





Occluded Scene Classification via Cascade Supervised Contrastive Learning

Jianming Xu , Yunfei Li , *Graduate Student Member, IEEE*, Qian Shi , *Senior Member, IEEE*, and Lin He , *Member, IEEE*

Abstract—Occlusion handling is crucial for improving the performance of convolutional neural networks (CNNs) in real-world remote sensing images, which are often captured in complex and unconstrained environments. In particular, occlusion scene classification has received significant attention due to its usefulness in various remote sensing tasks. However, existing methods are limited by their dependence on close-world learning, which assumes that all test cases are included in the training set. This is problematic because occlusion is too complex to be thoroughly annotated. To address this issue, we propose a novel contrastive learning-based CNN that can classify out-of-distribution occluded scenes without the need for occlusion annotation. Our approach uses a two-branch subnetwork to learn representations of unoccluded anchor images and occlusion-augmented images. We then employ cascade supervised contrastive learning to make the network's representations invariant to occlusion. Unlike standard contrastive learning, our method leverages category information to avoid incompact intraclass distribution and uses a cascade strategy to hierarchically learn occlusion-invariant representations. Finally, we use a multi-layer perceptron to classify the learned representations and assess representation quality. We evaluate our method on the UAVDT dataset and two simulated datasets, and the results demonstrate that our approach accurately characterizes occluded objects and achieves more precise classification in occlusion scenarios.

Index Terms—Convolutional neural network (CNN), contrastive learning, occlusion images, remote sensing.

I. INTRODUCTION

CONVOLUTIONAL neural networks (CNNs) have greatly impacted remote sensing (RS) and earth observation (EO) in the past few decades, with applications ranging from traffic management [1] and military surveillance [2] to precision agriculture [3]. One of the reasons for its success lies in the fact that CNNs have been adapted to the characteristics of RS images [4], [5], such as multiple scales [6], [7] and rotations [8], [9], [10].

Manuscript received 16 November 2022; revised 1 May 2023; accepted 5 May 2023. Date of publication 11 May 2023; date of current version 22 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62071184 and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011615 and Grant 2023A1515011887. (*Corresponding author: Lin He.*)

Jianming Xu, Yunfei Li, and Qian Shi are with the Guangdong Provincial Key Laboratory of Urbanization and Geosimulation, Center of Integrated Geographic Information Analysis, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: xujm27@mail2.sysu.edu.cn; liyf18213483@163.com; shixi5@mail.sysu.edu.cn).

Lin He is with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: helin@scut.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3274592

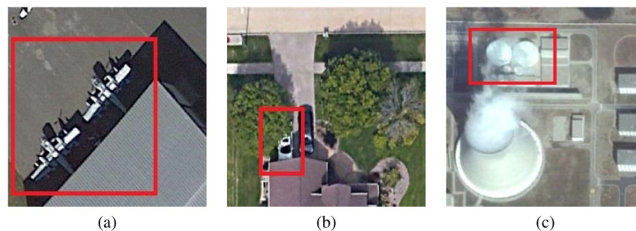


Fig. 1. (a)–(c) Example of occlusion on RS images.

However, existing methods still fall short in their performance on real-world RS images because the training sets they rely on typically contain all cases in the testing set [11]. In contrast, real-world RS images are unconstrained, and testing samples may be rare or unseen in the training set, making conventional methods ineffective [12]. For instance, recognizing targets that are partially occluded by uninterested objects (occlusion) represents a major challenge in interpreting unconstrained RS images [13]. Occlusion can cause significant CNN performance degradation [14], and its occurrence is often inevitable in RS due to buildings, trees, and mist cover, as shown in Fig. 1, which has attracted some attention in the RS community [15], especially among those who focus on low-altitude images, such as those collected by unmanned aerial vehicles (UAVs) [16], [17], [18], [19]. Despite significant efforts in this area, the exploration of occlusion in RS images is still limited.

