text{i} = \text{perm}(\hat{\text{y}}_\text{i})\} \tag{23}$$

where y_i and $\widehat{y_i}$ are the ground-truth label and predicted label of an image $\boldsymbol{x}_i \in \mathcal{D}^u$, respectively. N_t represents the total number of test samples, and P represents the set of all permutations, which is obtained by the Hungarian algorithm [58]. As the network is trained using multiple heads, we calculate evaluation metrics separately for each head. Both the average accuracy and the best head accuracy are reported. The best head is determined as the one that demonstrates the lowest training loss and is chosen to compute the result. In addition, we select normalized mutual information (NMI) [59] and adjusted rand index (ARI) [60] as metrics, which are commonly used to evaluate the effectiveness of clustering.

TABLE II
MSTAR DATASET DIVISION UNDER DIFFERENT KNOWN CLASS NUMBER SCENARIOS

Settings	Dataset Division	
	Known Classes	Unknown Classes
Class 5/5	2S1, BMP2, ZIL131 ZSU234, BTR60	BRDM2, BTR70, D7 T62, T72
Class 7/3	2S1, BMP2, ZIL131, ZSU234 BTR60, BTR70, T62	BRDM2, D7 T72

Implementation Details: To maintain consistency with the settings in [24], [27], [30], we use a ResNet18 encoder as the feature extractor E. The labeled classification head h is an $l2$-normalized linear layer with C^l output units, while the unlabeled classification head g is composed of a projection head with 2048 hidden units and 256 output units, followed by a $l2$-normalized linear layer with C^u output units. The entire experimental process uses single-channel SAR images as input, and the input images are uniformly cropped to a size of 96×96. During the pretraining phase, we first conduct self-supervised training for 1000 epochs using both labeled and unlabeled data, with a weight decay of 10^{-6} and a learning rate $lr = batchsize/256 \times base\_lr$, where $batchsize$ is set to 256 by default and $base\_lr$ is a base learning rate set to 0.5. The coefficients λ, μ, and ν in the loss function are set to 25, 25, and 1, respectively. And the threshold in Algorithm 1 is set to $\tau_{\text{pro}} = 0.06$, $\tau_{\text{bin}} = 100$, and $\tau_{\text{ratio}} = 0.5$. After that, supervised training is carried out for 200 epochs using only labeled data. In the NCD phase, we conduct training for 200 epochs, with a ten-epoch linear warm-up and cosine annealing ($base\_lr = 0.01, min\_lr = 0.001$) and a weight decay of 1.5×10^{-4}. The $batchsize$ is set to 256, and the temperature τ of the softmax layer is set to 0.1. The hyperparameter α in the process of solving equidistant points is set to 1.5. During the testing phase, we selected the corresponding unknown classes from the testing set to form the target dataset, where each image is center-cropped to a uniform size of 96×96. The trained model is then used to cluster the target dataset, and the resulting ACC_{u}, NMI, ARI, and other metrics are used to evaluate the clustering performance of the trained model.

B. Comparison With the State-of-The-Art

We compare our method with four two-stage methods (KCL [22], MCL [23], DTC [15], and OSR-NCD [24]) and four single-stage methods (AutoNovel [26], UNO [27], IIC [20], and CNT [30]), where DTC has three different forms. In this experiment, we set the overclustering factor to 3, the multihead number to 5, and the multicrop number to 4. The coefficients of $\ell_{uno-msp}$, ℓ_{inter}, ℓ_{intra}, ℓ_{mc}, and ℓ_{edl} are set to 1, 0.05, 0.01, 1, and 0.05, respectively. The settings of other methods remain consistent with their corresponding original codes. We set two experimental scenarios: 1) the number of labeled classes is equal to the number of unlabeled classes (five labeled classes / five unlabeled classes); and 2) the number of labeled classes is greater than the number of unlabeled classes (seven labeled classes / three unlabeled classes). The detailed settings are as shown in Table II.

