SSL-MBC: Self-Supervised Learning with Multi-Branch Consistency for Few-Shot PolSAR Image Classification

Wenmei Li, *Member, IEEE,*, Hao Xia, Bin Xi, Yu Wang, *Member, IEEE,* Jing Lu, Yuhong He

***Abstract*—Deep learning (DL) methods have recently made substantial advances in polarimetric synthetic aperture radar (PolSAR) image classification. However, supervised training relying on massive labeled samples is one of its major limitations, especially for PolSAR images that are hard to manually annotate. Self-supervised learning (SSL) is an effective solution for insufficient labeled samples by mining supervised information from the data itself. Nevertheless, fully utilizing SSL in PolSAR classification tasks is still a great challenge due to the data complexity. Based on the above issues, we propose a SSL model with multi-branch consistency (SSL-MBC) for few-shot PolSAR image classification. Specifically, the data augmentation technique used in the pretext task involves a combination of various spatial transformations and channel transformations achieved through scattering feature extraction. Additionally, the distinct scattering features of PolSAR data are considered as its unique multimodal representations. It is observed that the different modal representations of the same instance exhibit similarity in the encoding space, with the hidden features of more modals being more prominent. Therefore, a multi-branch contrastive SSL framework, without negative samples, is employed to efficiently achieve representation learning. The resulting abstract features are then fine-tuned to ensure generalization in downstream tasks, thereby enabling few-shot classification. Experimental results yielded from selected PolSAR datasets convincingly indicate that our method exhibits superior performance compared to other existing methodologies. The exhaustive ablation study shows that the model performance degrades when either the data augmentation or any branch is masked, and the classification result does not rely on the label amount.**

***Index Terms*—Polarimetric synthetic aperture radar (PolSAR), self-supervised learning (SSL), multimodal representation, image classification, few-shot.**

This work was funded by the National Natural Science Foundation of China under Grant 42071414, the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China (Grant No. KLSMNR-K202201), the Open Fund of State Key Laboratory of Remote Sensing Science (Grant No. OFSLRSS202202), the Key Laboratory of Land Satellite Remote Sensing Application, Ministry of Natural Resources of the People's Republic of China (Grant No. 202305), the China Postdoctoral Science Foundation under Grant 2019M661896, and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (Grant No. KYCX24_1219) *(Corresponding authors: Yuhong He.)*

Manuscript received April 19, 2021; revised August 16, 2021.

Wenmei Li, Hao Xia, Bin Xi are with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. Wenmei Li is also with the Health Big Data Analysis and Location Services Engineering Laboratory of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: liwm@njupt.edu.cn, 1023172909@njupt.edu.cn, 1023172908@njupt.edu.cn).

Yu Wang is with the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: yuwang@njupt.edu.cn).

Jing Lu is with Land Satellite Remote Sensing Application Center, MNR, No. 1 Baisheng Village, Haidian District, Beijing 100048,China (e-mail: luj@lasac.cn).

Yuhong He is with the Department of Geography, Geomatics and Environment, University of Toronto, 3359 Mississauga Road, Mississauga, ON L5L 1C6, Canada (e-mail: yuhong.he@utoronto.ca).

I. Introduction

POLARIMETRIC synthetic aperture radar (PolSAR), an active microwave imaging sensor, serves as the primary means of earth observation, and extracts fully polarimetric information about targets under all-weather and all-time conditions. Compared to conventional SAR, PolSAR incorporates polarimetric decompositions for a more complete portrayal of the target polarimetric scattering mechanism. PolSAR image classification predicts the true category corresponding to each pixel based on the information it contains, which is one of the major components of PolSAR image interpretation and has been applied widely in target detection [1], ecological protection [2], urban planning [3], and more.

Researchers have proposed a range of effective methods for traditional PolSAR image classification. These approaches mainly focus on two key aspects: scattering feature extraction and classifier enhancement. Scattering feature extraction is performed by coherent and incoherent decomposition. The former is based on the polarimetric scattering matrix [4], [5], while the latter uses polarization coherence matrix and covariance matrix as the basis, containing both eigenvalue-based and scattering model-based decomposition [6]–[8]. Scattering feature extraction provides more options for target parameters, however, selecting the most appropriate feature for a specific task depends on expert knowledge and experience. In terms of classifier improvement, machine learning algorithms such as support vector machine (SVM) [9], random forest [10], and decision tree [11] have been employed for PolSAR image classification. These methods solve complex nonlinear problems through autonomous learning, but the reliance on manual feature engineering is the main limitation.

The recent research has proven the effectiveness of deep learning (DL) technology in image classification tasks, with its ability to autonomously extract high-level abstract features without human intervention [12]. Classic DL methods primarily encompass deep belief network (DBN) [13], sparse auto encoder (SAE) [14], and convolutional neural network (CNN) [15], with the latter being widely used due to its efficient parameter sharing and modular architecture. The exceptional

performance of DL techniques has led to the applications of numerous DL-based classification methods to PolSAR images, such as real-valued CNN (RV-CNN) [16], complex-valued CNN (CV-CNN) [17], and vision transformer (VIT) [18]. While these methods produce satisfactory classification accuracy, they encounter several challenges:

1) These models need a large amount of labeled data for training, which is costly and often complicated by the difficulty of ensuring label quality.
2) The performance of DL models in classification tasks significantly declines when there is an insufficient number of labeled training samples.
3) Models trained on artificially annotated labels are prone to over-fitting. This phenomenon is characterized by the model converging to a solution that is only applicable to a specific task, resulting in poor generalization to new tasks.

To address these challenges, considerable studies have been conducted on few-shot scenarios to mitigate the dependency on extensive labels [19]. Self-supervised learning (SSL) emerges as a potential approach in this context, leveraging pretext tasks to mine supervisory signals from large volumes of unlabeled data, and train the network with this information to develop meaningful representations for subsequent classification tasks [20]. SSL not only overcomes the limitation of label quantity but also mines latent data correlations, demonstrating considerable potential in few-shot classification. In the context of PolSAR, some reasonable SSL-based few-shot PolSAR image classification frameworks have been designed [21], [22]. These frameworks complete the model training by reducing the distance between positive sample pairs and expanding the difference between negative pairs. However, when there are insufficient negative samples or the differences among samples are not significant, the generalization capability of the model will be undermined. And due to the need to consider negative samples, the computational resources required for this type of framework are also larger. Though a SSL framework without negative samples was proposed by [23], its pretext task only involved simple cropping and flipping of the original patches and did not consider the inherent characteristics of PolSAR data.

Polarimetric features are a series of attributes with clear physical significance extracted from rich scattering through different target decomposition methods, which can be considered unique multimodal representations for PolSAR data. Since the scattering theory for each pixel is unique, the multimodal representations for the same pixel are similar. Although chromatic distortion, a highly effective data augmentation method in optical images, has no physical meaning for PolSAR data. We consider whether it is possible to extract more important shared information from the multimodal representations of PolSAR data. It is worth noting that [22] have pointed out that mutual information between different feature representations can provide good prior knowledge for the model, but it only involves one pair of representations at each run and needs negative samples to avoid collapse. Inspired by multi-view [24], we believe that maximizing mutual information across more representation modes may lead to improved outcomes. The challenge lies in designing an efficient SSL framework to extract high-quality prior knowledge from multimodal representations of PolSAR data.

Based on the above facts, this study aims to design a SSL model with multi-branch consistency (SSL-MBC) that utilizes multimodal representations to solve the poor accuracy in PolSAR image classification with insufficiently labeled samples. Model training does not depend on artificial annotation labels in the designed pretext task. The shared information extracted from the multimodal representation of PolSAR data provides prior knowledge for the model. A multi-branch framework is proposed for the pretext task, and the entire process does not require negative samples. After that, the trained feature extraction model is transferred to the downstream classification task to achieve impressive performance. The main contributions of this paper are summarized as follows:

1) We innovatively use multimodal representations of PolSAR data to improve SSL performance, based on the fact that invariant features present in more modalities better represent the target's essential properties.
2) A SSL framework combined with multi-branch consistency (SSL-MBC) is proposed. This framework learns consistency from different modal representations of unlabeled PolSAR data through multiple branches without negative samples.
3) A rich pretext task is designed to provide training motivation to our framework, which contains diverse geometric transformation augmentations, as well as cross-modal variance in the coding space.

The subsequent structure of this article is as follows: Section II reviews the relevant work. Section III offers a comprehensive account of the proposed method. Section IV shows the experimental outcomes on diverse datasets and the corresponding analyses. Finally, the conclusions are offered in Section V.

II. Related Work

A. PolSAR Data Processing

When considering the combined horizontal and vertical polarization bases, the scattering properties of targets captured by PolSAR can be effectively depicted using the complex two-dimensional scattering matrix S:

$$S = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \tag{1}$$

where H and V are the horizontal and vertical modes, respectively. When the reciprocity assumption is satisfied, for single station $S_{HV} = S_{VH}$.

The scattering matrix S is vectored as K by the Pauli decomposition [25] as:

$$K = \frac{1}{\sqrt{2}}[S_{HH} + S_{VV}, S_{HH} - S_{VV}, 2S_{HV}]^T \tag{2}$$

where the superscript T represents the transpose operation. On this basis, the coherence matrix $\mathbf{T}$ after multilook is obtained as:

$$\mathbf{T} = \frac{1}{L}\sum_{i=1}^{L} K_i K_i^H = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} \tag{3}$$

where L is the number of looks and the superscript H is the conjugate transposition. Based on Eq. (3), the matrix T is characterized as a Hermitian matrix, exhibiting real values along its diagonal and complex values in its off-diagonal elements.