Occlusion handling methods can be broadly categorized into two types: recovery-based and invariant representation-learning-based methods [20]. Recovery-based methods mimic the information completion process of the human neural system [21], which retrieves the missing information in the image or feature space [22], [23]. While occlusion recovery has achieved remarkable progress with the recent advances in generative models such as autoencoder and generative adversarial networks [24], [25], information reconstruction may not be necessary or computationally efficient for many RS tasks that require only discriminative models to predict object attributes (e.g., classification, detection, and segmentation). On the other hand, occlusion-invariant representation learning [26], [27], [28], [29] is preferable for tasks such as those mentioned above, which aim to learn a representation space that is insensitive to occlusion. One of the key challenges in learning occlusion-insensitive representation spaces is to alleviate occlusion-induced high intraclass similarity, low interclass separability, and poor representation discriminative [30]. Several methods have been implemented to

address this issue, including feature alignment techniques [31], low-rank constraints for sparse representation, and attention models to compel learning of nonoccluded representations [15], [32], [33], [34], [35]. For example, He et al. [32] employed low-rank and sparse constraints to the representation to account for occlusion in infrared target tracking. Wang et al. [33] expanded the receptive field and concentrated more on the unoccluded areas by utilizing dilated attention cross-stage partial modules to learn occlusion-invariant embeddings in tree recognition. Ren et al. [34] proposed a deformable faster-RCNN by aggregating multilayer features to extract unoccluded parts and enhance the CNN's generalization ability to partially occluded objects.

Current occlusion-invariant representation learning methods are mainly built on the close-world learning paradigm, which necessitates that testing cases are available in the training set. However, real-world occlusion is complex and encompasses diverse spatial and spectral patterns that make it almost impossible to accommodate all occlusion cases in the training set, rendering close-world learning inadequate. To address these challenges, contrastive learning [36] offers an intuitive solution. Contrastive learning is a representation learning approach that aims to produce closer representations between the anchor image and its augmented image. This definition suggests that occlusion invariance can be achieved by reducing the distance between the anchor image and its occlusion-augmented image. Contrastive learning is also semisupervised, which eliminates the need for comprehensive manual occlusion annotations in close-world learning. Finally, modern contrastive learning models have demonstrated compelling generalization ability [37], [38], [39], [40], which may enhance CNNs' ability to better characterize real-world occlusion that is rare or unseen in the training set.

Although contrastive learning is theoretically suitable for occlusion handling, existing contrastive learning methods are insufficient for learning occlusion-invariant representations on real-world RS images due to certain limitations. On the one hand, conventional contrastive learning disregards category information, which is unnecessary for an occlusion-level semisupervised problem. In addition, images with the same class in an input batch are treated as negative, increasing their representation heterogeneity. This may result in loose intraclass distribution or even worsen the recognition performance of occlusion samples. On the other hand, contrastive learning typically restricts the final layer of CNNs, while intermediate layers should also be occlusion-invariant. CNNs learn representations hierarchically, which implies that intermediate layers also contain high-level representation. Neglecting the occlusion-invariant constraint of intermediate layers could weaken the representation's invariance to occlusion.

To fill the gap between contrastive learning and RS occlusion-invariant scene classification, we propose the cascade supervised contrastive learning (Cascade-SupCon) network. Our approach introduces category information to enhance the discriminability of representations. By defining images with the same class as positive and others as negative, Cascade-SupCon reduces interclass similarity and eases issues of high intraclass diversity. Moreover, our cascade strategy ensures that occlusion-invariant

representations are learned hierarchically, with constraints applied to intermediate layers rather than only the final layer. The main contributions of our work are summarized as follows.