Considering the scenario without the number of unlabeled classes in practical applications, we conduct experiments on estimating the number of unlabeled classes following the method described in Section III-B. The division of $\mathcal{D}^l$ and the estimation results are shown in Table III.

It can be seen that the number of unlabeled classes can be estimated without error, which indicates that the method can provide accurate estimation and serve the next steps.

To facilitate a more convenient and fair comparison, we uniformly assume that the number of unknown classes can be obtained for all comparative methods. The experimental results are shown in Table IV. It can be seen that our method has better results than the previous optimal baseline in both scenarios. Specifically, we notice that all two-stage methods face challenges in achieving satisfactory clustering results. This is primarily attributed to their reliance on transferring semantic knowledge of labeled data from the previous stage, where the imbalanced training in the transfer process often leads to collapse. AutoNovel [26] trains on both labeled and unlabeled data simultaneously. However, the results in both scenarios are worse than those of other single-stage methods. This is because the binary pseudolabels judged to be of the same class need to satisfy the same top-5 unit position index of the feature vectors, but very few extracted feature vectors from the same class SAR images meet this criterion. Therefore, the effective training times in each epoch are very few, which leads to the clustering effect mainly relying on the auxiliary training loss function.

In comparison to other methods, UNO-based methods (UNO, IIC, CNT, and our method) have achieved remarkable effectiveness, attributed to providing a unified and balanced training approach for both types of data. These methods exhibit notable enhancements in clustering unknown classes when the ratio of C^l to C^u increases. This improvement is due to the reduced number of unknown classes, leading to a lower risk of confusion between similar classes. In our method, the incorporation of equal distance loss allows the model to implicitly guide relationships between unlabeled classes based on the relationships between labeled classes. With the presence of more labeled classes contributing to increased interclass relationship information, the discriminability among unlabeled classes is further magnified. We compare the proposed method with the current state-of-the-art CNT and basic framework UNO (CNT mainly focuses on average clustering accuracy, without providing other clustering metrics). When $C^l = C^u = 5$, our method improves the average clustering accuracy by 6.06% compared to UNO and by 0.66% compared to CNT. When $C^l = 7$ and $C^u = 3$, our method enhances the average clustering accuracy by 6.39% compared to UNO and by 1.67% compared to CNT.

To visually demonstrate the advantages of our method, we use the t-SNE plot for comparison with the two baselines (UNO and IIC), as shown in Fig. 5. Each color represents an unlabeled class. As shown in the figure, when $C^l = C^u = 5$, the separation between unlabeled classes in our method is more distinct, indicating that the resulting representations have higher distinguishability. Furthermore, our method also achieves state-of-the-art results in both NMI and ARI, demonstrating the superiority of our approach in SAR NCD tasks.

TABLE III
UNKNOWN CLASSES NUMBER ESTIMATION RESULTS

Settings	Division of D^l			Metrics		
	$\mathcal{D}_T^l$	$\mathcal{D}_{probe,a}$	$\mathcal{D}_{probe,v}$	GT	Ours	Error
Class 5/5	2S1, BMP2, ZIL131	ZSU234	BTR60	5.0	5.0	0.0
Class 7/3	2S1, BMP2, ZIL131, ZSU234, BTR60	T62	BTR70	3.0	3.0	0.0

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSTAR FOR NOVEL CATEGORY DISCOVERY