Scattered feature extraction provides more detailed target information, whereas incoherent decomposition methods based on coherence and covariance matrices are more suitable for feature analysis in large scenes [26]. Cloude-Pottier decomposition [6] and Freeman-Durden decomposition [7] are classical methods based on eigenvalue decomposition and scattering models, respectively. Both of them are employed to extract statistical and physical features, with their applications widely used in land cover classification tasks.

In Cloude-Pottier decomposition, the coherence matrix T after eigenvalue decomposition is:

$$\mathbf{T} = U\Lambda U^H \tag{4}$$

where Λ is a diagonal matrix that includes three eigenvalues λ_1, λ_2, and λ_3 of the matrix $\mathbf{T}$. U consists of the column vectors associated with these eigenvalues. The entropy H, anisotropy A, and mean scattering angle $\overline{\alpha}$ are provided as follows:

$$H = \sum_{i=1}^{3} -P_i \log_3 P_i \ , \ P_i = \frac{\lambda_i}{\sum_{j=1}^{3}\lambda_j} \tag{5}$$

$$A = \frac{\lambda_2 - \lambda_3}{\lambda_2 + \lambda_3} \tag{6}$$

$$\overline{\alpha} = \sum_{i=1}^{3} P_i \alpha_i \tag{7}$$

where α_i is the scattering angle.

In Freeman-Durden decomposition, the formula for the coherence matrix T is as follows:

$$\mathbf{T} = f_v\langle T_{vol}\rangle + f_d T_{dbl} + f_s T_{odd} \tag{8}$$

where T_{vol}, T_{dbl}, and T_{odd} represents the volume, double, and odd scatter model, respectively. f_v, f_d, and f_s corresponds to the respective scattering components. $\langle\cdot\rangle$ is ensemble average operation. Specifically, T_{vol}, T_{dbl}, and T_{odd} are defined as follows:

$$\langle T_{vol}\rangle = \frac{1}{4}\begin{bmatrix} 2 & 2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

$$T_{dbl} = \begin{bmatrix} |\alpha|^2 & \alpha & 0 \\ \alpha^* & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{10}$$

$$T_{odd} = \begin{bmatrix} 1 & \beta^* & 0 \\ \beta & |\beta|^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{11}$$

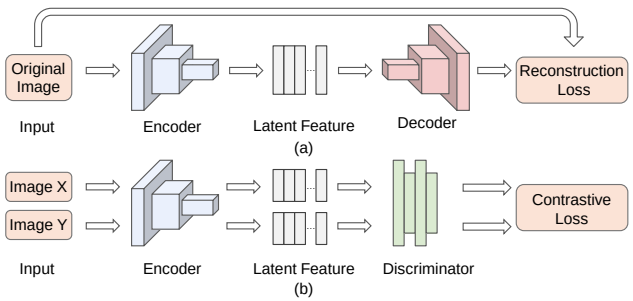

Fig. 1. Comparing generative SSL and contrastive SSL concepts. (a) Generative SSL. (b) Contrastive SSL.

where α and β are parameters for the double and odd scatter models. Therefore, the scattering powers P_v, P_d, and P_s for each model are calculated as follows.

$$P_v = \frac{3}{8} f_v \tag{12}$$

$$P_d = f_d(1 + |\alpha|^2) \tag{13}$$

$$P_s = f_s(1 + |\beta|^2) \tag{14}$$

B. Self-Supervised Learning

Due to the complex imaging mechanism of PolSAR, its pixel-level manual labeling relies on expert knowledge and a lot of manpower and time. This indicates that having substantial and accurate labeled data is unrealistic in practical tasks. However, DL end-to-end operation implies an insufficient prior hypothesis, which could cause overfitting or obtaining incorrect results with insufficient training samples [27].

SSL mines data intrinsic co-occurrence relationships as self-supervision signals and ensures that the learned high-level features are also effective for downstream tasks. It is one of the useful ideas for solving the few-shot problem. Based on the architecture and objectives of SSL, it can be broadly divided into two categories: generative SSL and contrastive SSL, and their rough conceptual diagrams are shown in Fig 1.

Generative SSL primarily includes generative adversarial networks (GAN) [28] and autoencoder (AE) [29], which generate new data from the original data, aiming to make the generated data as close as possible to the original data. Contrastive SSL enables the model to distinguish similar samples from dissimilar samples by comparing them, facilitating the acquisition of invariant features across distinct instances. In comparison to generative SSL, contrastive SSL presents simpler and more direct tasks, demonstrating greater effectiveness in computer vision classification tasks. Currently, a series of effective contrastive SSL frameworks, such as CMC [24], MoCo [30], SimCLR [31], BYOL [32], and SimSiam [33] have been proposed. Among them, BYOL and SimSiam discard negative samples, achieving enhanced efficiency while maintaining high performance.

Many researchers have yielded impressive results by designing classification methods based on SSL in PolSAR image classification domain. [21] first introduced the concept of SSL into PolSAR data processing, extracting mutual information

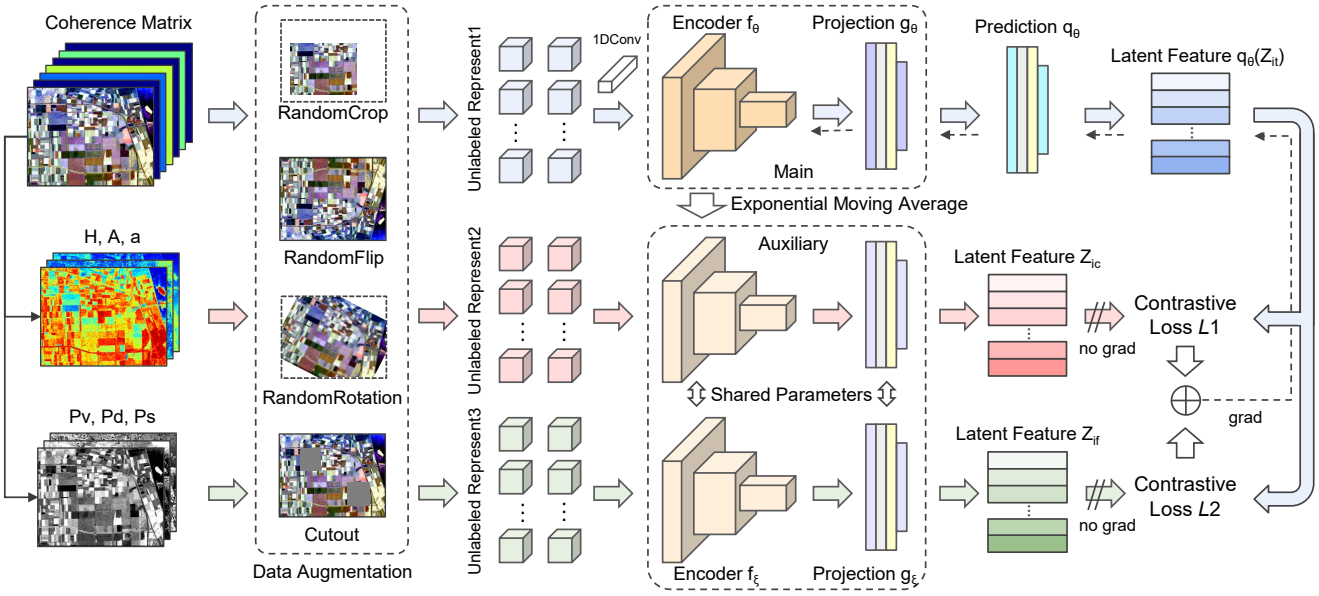

Fig. 2. Architecture of proposed SSL-MBC for Few-Shot PolSAR image classification. There is a main branch with a prediction head and two auxiliary branches in the multi-branch SSL.

from different representations of unlabeled data to provide prior knowledge for the model, and achieving notable results even under circumstances with limited sample availability. [22] combined a diversity stimulation mechanism with contrastive SSL and learned transferable representations from unlabeled PolSAR data using convolutional architecture. [23] investigated a SSL framework without negative samples that achieved excellent performance when combined with the mix-up regularization strategy. [34] utilized a distribution-inspired positive sample generation strategy for representation learning and designed a hybrid anti-imbalance scheme to solve the class imbalance problem. Moreover, methods based on SSL have been widely applied in region detection [35], ship identification [36], and other fields.

III. Proposed Method

We propose a SSL-MBC framework and a comprehensive overview of related techniques is provided in this section. The primary focus is based on the natural properties of PolSAR data, using multimodal representations of the same instance as learning pairs to avoid the dilemma that only spatial data augmentation is not easy to control. The designed multi-branch SSL framework unifies the consistency between branches to mine potential connections across modals while discarding negative sample pairs to cope with the computational increase in modal number. Furthermore, the specifics of the encoder designing and other components within the framework, along with the process of transferring the model to downstream tasks, are also described.

A. Framework of SSL-MBC

The SSL-MBC framework follows the pretrain-finetune paradigm commonly used in SSL methods. The model is first trained by constructing supervised information from massive unlabeled data via the pretext task, and then the trained model is transferred to labeled downstream tasks for fine-tuning to achieve satisfactory performance.