- 1) We propose a novel CNN method for occlusion-invariant scene classification based on an improved contrastive learning. Specifically, Our method incorporates category constraint and cascade strategy into SimCLR to achieve hierarchical occlusion-invariant representation learning, which enhance the CNN's classification ability in occlusion scenarios. Our proposed method is transplantable and has good out-of-distribution generalization ability, making it suitable for practical tasks such as object detection and segmentation, which rely on accurate characterization and classification of occluded objects. To the best of our knowledge, this is the first work to introduce contrastive learning into RS occlusion-invariant scene classification.
- 2) Our CNN operates in an occlusion-level semisupervised manner, enabling training without the need for occlusion annotations. This relieves the requirements of supervised models that rely on manual annotations.
- 3) We introduce several evaluation metrics for assessing the efficacy of occlusion-invariant representation learning. To enable a comprehensive evaluation of our CNN's occlusion handling capability, we use metrics that measure the spatial autocorrelation between occlusion and representation distribution, interclass separability, intraclass similarity, and representation discriminability.

The rest of this article is organized as follows. Section II describes the particulars of Cascade-SupCon. Section III presents the details of the datasets adopted in our experiments, whereas Section IV provides experimental results and their discussion. Finally, Section V concludes this article and offers a brief discussion of future works.

II. METHODOLOGY

Our Cascade-SupCon is an occlusion-invariant scene classification method based on an improved contrastive learning. As shown in Fig. 2, Cascade-SupCon consists of a pretext task module, a siamese representation learning subnetwork, a cascade supervised contrastive constraint, and a multilayer perceptron (MLP). The pretext task employs randomly positioned rectangle occlusion to simulate real-world occlusion, outputting original and occluded images for subsequent representation learning. The siamese subnetwork synchronously characterizes the original and occluded images. Then, the cascade supervised contrastive constraint compels the siamese subnetwork to learn occlusion-invariant representations. Finally, an MLP is introduced to classify the input scenes based on the learned representations and facilitates the quality evaluation of the learned representation. In the following, Cascade-SupCon is described in detail.

A. Pretext Task Module

In unconstrained environments, comprehensive occlusion annotations may be difficult due to the complexity of occlusion.

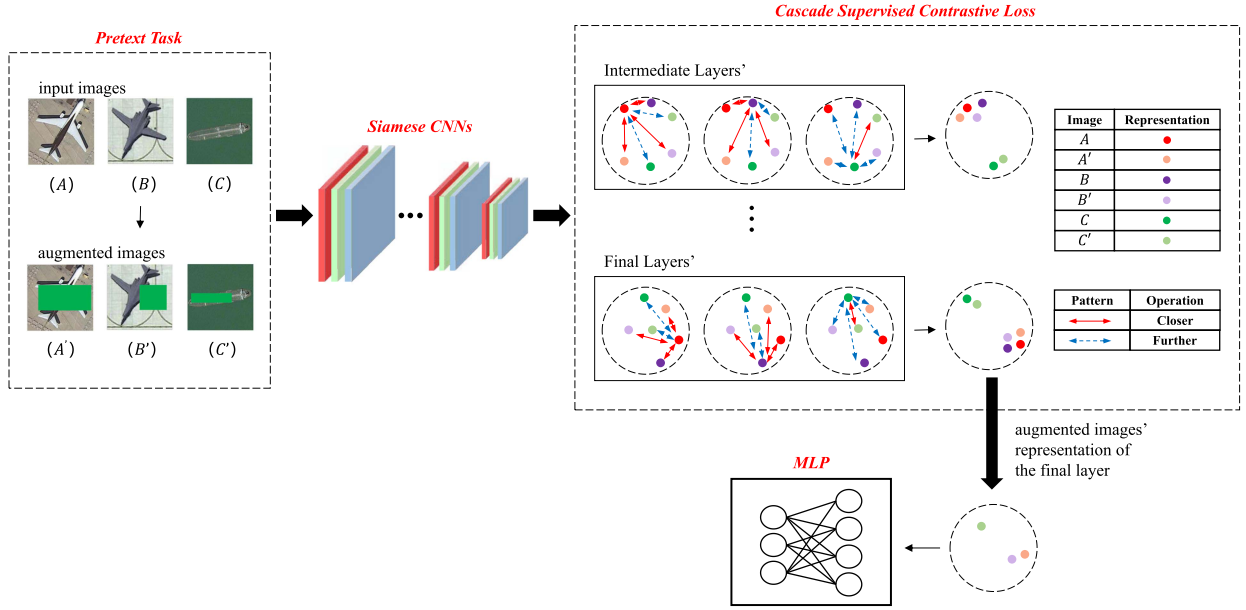


Fig. 2. Graphical illustration of training process of Cascade-SupCon.