Methods	Class 5/5			Class 7/3		
	ACC_u	NMI	ARI	ACC_u	NMI	ARI
KCL [22]	0.5399	0.2983	0.2462	0.6586	0.3826	0.3957
MCL [23]	0.4311	0.1789	0.1515	0.6183	0.2828	0.2757
DTC-Base [15]	0.3916	0.2451	0.1617	0.6022	0.2862	0.2913
DTC-TE [15]	0.3924	0.2247	0.1305	0.5847	0.2849	0.1445
DTC-π [15]	0.3990	0.3341	0.2309	0.6102	0.3170	0.3370
OSR-NCD [24]	0.4386	0.2490	0.1584	0.6734	0.4145	0.3029
AutoNovel [26]	0.5033	0.2241	0.2058	0.6650	0.4145	0.3361
UNO [27]	0.8920	0.7832	0.7597	0.9299	0.8028	0.8219
IIC [20]	0.8970	0.8108	0.7867	0.9581	0.8779	0.8710
CNT [30]	0.9510	\	\	0.9771	\	\
Our method	**0.9576**	**0.9062**	**0.9080**	**0.9938**	**0.9678**	**0.9810**

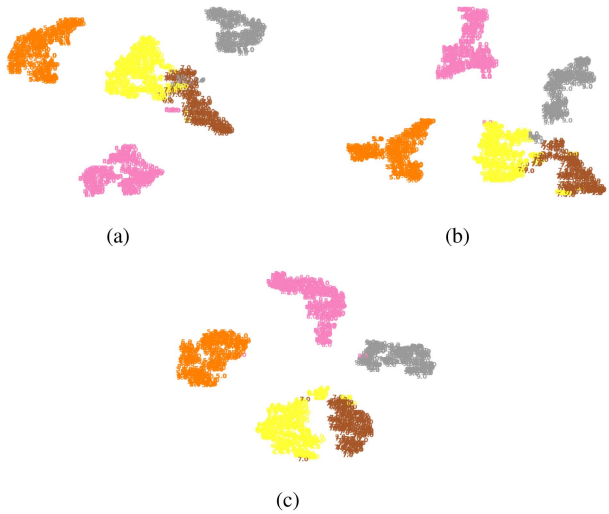


Fig. 5. Results of t-SNE visualization on the unlabeled dataset. (a) UNO. (b) IIC. (c) Our method.

TABLE V
COMPARISON OF FLOPS AND PARAMS FOR DIFFERENT METHODS

Methods	FLOPs	Params
UNO [27]	333.15 M	23.83 M
IIC [20]	333.15 M	23,83 M
Our method	336.32 M	26,99 M

Finally, we calculate the floating point operations (FLOPs) and parameters (Params) of the proposed method and compared them with two baseline methods (UNO and IIC), as shown in Table V. The FLOPs are calculated using a single 96×96 image as input. It can be seen that UNO and IIC have the same indicators because IIC adopts the basic framework of UNO and has the same number of Overclustering heads and unlabeled heads. Our method has higher complexity due to the increased number of heads.

C. Ablation Study

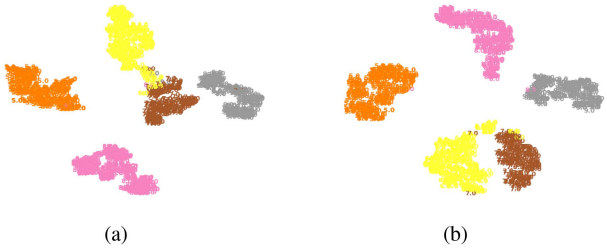
In the above experiments, we compare our method with other state-of-the-art methods. In this part, we conduct ablation studies to demonstrate the effectiveness of the proposed improved techniques and the influence of the choice of hyperparameters. Following the typical division of NCD tasks [15], [20], [26], [27], we select five classes as labeled classes (known classes) and the remaining five classes as unlabeled classes (unknown classes), i.e., $C^l = C^u = 5$.

Validity of Improved Techniques: To validate the contributions of each proposed technique to the overall NCD task, we conducted a study by systematically removing different components. The components involved in the study are interclass and intraclass constraints (IIC), self-supervised pretraining (SSP), EDL, noisy pair filtering (NPF), and multicrop consistency loss (MCCL). In this experiment, we set the multihead number to 5, the overclustering factor to 3, and the multicrop number to 4, with all other settings consistent with Section B. The results of the ablation study are shown in Table VI.