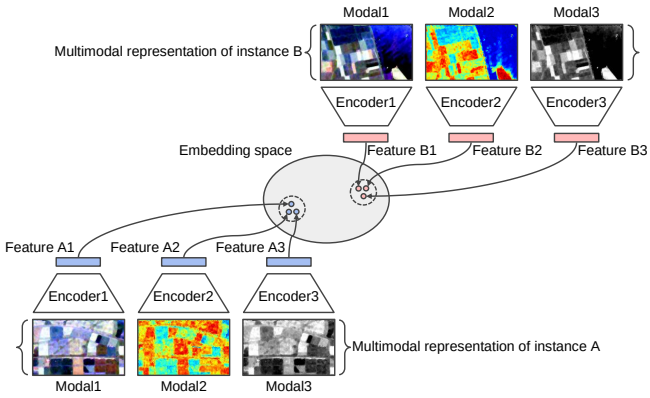

Fig. 3. The schematic diagram shows that different modal representations of the same instance are similar in the encoding space.

The overall architecture of SSL-MBC is presented in Fig. 2. Firstly, multi-branch representations are constructed by different decompositions (Cloud-Pottier decomposition and Freeman-Durden decomposition), and the data augmentation is performed separately. The obtained representations consider both rich spatial transformations and the natural properties of PolSAR data. Secondly, different modal representations are fed into the multi-branch SSL framework without negative samples for embedding and projection, accomplishing implicit knowledge learning by maintaining consistency across branches. Notably, the branch with prediction header is called the main branch, and the rest are auxiliary branches. Z_{it}, Z_{ic}, and Z_{if} are used to represent the hidden features extracted by different branches. Since the proposed method aims to address the problem of low and difficult-to-obtain labeled data, no

artificial labels are involved at all in SSL pretraining.

B. Multi-Branch Representations Construction

The efficacy of input representations significantly influences the acquisition of knowledge during self-supervised training processes, where X denotes the input space and R the space of representations. The training seeks to optimize a mapping $f : X \rightarrow R$ to minimize the objective function $\mathcal{L}$. For optical images, data augmentation combining geometric and color transformations has been proven to be crucial for learning effective representations [31]. While PolSAR images lack spectral features, the aforementioned perspective motivates us to generate enhanced PolSAR data by simultaneously altering the spatial and scattering dimensions. In the spatial dimension, certain effective transforms such as cropping and flipping can be directly applied from the optical image domain. Additionally, polarimetric features derived through eigenvalue decomposition using the covariance matrix yield comparable results to those computed in the scattering dimension of the original data. This characteristic holds significance not only for PolSAR data, but also distinguishes it from other types of data. Second, it is widely recognized that crucial elements are shared between multiple views [24]. Due to various scattering features in PolSAR data, we believe the most core signal is present in various features. Hence, our approach is not limited to one or two modalities but extended to more.

The build process of the pretext task is shown in the first half of Fig. 2. The elements in the upper triangular part of the covariance matrix, $\{H,\ A,\ \alpha\}$ obtained by the Cloude-Pottier decomposition, and $\{P_v,\ P_d,\ P_s\}$ derived from the Freeman-Durden decomposition are selected as multi-modal representations for PolSAR data. The Cloude decomposition statistically characterizes the target, providing a quantitative understanding of its scattering complexity and orientation. The Freeman decomposition yields components that distinctly represent the various physical scattering mechanisms inherent in the target, offering insights into its underlying physical properties. The distinctiveness of the input features across different decomposition branches enables the model to learn the generalized information better. They undergo several spatial transforms, namely random cropping, random flipping, random rotation, and cutout, respectively, to obtain the final unsupervised representations. Where cutout is to mask several rectangular regions randomly on a patch [37]. As shown in Fig. 3, different modal representations of the same instance in this task are considered similar after encoding processing. The goal of training the model in the self-supervised stage is to maximize the similarity of features extracted from different modal representations of the same instance. It is noted that since our approach does not involve negative samples, differences between modes of different instances are not considered.

C. Architecture of Encoder

Benefiting from the local connectivity and shared weights, CNN is still the most popular feature extractor in PolSAR classification [38], [39]. Typically, it consists of convolutional layers, nonlinear activation functions, pooling layers, and fully connected layers. To accomplish the intended objective, CNN must demonstrate a certain level of complexity to achieve satisfactory fitting. However, due to the nonlinear transformations, deeper networks will suffer from a degradation in performance due to the difficulty of achieving consistent transformations [40].

Residual blocks (ResBlock) provide a direct solution to deep model degradation [41]. Its underlying mapping $R(x)$ for input x is represented as:

$$R(x) = F(x) + x \tag{15}$$

where $F(x)$ denotes the mapping of all layers in the block for x. ResBlocks avoid gradient vanishing by forming a constant mapping through the cross-layer connection. Since PolSAR image classification is a pixel-level task, the small size patch around a particular pixel used for its representation is simultaneously fed into the encoder. Residual networks (ResNet) have demonstrated impressive performance in computer vision, particularly ResNet-18 and ResNet-50. However, these models exhibit notable feature degradation and excessive pooling as a result of their high number of layers. Consequently, they are not suitable for direct application to PolSAR image tasks [42].

As shown in the Fig. 4, a light ResNet suitable for processing PolSAR patches is designed as the encoder $f(\cdot)$ of this framework. It primarily comprises of three ResBlocks and a global average pooling layer, exhibiting a more profound structure compared to [21]–[23]. Each ResBlock consists of two convolutional layers and two batch normalization (BN) layers alternately, and the stride of the kernel in the first layer has a stride of 2 in order to accomplish feature downsampling. The number of convolutional layer kernels in the three ResBlocks are 32, 64, and 128, respectively, and the size of all kernels is 3×3.

D. Projection and Predictor

The presence of projection and predictive heads has been confirmed to improve SSL by transferring the loss of valid information [31]–[33]. For the encoder output $f(x)$, the projection head $g(\cdot)$ transforms it into $g(f(x))$, thus mapping it to the space where the contrastive loss is applied. Specifically, the purpose of SSL is to eliminate the effect of data augmentation on the input style while keeping the content unchanged. However, since the implementation of data augmentation is not task-independent, it is impractical to leave the content information unaffected during training [43]. The projection head

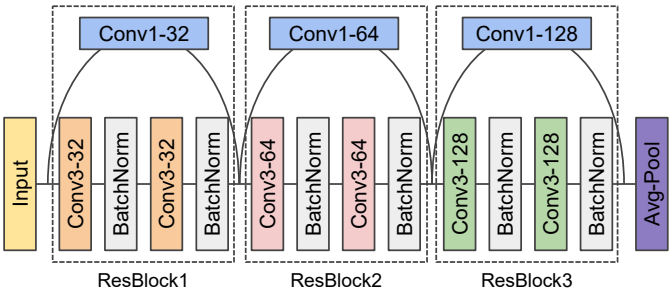

Fig. 4. The structure of the encoder.

encodes the encoder's backbone feature $f(x)$ into $g(f(x))$ for loss computation, which protects the diversity of $f(x)$ content to a certain extent and allows it to generalize better in downstream tasks. Given the discussion in [31], the projection head in this framework is nonlinear and comprises two linear connection layers, a BN layer and a ReLU nonlinear activation layer.

The prediction head $q(\cdot)$ further maps the projection head output $g(f(x))$ to $q(g(f(x)))$, which exists solely in the main branch and renders the model more 'flexible' in the sense. Whether the loss is symmetric or asymmetric, the SSL framework without negative samples collapses when the prediction head is absent [33]. In the proposed framework, the structure of projection head $q(\cdot)$ is the same as $g(\cdot)$.

E. Loss Function and Optimization

The goal of this SSL framework is to enable the encoder to learn effective features that generalize well in downstream tasks. For a batch with input $X = \{x_1, x_2, x_3, ..., x_n\}$, the multimodal representations of $x_i \in X$ are x_{it}, x_{ic}, and x_{if}. In the main branch each network parameter is defined by a set of weights θ with three components $f_\theta(\cdot)$, $g_\theta(\cdot)$, and $q_\theta(\cdot)$. The weights in both auxiliary branches are ξ and only $f_\xi(\cdot)$, $g_\xi(\cdot)$. From x_{it}, the main branch produces a feature $y_{it} = f_\theta(x_{it})$ and a projection $z_{it} = g_\theta(y_{it})$. To facilitate parameter migration, x_{it} reduces the channel depth to the same as x_{ic} and x_{if} by a 1D convolution before input to the main branch. Similarly, x_{ic} and x_{if} respectively undergo auxiliary branching to obtain features $y_{ic} = f_\xi(x_{ic})$, $y_{if} = f_\xi(x_{if})$ and projections $z_{ic} = g_\xi(y_{ic})$, $z_{if} = g_\xi(y_{if})$. Due to the presence of the predictor head $q_\theta(\cdot)$ in the main branch, we end up with $q_\theta(z_{it})$, z_{ic} and z_{if}.