This restrains the effectiveness of supervised models for occlusion handling, underscoring the need for the development of occlusion-level unsupervised/semisupervised methods. Here, we employ a training process involving pretext tasks [43] to form the occlusion-level semisupervised learning. This approach simplifies real-world tasks as specific augmentation, which facilitates the automatic generation of pseudo-labels and alleviates the need for occlusion annotations required in training models with real-world samples. As shown in Fig. 2, we simulate real-world occlusion by utilizing green rectangles of arbitrary size and location to occlude images, denoted as $Occ(\cdot)$. Regarding pseudo-label generation, traditional pretext tasks directly utilize data itself. However, such a strategy may be ineffective or even detrimental when applied to occlusion handling, as previously discussed in Section I. Therefore, we employ the category of images as pseudo-labels instead of what traditional approaches use. Then, the output of the pretext task module could be expressed as

$$\begin{aligned}
 A_1 &= \{(I_1, y_1), (I_2, y_2), \dots, (I_N, y_N)\} \\
 A_2 &= \{(Occ(I_1), y_1), (Occ(I_2), y_2), \dots, (Occ(I_N), y_N)\}
 \end{aligned} \tag{1}$$

where I_i ($i = 1, \dots, N$) denotes the i th image in a minibatch of N samples and y_i denotes the corresponding one-hot encoded label.

B. Siamese Subnetwork

We attain occlusion invariance by contrasting the representations of unoccluded and occluded views of a target. This is accomplished through our siamese subnetwork, which comprises a two-branch structure that learns the representations of two input sets (i.e., A_1 and A_2). More specifically, we utilize convolutional layers of ResNet-50 [44] to construct two branches

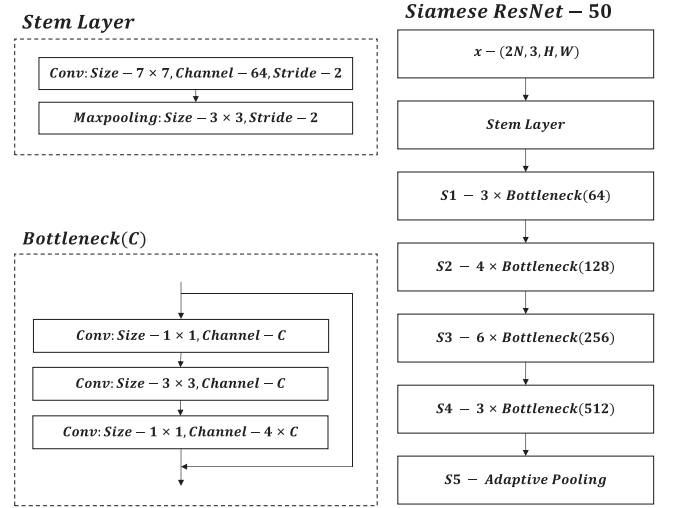


Fig. 3. Architecture of siamese ResNet-50.

of the siamese network (i.e., siamese ResNet-50). As shown in Fig. 3, the network is composed of stem layer, stage layers (i.e., S1, S2, S3, and S4), and an adaptive pooling layer (i.e., S5). The stem layer is constructed by $64 \ 7 \times 7$ convolution kernels and a 3×3 max-pooling layer. Each stage layer consists of several identical bottlenecks that share the same parameter C . For example, the S1 layer is a composition of three bottlenecks that all contain a shortcut connection, $C \ 1 \times 1$, $C \ 3 \times 3$, and $4C \ 1 \times 1$ convolution kernels.