The first and second lines of the experimental results correspond to the work of UNO [27] and IIC [20], respectively, providing us with baselines for comparison. It can be observed that incorporating self-supervised pretraining into the work of IIC results in an overall improvement in clustering performance, indicating that the pretraining method based on VICReg effectively enhances the model's understanding of the semantic information in the SAR image. Furthermore, the introduction of EDL during the NCD phase further improves clustering accuracy. On this basis, implementation of the NPF process filters out noisy view pairs from both the self-supervised

TABLE VI
STUDY OF THE VALIDITY OF EACH IMPROVED TECHNIQUE

Enhancement methods					Metrics			
IIC	SSP	EDL	NPF	MCCL	ACC_u	NMI	ARI	ACC_l
✗	✗	✗	✗	✗	0.8920	0.7832	0.7597	0.9494
✓	✗	✗	✗	✗	0.8970	0.8108	0.7867	0.9299
✓	✓	✗	✗	✗	0.9130	0.8445	0.8296	0.9833
✓	✓	✓	✗	✗	0.9264	0.8688	0.8474	0.8996
✓	✓	✓	✓	✗	0.9516	0.8813	0.8927	0.9848
✓	✓	✗	✓	✓	0.9457	0.8793	0.8790	**0.9935**
✓	✓	✓	✓	✓	**0.9576**	**0.9062**	**0.9080**	0.9906



(a) (b)

Fig. 6. Experiments with and without equal distance loss, with visualization effects on unlabeled data. (a) Our method without equal distance loss. (b) Our method including equal distance loss.

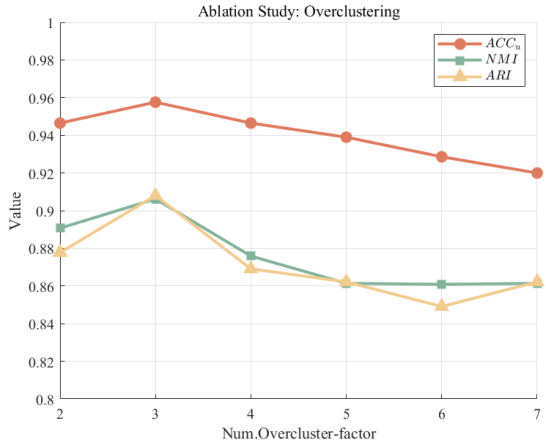


Fig. 7. Impact of the change in overclustering factor value on clustering performance.

pretraining and NCD phases, thereby enhancing the effectiveness of pretraining and the efficiency of knowledge transfer, resulting in superior performance. In addition, the incorporation of MCCL constrains the consistency between multiple views in the latent space, consequently boosting accuracy by 0.6% and achieving optimal model performance. Comparative analysis of the final two lines of Table VI reveals that the introduction of EDL yields a noteworthy average clustering accuracy increase of 1.19%, along with enhancements of 2.69% for NMI and 2.9% for ARI. To present the performance of EDL more intuitively, we use t-SNE to visualize the results obtained before and after the introduction of EDL, as shown in Fig. 6, where each color represents an unlabeled class.

From the visual representations, it is evident that in the latent space, distinguishing between the brown and yellow classes is more challenging compared to other classes. This difficulty arises from the inherent similarity between these two classes and the influence of labeled data on guiding clustering. With the introduction of EDL, the distances between the two classes in the latent space expand, enabling clear differentiation. Moreover, the distance between the gray and brown classes also increases, signifying a notable decrease in the risk of misclassification. In summary, the utilization of EDL visibly enhances the distinctiveness between unlabeled classes, ultimately improving the effectiveness of clustering.

Overclustering Factor: The implementation of overclustering has been proven to greatly boost the quality of representation in UNO [27] and IIC [20]. To investigate the impact of the choice of overclustering factor on NCD performance, we conduct a comparison of experimental results with varying overclustering factor values while keeping the other hyperparameters constant. The experimental results are shown in Fig. 7.