Considering the computational burden and the fact that both decompositions are performed based on the coherence matrix, $q_\theta(z_{it})$ obtained from representation x_{it} is considered as 'core' in this framework. The 'core' representation distinguishes itself from other representations by requiring optimization and calculating individual losses with them. Specifically, the following mean squared error loss $\mathcal{L}(x_{it}, x_{ic})$ is used to measure the similarity between l_2-normalized $q_\theta(z_{it})$ and z_{ic}:

$$\mathcal{L}(x_{it}, x_{ic}) = 2 - 2 \cdot \frac{\langle q_\theta(z_{it}), z_{ic} \rangle}{||q_\theta(z_{it})||_2 \cdot ||z_{ic}||_2} \tag{16}$$

According to Eq. (16), the loss $\mathcal{L}(x_{it}, x_{if})$ between $q_\theta(z_{it})$ and z_{if} is similarly obtained. Therefore, the final loss $\mathcal{L}_{\text{final}}$ is generalized as:

$$\mathcal{L}_{\text{final}} = \sum_{n=ic,if} \mathcal{L}(x_{it}, x_n) = \mathcal{L}(x_{it}, x_{ic}) + \mathcal{L}(x_{it}, x_{if}) \tag{17}$$

Based on the aforementioned results, stochastic gradient descent (SGD) is employed on every batch of samples using back propagation to minimize the loss during model training. Notably, only the main branch parameter θ is updated in the gradient propagation. The auxiliary branch provides regression objective for the main branch, and its parameter ξ is updated by the exponential moving average strategy (EMA) with decay rate ρ [44]. As shown in Fig. 2, the framework parameter update can be summarized as:

$$\theta = optimizer(\theta, \mathcal{L}_{\text{final}}, \eta) \tag{18}$$

$$\xi = \rho\xi + (1 - \rho)\theta \tag{19}$$

where $optimizer$ is the specified optimizer. η is the learning rate and $\rho \in [0, 1]$. EMA allows $f_\xi(\cdot)$, $g_\xi(\cdot)$ to be updated from $f_\theta(\cdot)$, $g_\theta(\cdot)$ with a momentum state, which ensures a consistent and stable output. This strategy effectively maintains the dissimilarity of features between the primary and auxiliary branches, thereby preventing the model from adopting a simplistic solution. Furthermore, it facilitates iterative partial parameter updating.

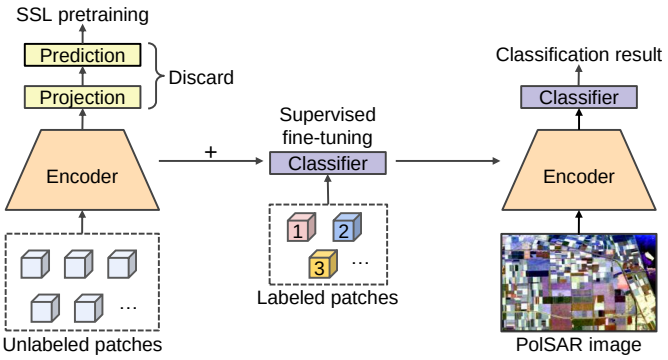

Fig. 5. The final fine-tuning process. The encoder is migrated to the downstream task and only a few labeled samples are used to train the new classifier.

F. Final Fine-Tuning

During the SSL training phase, the encoder is equipped with the ability to extract generalized discriminative features through the pretext task. However, supervised fine-tuning the trained model is required to enhance its applicability before migrating it to downstream tasks.

The final fine-tuning process is shown in Fig. 5. In fine-tuning, the encoder $f_\theta(\cdot)$ in the main branch of SSL-MBC is constantly transferred to the downstream task for feature extraction, but the projection and prediction heads are discarded. In addition, a new linear fully connected layer with randomly initialized parameters is added to the encoder back as a classifier, whose neuron count equals category count. Only a few labeled samples are used for supervised training, which is perfectly adequate for classifiers with low complexity. Remarkably, the migrated encoder parameters are untrained and completely unchanged during the fine-tuning phase.

IV. Experimental Results and Analysis

A. Description of Experimental Datasets

Three representative PolSAR datasets Flevoland, San Francisco, and Oberpfaffenhofen are chosen to conduct the experiments. Specific descriptions of the individual dataset are given below:

1) Flevoland: As shown in Fig. 6, the Flevoland dataset was obtained by the ARISAR system of NASA/JPL laboratory

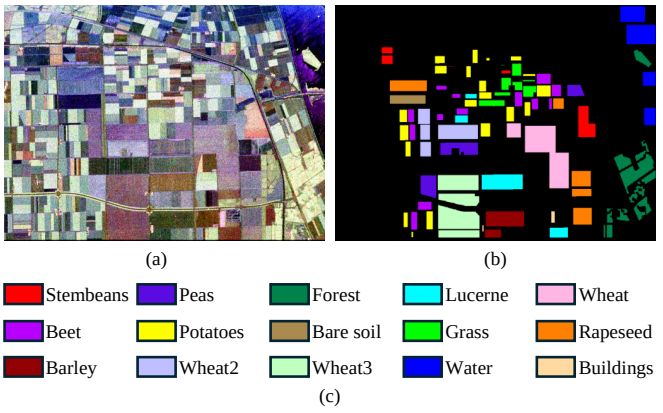

Fig. 6. Flevoland dataset. (a) Pauli-RGB image. (b) Ground truth. (c) Category legend.

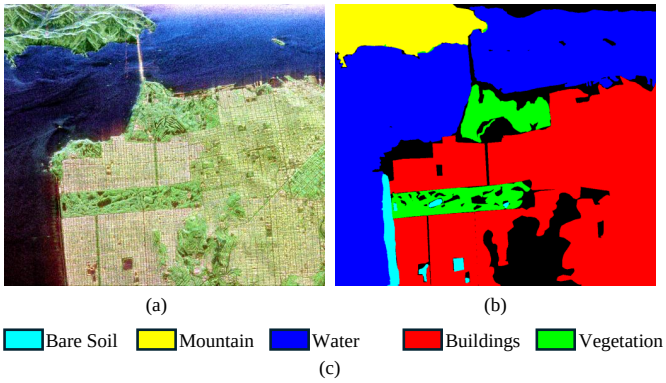

Fig. 7. San Francisco dataset. (a) Pauli-RGB image. (b) Ground truth. (c) Category legend.

in 1989 by detecting the ground over the Flevoland region of Poland, which consists of 1024×750 pixels. There are 15 land objects in the labeled area, namely stembeans, peas, forest, lucerne, wheat, beet, potatoes, bare soil, grass, rapeseed, barley, wheat2, wheat3, water, and buildings.

2) San Francisco: As depicted in Fig. 7, the San Francisco dataset was obtained by the RadarSat-2 satellite over the San Francisco region of the United States in 2008, which consists of 1024×900 pixels. There are 5 types of ground features within the labeled area, including bare soil, mountain, water, buildings, and vegetation.

3) Oberpfaffenhofen: As shown in Fig. 8, the Oberpfaffenhofen dataset was collected by the German Aerospace Center during ground exploration over the Oberpfaffenhofen region of Germany, which consists of 1300×1200 pixels. There are 3 surface types within the marked area, comprising built-up areas, wood land, and open areas.

B. Experimental Setting

Due to the inherent limitations of the image system, the raw PolSAR images exhibit severe speckle noise [45]. The data preprocessing has been performed before formal sampling. First, a refined Lee filtering with a 7×7 window is used to filter each dataset [46]. Then, Cloude-Pottier decomposition and Freeman-Durden decomposition are applied to obtain the

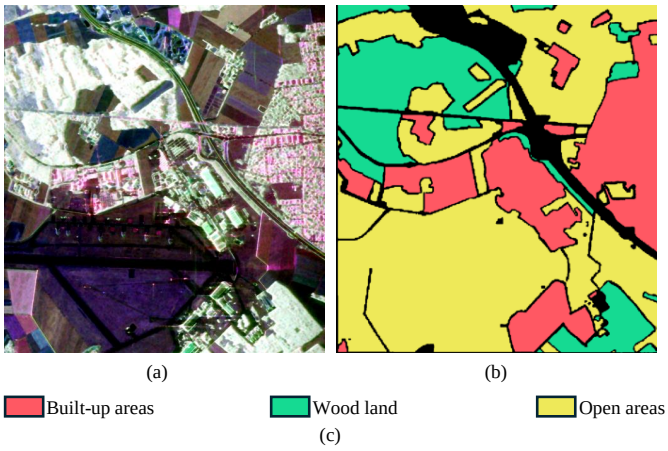

Fig. 8. Oberpfaffenhofen dataset. (a) Pauli-RGB image. (b) Ground truth. (c) Category legend.

corresponding scattering features. Additionally, the data are standardized in each scattering dimension.

In data augmentation, a patch is first randomly cropped at a ratio between [0.8, 1] and restored to the initial size. It is then flipped horizontally and vertically with a chance of 0.5, and rotated at an angle between [-30°, 30°]. Finally, two 2×2 regions in the patch are randomly masked out.

The details of the encoder have been displayed in Fig. 4. For projection and prediction heads, the dimensions of their linear connected layers are 128 and 32, respectively. In the SSL phase, 20% of the unlabeled samples from each dataset are selected for representation learning. The optimizer is a SGD with η of 0.01, momentum of 0.9, and weight decay of 0.0001, which performs 100 epochs and reduces n by half at epoch=60. In addition, the batch size is 512 and the decay rate ρ is 0.996.

For fine-tuning, we use only 50 labeled samples per class to form the few-shot case. The sampling rates in the above three datasets are 0.47%, 0.03%, and 0.01%, respectively. An Adam optimizer with η of 0.01 is used to train the newly added classifier with 100 epochs and the batch size is 64.

To demonstrate the effectiveness of the proposed methods, six advanced methods are chosen for comparison, including SVM [47], MLP [48], CNN [16], CV-CNN [49], PCLNet [22], SSPRL [23], PiCL [34]. Among them, SVM is a machine learning method. MLP, CNN, and CV-CNN are DL methods in the supervised paradigm. PCLNet, SSPRL, and PiCL are state-of-the-art SSL methods for PolSAR classification. Furthermore, four common metrics, namely overall accuracy (OA), average accuracy (AA), kappa coefficient, and class accuracy, are employed to assess the results. Each method was individually replicated ten times across every dataset, enabling the calculation of mean values for four designated metrics, as well as the variance for the first three metrics.