In order to reduce the amount of parameters, two branches are weight-shared. The input $A_1 \cup A_2$ is processed by a stem layer, four stage layers, and an adaptive pooling layer in sequence. The S_i layer would generate a set of representations as

$$x_i = \{x_i^1, \dots, x_i^N, x_i^{N+1}, \dots, x_i^{2N}\} \tag{2}$$

where x_i^j ($j = 1, \dots, N$) indicates the representation of the j th original image, and x_i^{N+j} ($j = 1, \dots, N$) stands for the representation of the j th augmented image. The final output x of siamese ResNet-50 could be represented as

$$x = \{x_1, x_2, x_3, x_4, x_5\} \quad (3)$$

where x_i ($i = 1, 2, 3, 4, 5$) refers to the representation set of S_i layer.

C. Cascade Supervised Contrastive Constraint

After applying Siamese ResNet-50, we enhance the network's occlusion invariance by implementing a cascade supervised contrastive constraint. Contrastive constraint usually involves a similarity computation and a contrastive loss; given two representations of the same layer x_i^j and x_i^k , the similarity between them could be computed as

$$d(x_i^j, x_i^k) = \frac{x_i^{jT} x_i^k}{\|x_i^j\| \|x_i^k\|}. \quad (4)$$

The design of the loss function is crucial in achieving occlusion invariance. In the following, we first introduce the conventional contrastive learning loss function and its limitations in handling occlusion. Then, we explain the design of the cascade supervised contrastive loss function. Conventional contrastive learning is class-level semisupervised; therefore, only the anchor image and the augmented image are considered a positive pair. In addition, the majority of conventional contrastive loss functions usually operate on the final layer of a siamese neural network. Consequently, given a representation of the adaptive pooling layer x_5^j , the positive sample would be x_5^{N+j} . To learn a representation space where positive samples are closer and negative samples are further away, conventional losses [37], [39], [41], [42] could be expressed as

$$l_5^j = -\log \frac{\exp(d(x_5^j, x_5^{N+j})/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{j \neq k} \exp(d(x_5^j, x_5^k)/\tau)} \quad (5)$$

where $\mathbb{1}_{k \neq i}$ denotes an indicator function output 1 if $k \neq i$ and only if $k \neq i$ and τ controls the model sensitivity towards hard negative samples. When τ is higher, two negative but similar samples would cause greater loss. Minimizing such losses would decrease the representation distance between the original image and the corresponding occlusion augmented image, introducing pretext task invariance to the network. With the generalization ability of contrastive learning, the invariance of rectangle occlusion could even extend to real-world occlusion.

Although conventional contrastive losses introduce certain degree of occlusion invariance, such a configuration is not enough to handle occlusion. On the one hand, the manner of defining positive and negative sample pairs could be detrimental to occlusion handling. As shown in Fig. 4(a), the representation of the occluded image (e.g., the pink point) is located far away from their anchor image (e.g., the red point). By minimizing conventional losses, the similarity between the representation of the anchor and augmented images would increase, as shown in Fig. 4(b). Nonetheless, images within the same category are

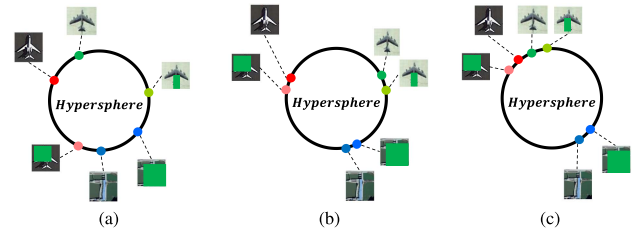


Fig. 4. Graphical illustration of the motivation to our work. The occlusion effects on representation space are described in (a). The optimization objectiveness of contrastive learning is described in (b). The optimization objectiveness of Cascade-SupCon is described in (c).

considered negative and the similarity of their representation would be decreased. For example, though the red green points both stand for airplane representation, their distance would even increase after the loss optimization. This positive sample definition could loosen the intraclass distribution and may worsen the high-intraclass diversity problem caused by occlusion. On the other hand, the constraint of conventional contrastive losses is inadequate. They only constrain the representation of the final layer to be occlusion invariant. Intermediate layers could also generate high semantic representation, which implies that they should also be occlusion invariant.