We observe that as the value increases, all three metrics show a trend of initially rising and then declining. This is because the introduction of overclustering provides finer partitioning, which helps the model perceive the fine-grained differences between images. However, excessively fine partitioning forces interclass distances to increase, leading to a decrease in clustering performance. Therefore, selecting the sweet spot of overclustering in training can effectively assist clustering tasks.

Multihead Number: All methods based on UNO employ multihead techniques [20], [27], [30], but there has been no research conducted on the optimal number of multiheads. Therefore, we conduct experiments by varying the number of unlabeled heads (the number of overclustering heads increases synchronously), and the results are presented in Fig. 8. We notice that as the number of unlabeled heads increases, the clustering performance first increases, as more classification heads mitigate the case of converging to suboptimal clustering configurations. As the number of unlabeled heads continues to increase, the benefits of clustering will reach a peak and then gradually decrease, converging towards the results obtained when multihead was not utilized.

Multicrop Number: In this article, we introduce the multicrop technique during the NCD phase. As all generated unlabeled-class augmented views participate in swapped prediction, the number of multicrops has a significant impact on intraclass loss, interclass loss, and multicrop consistency loss. Therefore, we

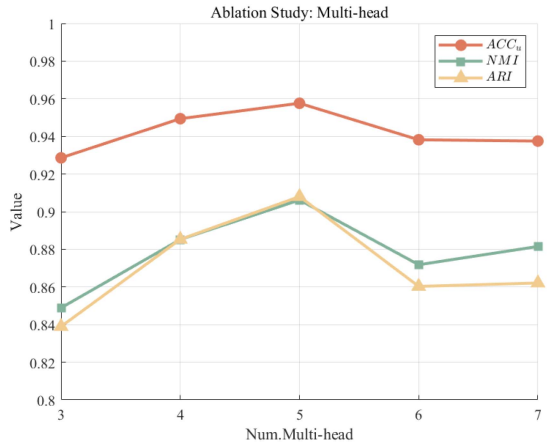


Fig. 8. Impact of the change in multihead number on clustering performance.

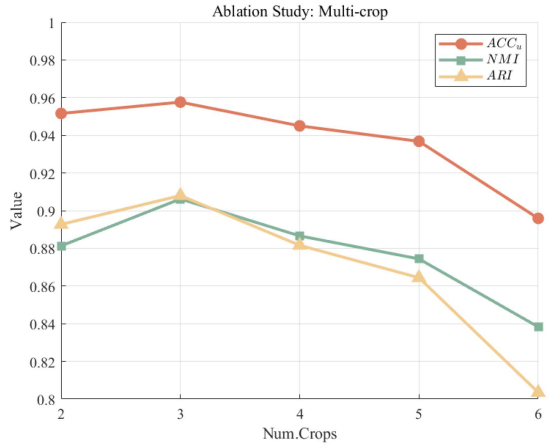


Fig. 9. Impact of the change in multicrop number on clustering performance.

need to balance the subtasks generated by multiple loss functions and find the optimal number of cropping views that are most beneficial for the overall NCD task. We vary the number of multicrops, and the experimental results are shown in Fig. 9.

It can be observed that with the increase in the number of multicrops, the clustering performance shows a trend of initially rising and then rapidly declining. The improvement in performance is due to the implicit contrast of more views, which enhances the model's understanding of the target and perception of subtle differences. The decline is attributed to the uncertainty introduced by too many views, disrupting the stable convergence of training. Therefore, we need to find the balance point between these two effects to achieve optimal clustering performance.

V. Discussion

Compared to baseline methods and state-of-the-art methods, our method demonstrates superior clustering performance across multiple metrics. We believe this is primarily attributed to three key factors: First, addressing the collapse issue when applying the NCD method to SAR image datasets. Second, better constraining the output consistency of multiple augmented views generated from the same sample. Third, effectively transferring knowledge of known class relationships to improve the distinguishability between unknown classes. Although our method achieves better results, it still has limitations.