C. Experimental Results of Flevoland

The experimental results of each approach on the Flevoland dataset are shown in Fig. 9. Due to the variety of targets in this dataset, the results of each classification method relying

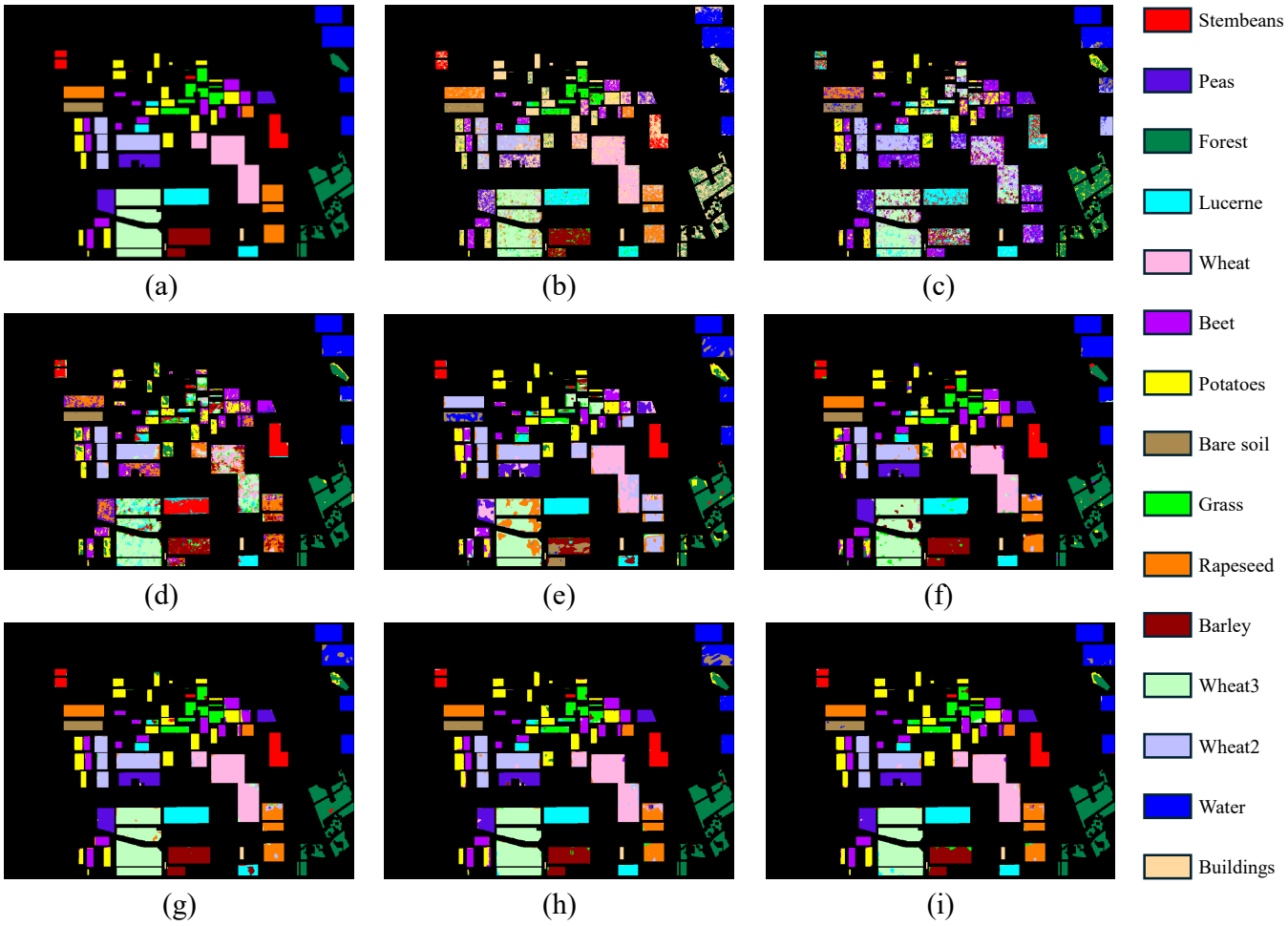

Fig. 9. Classification maps of the Flevoland dataset. (a) Ground truth. (b) SVM. (c) MLP. (d) CNN. (e) CV-CNN. (f) PCLNet. (g) SSPRL. (h) PiCL. (i) SSL-MBC.

on supervised training are not satisfactory in the case of small samples. Specifically, SVM and MLP have substantial misclassified pixels in each type of region. Both CNN and CV-CNN are limited by insufficient training samples and have wide consecutive errors in some species such as beet and wheat, which is the main factor for their poor performance. Several SSL-based methods have the ability to extract discriminative features that distinguish between categories through pretraining, thus giving more accurate results. Although PCLNet, SSPRL, and PiCL have achieved a certain level of accuracy, obvious error patches are still noticed. Our SSL-MBC intuitively achieves the best results with relatively few misclassified pixels and consistent performance across all classes.

The quantitative metrics of the selected methods are shown in the Table I. The SSL-MBC proposed in this study obtains the best values on all three metrics, OA, AA, and Kappa, which corresponds to the results presented in Fig. 9. It is important to highlight that SSL-MBC demonstrates not only the highest average but also the lowest variability across the aforementioned metrics. This implies that it excels in performance while simultaneously ensuring strong stability.

D. Experimental Results of San Francisco

The experimental results of each approach on the San Francisco dataset are shown in Fig. 10. Due to lacking labeled data, the model cannot achieve a stable solution, resulting in unsatisfactory results for all supervised training methods. There are a substantial number of misclassification points in the categories of mountain, building, water, and vegetation, which have a large distribution range. The performance of PCLNet shows significant improvement over the previously mentioned methods; however, it still encounters challenges in differentiating between building and vegetation pixels. Affected by the imbalance of category samples, SSPRL has large misclassifications around the bare soil category. PiCL presents a result map that is notably cleaner, but there still remains a significant cluster of errors observed on the right side of the sampled area. Compared to the above methods, SSL-MBC shows a more reliable performance and provides excellent identification of regions that are commonly susceptible to incorrect judgment.

The quantitative metrics of the selected methods are shown in the Table II. It is clear that the designed SSL-MBC achieves the most accurate and stable results. It achieves the highest OA of 96.71% and 95.38% on bare soil and buildings, respectively.

TABLE I
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE FLEVOLAND DATASET

Class	SVM	MLP	CNN	CV-CNN	PCLNet	SSPRL	PiCL	SSL-MBC
Strambeans	43.00	40.44	79.86	83.86	96.67	98.05	99.15	**99.17**
Peas	54.74	52.93	45.68	67.35	98.11	**98.69**	97.48	98.61
Forest	17.62	80.81	87.15	79.95	91.67	97.05	97.19	**98.20**
Lucerne	87.19	85.11	48.74	75.35	91.48	93.43	**96.06**	95.13
Wheat	65.90	25.86	47.67	71.89	88.05	91.73	91.68	**95.18**
Beat	67.76	56.51	73.27	83.05	94.58	95.04	96.39	**97.21**
Potatoes	18.39	63.31	67.44	76.44	84.75	87.27	95.14	**98.25**
Bare Soil	98.07	89.95	99.86	71.98	**99.98**	99.97	99.12	98.61
Grass	83.93	14.74	31.41	28.43	92.03	**94.90**	88.96	90.09
Rapeseed	72.23	30.37	49.73	18.50	88.90	84.67	92.99	**93.65**
Barley	91.55	34.03	64.08	68.05	97.73	**98.49**	95.85	97.81
Wheat2	74.59	78.30	84.28	84.07	92.19	95.51	**95.97**	95.69
Wheat3	80.69	64.45	68.28	66.05	90.64	95.08	**98.13**	97.82
Water	91.26	79.88	96.85	95.46	**98.62**	82.29	89.94	97.65
Buildings	98.32	76.73	96.85	98.21	99.09	**99.47**	98.79	99.12
OA(%)	64.45±0.25	57.48±3.27	67.08±1.28	70.14±2.59	92.09±0.22	92.71±0.28	95.10±0.18	**96.72**±0.04
AA(%)	69.68±0.08	58.43±2.53	69.41±3.49	71.24±0.19	93.63±0.15	94.11±0.11	95.52±0.19	**96.81**±0.06
Kappa(×100)	62.17±0.24	53.78±3.37	64.20±1.59	67.49±2.87	91.37±0.26	92.06±0.33	94.65±0.22	**96.42**±0.05

TABLE II
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE SAN FRANCISCO DATASET

Class	SVM	MLP	CNN	CV-CNN	PCLNet	SSPRL	PiCL	SSL-MBC
Bare Soil	87.54	88.28	88.75	86.37	87.50	94.19	95.61	**96.71**
Mountain	76.46	77.30	74.90	87.01	92.49	**94.99**	93.43	94.05
Water	92.17	86.65	87.42	89.52	93.10	92.58	**95.94**	95.19
Buildings	87.44	82.38	85.62	84.65	89.56	89.64	92.31	**95.38**
Vegetation	75.86	77.65	80.54	80.70	83.27	**89.87**	85.91	87.42
OA(%)	87.75±0.83	83.52±3.81	85.23±3.08	86.60±2.34	90.79±0.17	91.36±0.90	93.52±0.09	**94.69**±0.12
AA(%)	83.89±0.44	82.45±1.92	83.44±6.76	85.65±2.83	89.18±4.30	92.25±0.87	92.64±0.21	**93.57**±0.08
Kappa(×100)	81.38±1.74	75.89±6.95	78.12±6.31	80.11±4.67	86.03±0.32	86.98±1.91	90.08±0.19	**91.83**±0.25

TABLE III
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE OBERPFAFFENHOFEN DATASET

Class	SVM	MLP	CNN	CV-CNN	PCLNet	SSPRL	PiCL	SSL-MBC
Built-up Areas	63.89	52.85	70.48	70.56	**82.53**	75.06	77.36	81.90
Wood Land	82.79	89.55	83.33	74.68	86.65	**93.11**	89.08	91.90
Open Areas	83.10	93.15	91.63	**95.49**	89.41	91.55	93.79	93.77
OA(%)	78.24±0.64	82.40±2.11	84.78±2.92	85.34±0.26	87.17±0.11	87.72±0.19	88.80±0.37	**90.45**±0.10
AA(%)	76.59±0.83	78.52±1.91	81.82±1.80	80.24±0.60	86.20±0.47	86.57±0.18	86.75±0.50	**89.19**±0.30
Kappa(×100)	63.99±1.31	69.97±5.02	74.05±6.93	74.64±0.81	78.47±0.35	79.33±0.54	80.96±1.00	**83.83**±0.29

This indicates that abstract features derived from multimodal representations possess greater significance and exhibit a stronger ability to generalize in subsequent classification tasks.