As discussed above, conventional contrastive losses may result in a loose intraclass distribution. In our approach, images belonging to the same class are considered as positive, while those belonging to different classes are considered as negative. Although this configuration requires category annotation, it still falls under the category of occlusion-level semisupervised learning. We improve the conventional contrastive loss as defined in (6) in terms of supervised contrastive learning (SupCon) [45], forming loss as

$$\begin{aligned} l_5^j &= \frac{1}{|P(x_5^j)|} \sum_{x_5^p \in P(x_5^j)} l_{i,p} \\ &= \frac{-1}{|P(x_5^j)|} \sum_{x_5^p \in P(x_5^j)} \log \frac{\exp(d(x_5^j, x_5^p)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(d(x_5^j, x_5^k)/\tau)} \end{aligned} \quad (6)$$

where $P(x_5^j)$ is a set of representations standing for the same category as x_5^j , and $|P(x_5^j)|$ is the cardinal number of $P(x_5^j)$. First, optimizing such loss would also introduce pretext task invariance because $x_5^{N+j} \in P(x_5^j)$. Moreover, the same category representation is defined as positive, which helps decreasing the intraclass diversity. The loss of the final layer takes the following form:

$$L_5^{\text{con}} = \frac{1}{2N} \sum_{i=1}^{2N} l_5^j. \quad (7)$$

In order to endow all high semantic representations with occlusion invariance, we apply contrastive loss to intermediate layers. Such a way significantly improves the performance of RS occlusion scene classification. Then, the Cascade-SupCon

loss L^{con} is formulated as

$$L^{\text{con}} = \sum_{i=1}^5 a_i L_i^{\text{con}} \quad (8)$$

where L_i^{con} refers to the contrastive loss of the i th layer, and $a_i \in \{0, 1\}$ decides the contribution of L_i^{con} .

D. Multilayer Perceptron

The final step of Cascade-SupCon is classification with MLP. It is an application of cascade supervised contrastive constraint and facilitates the evaluation of learned representation. As augmentation benefits classification, we take the augmented representation $x_5^{\text{aug}} = \{x_5^{N+1}, \dots, x_5^{2N}\}$ of the adaptive pooling layer as input. In order to simultaneously generate occlusion-invariant representations and produce accurate class predictions, we design the final loss as the sum of cascade supervised contrastive loss L^{con} and cross-entropy classification loss L^{cls} , which could be formulated as

$$L = L^{\text{con}} + L^{\text{cls}}$$

$$L^{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{i,j} \log(y_{i,j}') \quad (9)$$

where $y_{i,j} \in \{0, 1\}$ denotes that the i th image in the minibatch belongs to the j th class and $y_{i,j}' \in [0, 1]$ refers to the probability that the i th image in minibatch belongs to the j th class.

III. DATASETS

We utilize UAVDT [46] and simulation datasets to evaluate the performance of Cascade-SupCon. The UAVDT dataset is utilized to assess the network performance in real-world occlusion handling, while the simulated datasets are used for further network behavior analysis due to the lack of fine-grained occlusion annotations in the real-world data. In this section, we introduce the UAVDT and the simulation datasets separately.

A. UAVDT Dataset

UAVDT is developed to facilitate a range of UAV computer vision applications in unconstrained environments, including those with various types of occlusion, weather conditions, and camera viewpoints. To evaluate the performance of Cascade-SupCon, we transform the dataset into a scene classification format. Specifically, we train the model on a set of 15 000 unoccluded samples and test it on a set of 6429 unoccluded and 300 occluded samples. This enables us to assess the network's ability to classify real-world RS images with and without occlusion, even though it has not received specific training on occluded samples.