We have confined our research scope to same-domain SAR images, where labeled data and unlabeled data share similar deep semantic information. However, in real-world applications, unlabeled data may belong to a different domain than the labeled data, such as ground targets versus ship targets. In the future, we will focus on the study of cross-domain novel category discovery to enhance the generalization of the proposed method. Furthermore, we will consider the complex scenario where the target unlabeled dataset contains both known and unknown classes.

VI. Conclusion

In this article, we propose a novel SAR image NCD method based on the UNO framework, which is comprehensively improved in both the pretraining and category discovery phases, achieving state-of-the-art clustering performance in two typical class division settings. In the pretraining phase, we propose a simple and effective noisy pair filtering algorithm to mitigate the impact of noisy pairs with significant semantic differences on the model. In the novel category discovery phase, we introduce multicrop consistency loss and equal distance loss to impose constraints on the representation from both intraclass view relationships and interclass distinctiveness in the latent space, further enhancing the clustering effectiveness. The experimental results on the MSTAR dataset demonstrate the effectiveness of the proposed method. Furthermore, through ablation studies on proposed techniques and key hyperparameters, we have illustrated the impact of the aforementioned improvements on the entire NCD task, as well as the effects of different hyperparameter values on the ultimate results.

References

[1] X. Qin, B. Deng, H. Wang, Y. Zeng, X. Chen, and H. Xiao, "An equalized ADMM-based high-resolution autofocusing imaging algorithm for THz-SAR ground moving targets," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8450–8460, 2024.

[2] M. Manzoni, S. Tebaldini, A. V. Monti-Guarnieri, and C. M. Prati, "Multipath in automotive MIMO SAR imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5202612.

[3] G. Fracastoro, E. Magli, G. Poggi, G. Scarpa, D. Valsesia, and L. Verdoliva, "Deep learning methods for synthetic aperture radar image despeckling: An overview of trends and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 29–51, Jun. 2021.

[4] C. Li, L. Du, Y. Li, and J. Song, "A novel SAR target recognition method combining electromagnetic scattering information and GCN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4508705.

[5] Y. Li, L. Du, and D. Wei, "Multiscale CNN based on component analysis for SAR ATR," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5211212.

[6] C. Wang et al., "SAR ATR under limited training data via MobileNetV3," in *2023 IEEE Radar Conf.*, 2023, pp. 1–6.

[7] H. Lin, H. Wang, F. Xu, and Y.-Q. Jin, "Target recognition for SAR images enhanced by polarimetric information," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5204516.

[8] F. Gao et al., "SAR target incremental recognition based on features with strong separability," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5202813.

[9] F. Ma, F. Zhang, D. Xiang, Q. Yin, and Y. Zhou, "Fast task-specific region merging for SAR image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5222316.

[10] F. Zhang, X. Sun, F. Ma, and Q. Yin, "Superpixelwise likelihood ratio test statistic for polsar data and its application to built-up area extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 233–248, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271624000546

[11] F. Gao, X. Han, J. Wang, J. Sun, A. Hussain, and H. Zhou, "SAR ship instance segmentation with dynamic key points information enhancement," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 11365–11385, 2024.

[12] F. Zhang, Y. Yu, F. Ma, and Y. Zhou, "A physically realizable adversarial attack method against SAR target recognition model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 11943–11957, 2024.

[13] Z. LI, J. Otholt, B. Dai, D. Hu, C. Meinel, and H. Yang, "A closer look at novel class discovery from the labeled set," in *NeurIPS 2022 Workshop Distrib. Shifts: Connecting Methods Appl.*, 2022, pp 1–20. [Online]. Available: https://openreview.net/forum?id=8-TFK-fmQSq

[14] Y. Sun, Z. Shi, Y. Liang, and Y. Li, "When and how does known class help discover unknown ones? provable understandings through spectral analysis," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 33014–33043. [Online]. Available: https://openreview.net/forum?id=JHodnaW5WZ

[15] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *2019 IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8400–8408.