E. Experimental Results of Oberpfaffenhofen

The experimental results of each approach on the Oberpfaffenhofen dataset are shown in Fig. 11. SVM and MLP identified numerous build-up areas as wood land and open areas. CNN and CV-CNN suffer from severe misclassification in all three categories included in the dataset, especially in the middle and bottom right regions where the classes are more complex. On the contrary, the result plots of the SSL-trained PCLNet, SSPRL, and PiCL demonstrate superior quality. Although they do not show large errors, the presence of scattered error points still limits their accuracy. Notably, the SSL-MBC classification map appears cleaner due to its precision, which is evident in the open areas.

The quantitative metrics of the selected methods are shown in the Table III. Because there are few categories in oberpfafenhofen dataset and each kind of pixel is many, the OA of all methods reaches a certain level. Although our SSL-MBC does not reach the highest accuracy in any of the three categories, the overall OA, AA and Kappa are optimal, reaching 90.45%, 89.19%, and 83.83% respectively. In particular, SSL-MBC is advantageous in AA, which means that it possesses discriminative features that are practical for all classes and not limited to one class.

F. Ablation Study

In order to exhaustively illustrate the effectiveness of data augmentation and multi-branch design in the SSL-MBC

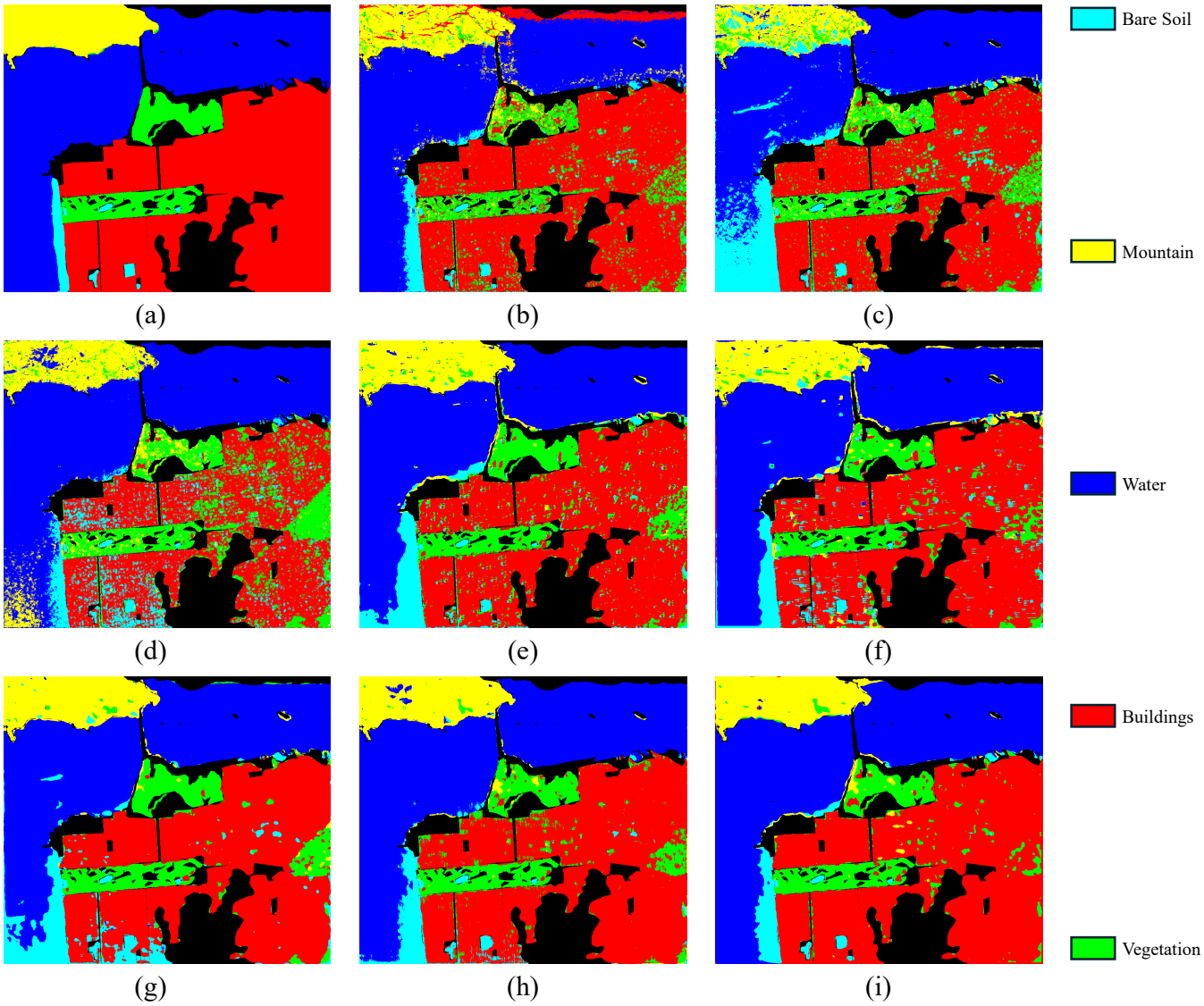

Fig. 10. Classification maps of the San Francisco dataset. (a) Ground truth. (b) SVM. (c) MLP. (d) CNN. (e) CV-CNN. (f) PCLNet. (g) SSPRL. (h) PiCL. (i) SSL-MBC.

framework, we have conducted comprehensive ablation experiments. Designate data augmentation, the Cloude-Pottier decomposition branch, and the Freeman-Durden decomposition branch with the symbols ”A,” ”C,” and ”F,” respectively. The base frame with all the above parts removed is the “Base”, at which point it does not make sense to learn a representation between the input and itself. We gradually add other modules to “Base” and conduct separate experiments on each dataset.

The ablation results for all datasets are shown in the Table IV. Although the specific values are different, the trends presented in each result are similar. According to ”Base+C+F”, data augmentation is crucial in improving the training effect, and when it is masked, the classification accuracy decreases in all scenarios. Combining multiple augmentations makes the comparison task more difficult, resulting in a higher quality of features learned by the model. When the model is not fine-tuned with a limited number of labels, the newly added classifier parameters are randomized. As a result, the classification outcomes are purely random. Besides, more branch consistency means better performance. Specifically, ”Base+A+C+F” based on multimodal representations is better than ”Base+A+C” and ”Base+A+F” which mask one modality, while they are better than ”Base+A” that use only a single modality. This supports the viewpoint of the paper, which states that the inclusion of key attributes in the modal representations of the target enhances its overall credibility. Further constraints are imposed on the invariance of hidden features by incorporating multiple modalities. This approach eliminates the dependence on a single or limited number of modalities, reducing the potential for contingency. Consequently, the model can prioritize stable, fundamental, and widely applicable features.

G. Impact of Labeled Sample Size

In the fine-tuning phase, a limited number of labeled samples are given to facilitate supervised training of the model’s