B. Simulated Datasets

Several RS image datasets include occluded samples. However, they do not annotate whether the objects are occluded or elaborately describe the occlusion status, such as occluder's sizes and types. This situation hinders the further analysis of the

network occlusion handling mechanism. The computer vision community usually uses simulated occlusion to aid network analysis because simulation methods can fully control occlusion status. The popular PASCAL 3D+ [47] and Occluded-COCO-Vehicles [14] are simulated datasets, yet they expedite the progress of occlusion handling massively. Therefore, we generate simulated datasets based on DIOR [48] and LEVIR [49], namely, DIOR-Occ and LEVIR-Occ, to support further analysis of the proposed method.

As shown in Fig. 5, we simulate occlusion by sticking various patterns as occluders. Since the real-world occlusion is complex, our simulation is designed to have the following characteristics.

- 1) The occlusion on an object is randomly simulated, where the position of the occlusion is not predetermined, and therefore, the occlusion is randomly spatially distributed. The randomness of simulated occlusion is an imitation of real-world occlusion, which has been adopted in most available occlusion datasets [14], [47]
- 2) The patterns' size is diverse, which mimics multiscale occlusion. The sizes of patterns are divided into four occlusion levels, namely, L0 (no occlusion), L1 (20–40% area of the object is occluded), L2 (40–60% area of the object is occluded), and L3 (60–80% area of the object is occluded).
- 3) The types of occluders are abundant, including rectangles, clouds, trees, and camouflages. Besides, the shapes and colors of patterns are diverse, even when those patterns are of the same type. For example, we introduce different shapes of clouds, such as sparse and dense clouds. The colors of clouds are also multifarious, like dark and white clouds.

In our experiments, we use 70% of the data as a training set and 30% as a test set. For DIOR-Occ, it has four types of objects (airplanes, ships, storage tanks, and windmills), with 17 635 samples in the training set and 7560 samples in the test set. LEVIR-Occ includes three types of objects (airplanes, ships, and storage tanks), with 7718 samples in the training set and 3310 samples in the test set. To assess the network's generalization ability on out-of-distribution occlusion samples, the training set of both DIOR-Occ and LEVIR-Occ only contains L0 data, while the test set includes four types of occlusion (rectangle, tree, cloud, and camouflage) and four levels of occlusion (L0, L1, L2, and L3).

IV. EXPERIMENTS

In this section, we present the results and discussion of experiments being conducted on the UAVDT and simulated datasets to showcase the effectiveness of the proposed Cascade-SupCon. The experiments on the UAVDT dataset aim to demonstrate the usefulness of the proposed method in real-world occlusion handling, while the experiments on the simulated datasets illustrate the behavior and mechanism of Cascade-SupCon. The rest of this section is organized as follows. Section IV-A provides the definition and intuitive meaning of the metrics adopted in our experimental result analysis. Section IV-B details the networks, parameter settings, and computation environment being used in