[16] D. Berthelot et al., "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HklkeR4KPB

[17] D. Berthelot et al., "Mixmatch A. holistic approach to semi-supervised learning," in *Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds.. Curran Associates, Inc., 2019, vol. 32, pp. 5049–5059.

[18] X. Wang, L. Lian, and S. X. Yu, "Unsupervised selective labeling for more effective semi-supervised learning," in *Computer Vis.–ECCV 2022: 17th Eur. Conf.*, Tel Aviv, Israel, 2022, pp. 427–445, doi: 10.1007/978-3-031-20056-4_25.

[19] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu, "Simmatch: Semi-supervised learning with similarity matching," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14451–14461.

[20] W. Li, Z. Fan, J. Huo, and Y. Gao, "Modeling inter-class and intra-class constraints in novel class discovery," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 3449–3458.

[21] K. J. Joseph et al., "Spacing loss for discovering novel categories," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 3760–3765.

[22] Y.-C. Hsu, Z. Lv, and Z. Kira, "Learning to cluster in order to transfer across domains and tasks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 4926–4945. [Online]. Available: https://openreview.net/forum?id=ByRWCqvT-

[23] Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, and Z. Kira, "Multi-class classification without multi-class labels," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 5183–5198. [Online]. Available: https://openreview.net/forum?id=SJzR2iRcK7

[24] M. Chen, J.-Y. Xia, T. Liu, L. Liu, and Y. Liu, "Open set recognition and category discovery framework for SAR target classification based on K-contrast loss and deep clustering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3489–3501, 2024.

[25] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2022.

[26] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, "Automatically discovering and learning new visual categories with ranking statistics," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 10052–10064.

[27] E. Fini, E. Sangineto, S. Lathuilière, Z. Zhong, M. Nabi, and E. Ricci, "A unified objective for novel class discovery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9264–9272.

[28] H. Chi et al., "Meta discovery: Learning to discover novel classes given very limited data," in *Proc. Int. Conf. Learn. Representations*, 2022.

[29] Z. Zhong, E. Fini, S. Roy, Z. Luo, E. Ricci, and N. Sebe, "Neighborhood contrastive learning for novel class discovery," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10862–10870.

[30] H. Huang, F. Gao, J. Sun, J. Wang, A. Hussain, and H. Zhou, "Novel category discovery without forgetting for automatic target recognition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4408–4420, 2024.

[31] Z. Zang et al., "Boosting novel category discovery over domains with soft contrastive learning and all in one classifier," in *2023 IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11824–11833.

[32] J. Zheng, W. Li, J. Hong, L. Petersson, and N. Barnes, "Towards open-set object detection and discovery," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 3960–3969.

[33] S. Roy, M. Liu, Z. Zhong, N. Sebe, and E. Ricci, "Class-incremental novel class discovery," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp 317–333.

[34] J. Chen, Z. Zhao, S. Zheng, L. Zhang, K. Qiu, and X. Yang, "PCSP: A novel class discovery algorithm for radio signal classification," *IEEE Trans. Cogn. Commun. Netw.*, vol. 10, no. 4, pp. 1190–1203, Aug. 2024.

[35] Y. Zhao, Z. Zhong, N. Sebe, and G. H. Lee, "Novel class discovery in semantic segmentation," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4330–4339.

[36] Y. Liu and T. Tuytelaars, "Residual tuning: Toward novel category discovery without labels," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7271–7285, Oct. 2023.