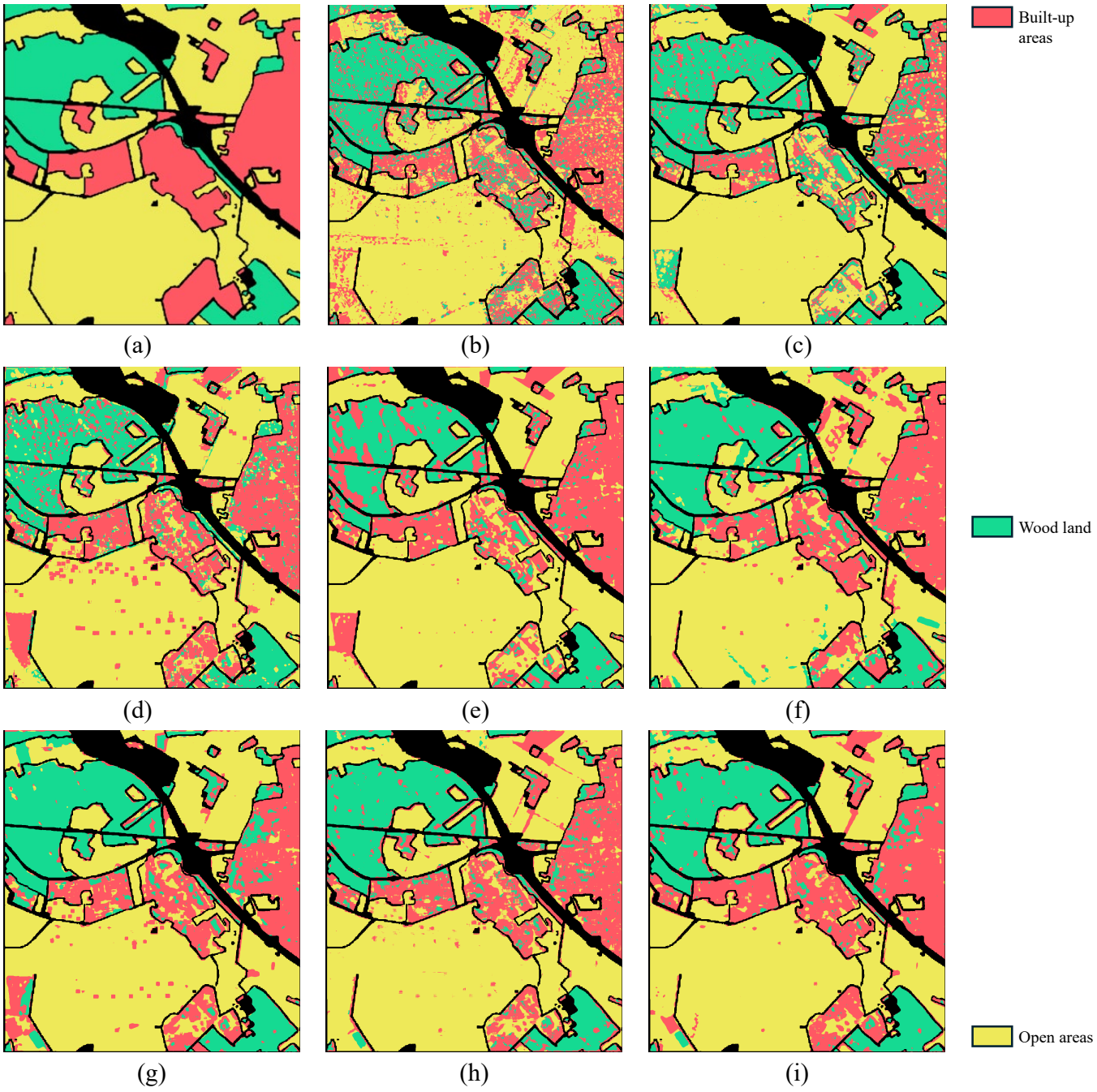

Fig. 11. Classification maps of the Oberpfaffenhofen dataset. (a) Ground truth. (b) SVM. (c) MLP. (d) CNN. (e) CV-CNN. (f) PCLNet. (g) SSPRL. (h) PiCL. (i) SSL-MBC.

classifier. To explore the effect of the labeled sample size on the results, we conducted independent experiments at sample sizes of 10, 20, 30, 50, 150, 200, and 300 per category concerning the settings in [22]. Additionally, an encoder that is not pre-trained in SSL is utilized for comparison, which is referred to as "W/O SSL-MBC". In "W/O SSL-MBC", the encoder parameters are only randomly initialized without migration and are updated together with the classifier during the training process.

The comparison results on three datasets are shown in Fig. 12. Intuitively, the performance of the encoder without SSL is greatly degraded on either dataset, which is especially obvious when the sample size is not more than 50. On the Flevoland dataset, due to the complexity of its object types, the encoder still fails to have a satisfactory performance when labeled samples per class reach 300. This indicates that SSL-MBC is necessary for the encoder to obtain excellent classification performance. Moreover, according to the curve variation, SSL-MBC maintains high accuracy even if only 10 labeled samples are available for fine-tuning. The rise in labeled sample count is not significant for the performance improvement of SSL-MBC, suggesting that our method does not rely on a large

TABLE IV
ABLATION STUDIES ON ALL SELECTED DATASETS

Methods	Flevoland			SanFrancisco			Oberpfaffenhofen		
	OA(%)	AA(%)	Kappa×100	OA(%)	AA(%)	Kappa×100	OA(%)	AA(%)	Kappa×100
Base	-	-	-	-	-	-	-	-	-
Base+A	93.29	93.89	92.69	90.74	90.72	86.00	85.78	83.03	76.08
Base+A(w/o fine-tuning)	6.80	10.64	1.12	26.81	29.86	9.37	55.02	39.74	26.04
Base+A+C	94.69	94.61	94.20	92.07	90.42	87.93	87.88	85.56	79.48
Base+A+F	94.20	94.48	93.67	91.76	89.86	87.42	88.80	87.47	81.11
Base+C+F	94.65	95.47	94.17	93.92	93.44	90.70	88.89	87.85	81.26
Base+A+C+F	96.83	96.89	96.55	94.69	93.75	91.83	90.45	89.19	83.83

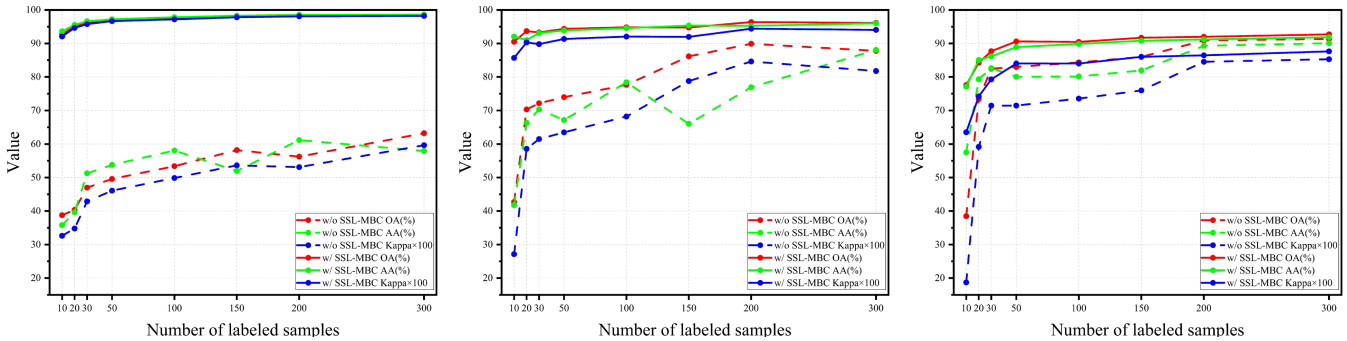

Fig. 12. Performance comparison of encoder trained with different numbers of labeled samples. The solid and dashed lines represent the results with and without SSL-MBC, respectively. (a) Flevoland. (b) San Francisco. (c) Oberpfaffenhofen.

amount of training data. Finally, each metric of "W/ SSL-MBC" outperforms "W/O SSL-MBC" in all cases, which partly implies that SSL mines better supervisory information than manual labeling.

H. Embedding Feature Visualization

To further illustrate the elevation of the encoder features by SSL-MBC, the high-dimensional features are projected to the 2D plane for visualization via t-SNE [50]. Specifically, we compare the results of encoder feature visualization with and without SSL. Due to the excessive sample volume in the San Francisco and Oberpfaffenhofen datasets, only 10,000 samples of each class were randomly selected for visualization.

The t-SNE feature visualization for each dataset are shown in Fig. 13, where scatter colors are the same as ground truth label colors. It is evident that the embedded features produced by the encoder trained by SSL-MBC are robust, implying that the features are more compact and separable. Encoders following supervised training using only a few labeled samples fail to converge or overfit due to insufficient training. Its features are relatively dispersed and there is substantial overlap among various point classes. In Fig. 13, we mark some heavily overlapping regions with red boxes to make it more intuitive.

V. CONCLUSION

To address the issue of limited samples in PolSAR classification tasks, a new approach called SSL-MBC is proposed. In pretext task design, our approach emphasizes the importance of data augmentation through simultaneous transformations in both spatial and channel dimensions, treating the extraction of scattering features as a transformation within the scattering dimension of PolSAR images. This task takes advantage of scattered features, the inherent characteristic of PolSAR data, and terms it as multimodal representations. Furthermore, the diverse modal representations of a given instance exhibit similarity in the embedding space, with features consistently present across multiple modalities considered central. Therefore, we perform SSL training with multi-branch consistency. Since branch expansion increases computational complexity, a SSL architecture without negative samples is adopted, which consists of a main branch with a prediction head and auxiliary branches. Among them, the projection and prediction head optimally preserve the information contained in the encoder features, and the EMA update strategy maintains the parameter difference between the main and auxiliary branches as well as the consistency of outputs. Finally, high accuracy on downstream tasks is achieved by fine-tuning based on the SSL pretraining results. Comprehensive experiments are conducted on three authoritative PolSAR datasets. Supervised training-based methods suffer from underfitting due to insufficient labels. Other recent SSL-based methods fail to adequately mine the hidden information in multimodal data. And our method has the best performance. Furthermore, an ablation study confirms the effectiveness of all modules and settings within the framework.

Although our work holds great significance for future PolSAR SSL studies, the performance is still affected by branch construction and few samples in the target domain. In practical applications, selecting more appropriate scattering features or

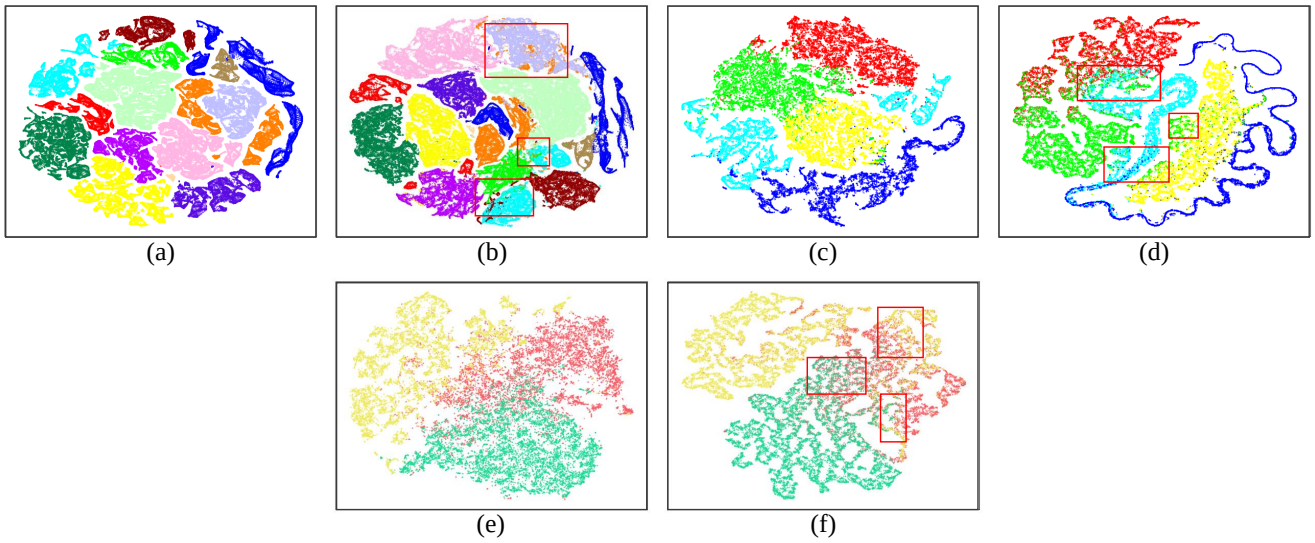

Fig. 13. The feature visualization for each dataset. (a) With SSL-MBC on Flevoland. (b) Without SSL-MBC on Flevoland. (c) With SSL-MBC on San Francisco. (d) Without SSL-MBC on San Francisco. (e) With SSL-MBC on Oberpfaffenhofen. (f) Without SSL-MBC on Oberpfaffenhofen.

considering more modal representations may lead to even more improved results. This ensures the practicality and extensibility of our study. Additionally, exploring the potential of PolSAR image classification without relying on target domain samples could be a novel direction. Using domain generalization and similar methods can achieve precise classification even when the target domain is not visible during the training phase.

REFERENCES

[1] G. Gao, Q. Bai, C. Zhang, L. Zhang, and L. Yao, "Dualistic cascade convolutional neural network dedicated to fully polsar image ship detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 663–681, 2023.

[2] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric sar imagery in complex land cover ecosystem," *ISPRS journal of photogrammetry and remote sensing*, vol. 151, pp. 223–236, 2019.

[3] Guofeng, Liang, Shouzhen, Chen, Jinsong, Qingquan, and Hongzhong, "The impacts of building orientation on polarimetric orientation angle estimation and model-based decomposition for multilook polarimetric sar data in urban areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5520–5532, 2016.

[4] E. Krogager, "New decomposition of the radar target scattering matrix," *Electronics Letters*, vol. 18, no. 26, pp. 1525–1527, 1990.

[5] W. Cameron and L. Leung, "Feature motivated polarization scattering matrix decomposition," in *IEEE International Conference on Radar*, 1990, pp. 549–557.

[6] S. Cloude and E. Pottier, "An entropy based classification scheme for land applications of polarimetric sar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 1, pp. 68–78, 1997.

[7] A. Freeman and S. Durden, "A three-component scattering model for polarimetric sar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 36, no. 3, pp. 963–973, 1998.

[8] Y. Yamaguchi, T. Moriyama, M. Ishido, and H. Yamada, "Four-component scattering model for polarimetric sar image decomposition," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 8, pp. 1699–1706, 2005.

[9] X. Wang, L. Zhang, N. Wang, and B. Zou, "Joint polarimetric-adjacent features based on lcsr for polsar image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 6230–6243, 2021.

[10] L. He, X. He, F. Hui, Y. Ye, T. Zhang, and X. Cheng, "Investigation of polarimetric decomposition for arctic summer sea ice classification using gaofen-3 fully polarimetric sar data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 3904–3915, 2022.

[11] Z. Qi, A. G.-O. Yeh, X. Li, and X. Zhang, "A three-component method for timely detection of land cover changes using polarimetric sar images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 107, pp. 3–21, 2015.

[12] W. Li, H. Xia, J. Zhang, Y. Wang, Y. Jia, and Y. He, "Complex-valued 2d-3d hybrid convolutional neural network with attention mechanism for polsar image classification," *Remote Sensing*, vol. 16, no. 16, p. 2908, 2024.

[13] S. Kamada and T. Ichimura, "Automatic extraction of road networks by using teacher-student adaptive structural deep belief network and its application to landslide disaster," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[14] R. Gui, X. Xu, R. Yang, K. Deng, and J. Hu, "Generalized zero-shot domain adaptation for unsupervised cross-domain polsar image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 270–283, 2021.

[15] W. Li, J. Zhang, H. Xia, Q. Liu, Y. Wang, Y. Jia, and Y. Chen, "Cross-scene building identification based on dual-stream neural network and efficient channel attention mechanism," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[16] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric sar image classification using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1935–1939, 2016.

[17] L. Zhang, H. Dong, and B. Zou, "Efficiently utilizing complex-valued polsar image data via a multi-task deep learning framework," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 157, pp. 59–72, 2019.

[18] H. Dong, L. Zhang, and B. Zou, "Exploring vision transformers for polarimetric sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[19] H. Bi, J. Sun, and Z. Xu, "A graph-based semisupervised deep learning model for polsar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2116–2132, 2019.

[20] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1422–1430.

[21] B. Ren, Y. Zhao, B. Hou, J. Chanussot, and L. Jiao, "A mutual information-based self-supervised learning model for polsar land cover classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9224–9237, 2021.

[22] L. Zhang, S. Zhang, B. Zou, and H. Dong, "Unsupervised deep representation learning and few-shot classification of polsar images," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–16, 2022.

[23] W. Zhang, Z. Pan, and Y. Hu, "Exploring polsar images representation via self-supervised learning and its application on few-shot classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[24] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision*. Springer, 2020, pp. 776–794.

[25] S. Cloude and E. Pottier, "A review of target decomposition theorems in radar polarimetry," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 34, no. 2, pp. 498–518, 1996.

[26] W. Xue-song and C. Siwei, "Polarimetric synthetic aperture radar interpretation and recognition: Advances and perspectives," vol. 9, no. R19109, 2020, pp. 259–276.

[27] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2023.

[28] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Computer ence*, 2015.

[29] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2013.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[32] J.-B. Grill, F. Strub, F. Altch'e, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *ArXiv*, vol. abs/2006.07733, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219687798

[33] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[34] Z. Kuang, H. Bi, F. Li, C. Xu, and J. Sun, "Polarimetry-inspired contrastive learning for class-imbalanced polsar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–19, 2024.

[35] P. Han, Y. Peng, Z. Cheng, D. Liao, and B. Han, "Sel-net: A self-supervised learning-based network for polsar image runway region detection," *Remote Sensing*, vol. 15, no. 19, 2023.

[36] W. Qiu, Z. Pan, and J. Yang, "Few-shot polsar ship detection based on polarimetric features selection and improved contrastive self-supervised learning," *REMOTE SENSING*, vol. 15, no. 7, APR 2023.

[37] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017.

[38] Q. Zhang, C. He, B. He, and M. Tong, "Learning scattering similarity and texture-based attention with convolutional neural networks for polsar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[39] J. Shi, T. He, S. Ji, M. Nie, and H. Jin, "Cnn-improved superpixel-to-pixel fuzzy graph convolution network for polsar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[40] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

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

[42] M. Zhao, Y. Cheng, X. Qin, W. Yu, and P. Wang, "Semi-supervised classification of polsar images based on co-training of cnn and svm with limited labeled samples," *Sensors*, vol. 23, no. 4, p. 2109, 2023.

[43] K. Gupta, T. Ajanthan, A. v. d. Hengel, and S. Gould, "Understanding and improving the role of projection head in self-supervised learning," *arXiv preprint arXiv:2212.11491*, 2022.

[44] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[45] X. Nie, X. Huang, B. Zhang, and H. Qiao, "Review on polsar image speckle reduction and classification methods," *Acta Automatica Sinica*, vol. 45, no. 8, pp. 1419–1438, 2019.

[46] J.-S. Lee, M. R. Grunes, and G. De Grandi, "Polarimetric sar speckle filtering and its implication for classification," *IEEE Transactions on Geoscience and remote sensing*, vol. 37, no. 5, pp. 2363–2373, 1999.

[47] C. Lardeux, P. L. Frison, C. Tison, J. C. Souyris, B. Stoll, B. Fruneau, and J. P. Rudant, "Support vector machine for multifrequency sar polarimetric data classification," *IEEE Transactions on Geoscience Remote Sensing*, vol. 47, no. 12, pp. 4143–4152, 2009.

[48] R. Hänsch, "Complex-valued multi-layer perceptrons–an application to polarimetric sar data," *Photogrammetric Engineering & Remote Sensing*, vol. 76, no. 9, pp. 1081–1088, 2010.

[49] Z. Zhang, H. Wang, F. Xu, and Y. Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 12, pp. 1–12, 2017.

[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.