[37] Z. LI, J. Otholt, B. Dai, D. Hu, C. Meinel, and H. Yang, "Supervised knowledge may hurt novel class discovery performance," *Trans. Mach. Learn. Res.*, 2023. [Online]. Available: https://openreview.net/forum?id=oqOBTo5uWD

[38] Y. Sun, Z. Shi, and Y. Li, "A graph-theoretic framework for understanding open-world semi-supervised learning," in *Proc. Int. Adv. Conf. Neural Inf. Process. Syst.*, 2023, pp. 23934–23967. [Online]. Available: https://openreview.net/forum?id=ZITOHWeAy7

[39] Z. Zhong, L. Zhu, Z. Luo, S. Li, Y. Yang, and N. Sebe, "Openmix: Reviving known knowledge for discovering novel visual categories in an open world," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9457–9465.

[40] L. Dai, W. Guo, Z. Zhang, and W. Yu, "Discovering novel categories in sar images in open set conditions," in *IGARSS 2022–2022 IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 1932–1935.

[41] H. Bao, L. Dong, S. Piao, and F. Wei, "BEit: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=p-BhZSz59o4

[42] V. Sehwag, M. Chiang, and P. Mittal, "SSD: A unified framework for self-supervised outlier detection," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=v5gjXpmR8J

[43] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *2015 IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1422–1430.

[44] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Computer Vis.–ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 649–666.

[45] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.

[46] Z. Xie et al., "Simmim: A simple framework for masked image modeling," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9643–9653.

[47] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988.

[48] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 2245–2260. [Online]. Available: https://openreview.net/forum?id=S1v4N2l0-

[49] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[50] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[51] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*. Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 22243–22255.

[52] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[53] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 9912–9924.

[54] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Adv. Neural Inf. Process. Syst.*, C. L. Burges, M. Bottou Welling, Z. Ghahramani, and K. Weinberger, eds. Curran Associates, Inc., 2013, vol. 26, pp. 2292–2300. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf
[55] J. Macqueen, "Some methods for classification and analysis of multivariate observations," *Proc. Symp. Math. Statist. Probability, 5th*, vol. 1, pp. 281–297, 1967.
[56] J. de Leeuw, "Applications of convex analysis to multidimensional scaling," in Department of Statistics, UCLA, 2005.
[57] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6706–6716.
[58] H. W. Kuhn, "The hungarian method for the assignment problem," *Nav. Res. Logistics*, vol. 52, no. 1-2, pp. 7–21, 2010.
[59] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Mar. 2003, doi: 10.1162/153244303321897735.
[60] S. Wang, J. Yang, J. Yao, Y. Bai, and W. Zhu, "An overview of advanced deep graph node clustering," *IEEE Trans. Computat. Social Syst.*, vol. 11, no. 1, pp. 1302–1314, Feb. 2024.

Mingyao Chen received the B.E. degree in information and communication engineering from the College of Information and Communication, National University of Defense Technology, Changsha, China, in 2022, where he is currently working toward the master's degree in information and communication engineering with the College of Electronic Science and Technology.

His research interests include radar automatic target recognition, pattern recognition, and open set recognition.

Tianpeng Liu received the B.E., M.E., and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2008, 2011, and 2016 respectively.

He is currently an Associate Professor with the College of Electronic Science and Technology, Changsha. He has authored or coauthored numerous papers in respected journals, including IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS and International Conference on Signal Processing. His research interests include radar signal processing, electronic countermeasure, and cross-eye jamming.

Li Liu (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2012.

She is currently a Full Professor with NUDT. She has held visiting appointments with the University of Waterloo, Waterloo, ON, Canada, The Chinese University of Hong Kong, Hong Kong, and the University of Oulu, Oulu, Finland, respectively. Her research interests include computer vision, pattern recognition, and machine learning.

Dr. Liu was a Co-Chair of many International Workshops along with major venues, such as CVPR and ICCV. She was the Leading Guest Editor of the special issues for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and *International Journal of Computer Vision*. She was also the Area Chair of several respected international conferences. She is currently an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Pattern Recognition*. Her papers currently have more than 10 000 citations, according to Google Scholar.