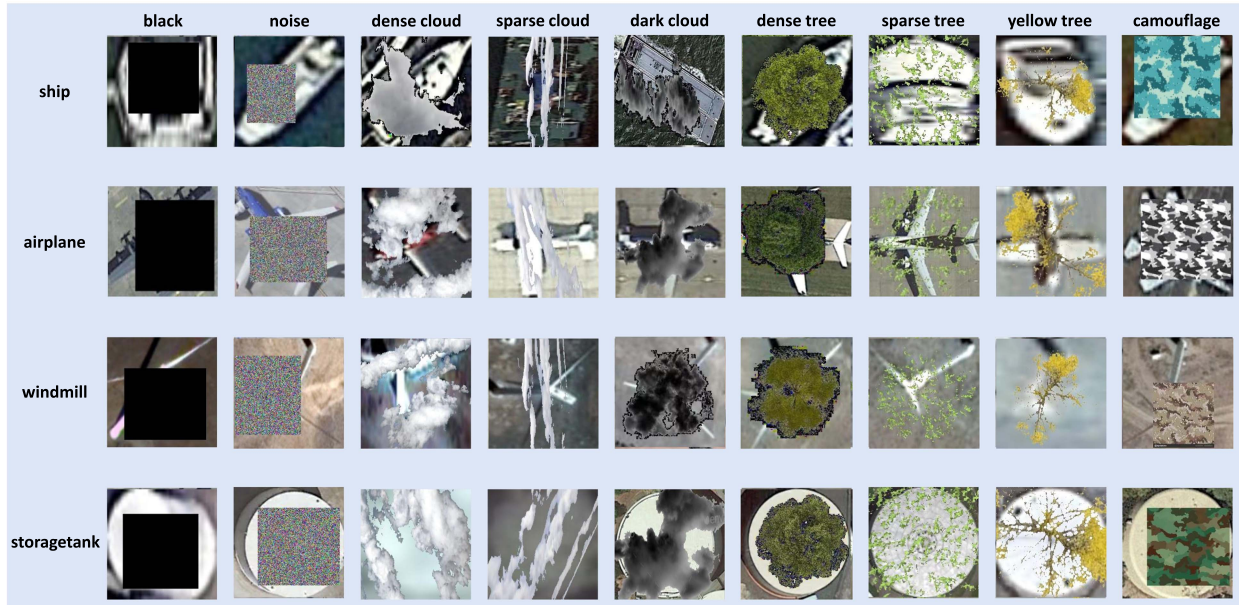


Fig. 5. Examples of different occlusion types and categories on the simulated DIOR-Occ dataset.

the experiments. In Section IV-C, we evaluate the classification performance on various datasets, while Sections IV-D–IV-F analyze the network behavior from intraclass, interclass, and cascade strategy angles, respectively.

A. Evaluation Metrics

This section introduces five quantitative evaluation metrics adopting for analyzing experimental results.

1) *Classification Accuracy*: The classification accuracy is used to assess the scene classification performance. In the experiment, we measure the classification accuracy of different categories and different occlusion levels to analyze the occlusion robustness of the model. Meanwhile, we comprehensively measure the model's performance using the overall classification accuracy.

2) *Global Moran Index (GMI)*: The GMI [50] initially measures the spatial autocorrelation between geographic entities. In our experiments, we utilize GMI to measure the dependence of intraclass representation distribution on occlusion levels and types. The value of GMI ranges from -1 to 1 . When the GMI approaches 0 , the data distribution does not rely on occlusion status and distributes randomly. When the GMI approaches 1 , the data distribution shows positive autocorrelation, meaning that the objects with similar occlusion status gather together. On the contrary, the data distribution shows negative autocorrelation when the GMI nears -1 , representing that data with similar occlusion status repel each other. In short, the network that produces a GMI closer to 0 can be recognized as more occlusion invariant because it generates representations that are less affected by occlusion.

3) *Trace of Interclass Covariance Matrix (TSB)*: TSB refers to differences between interclass representations. The interclass

covariance matrix is the covariance matrix of each class's mean vectors (representations), and TSB is the trace of it. When TSB is larger, the difference between interclass representations is more significant, which facilitates learning robust discrimination functions.

4) *Trace of Intraclass Covariance Matrix (TSW)*: TSW measures the differences of the intraclass representations. The intraclass covariance matrix of a category is defined as the covariance matrix of such class's representations. TSW is defined as the prior probability-weighted sum of the intraclass covariance matrix of each class. The smaller the TSW, the more compact the intraclass distribution, reducing the chance of misclassifications.

5) *Geometric-Based Separability Criterion*: It is inadequate to measure representation separability using TSB and TSW alone because separability depends simultaneously on inter- and intraclass distribution. We use the geometric separability criterion (J) to measure the overall separability of the representations. J is defined as TSB/TSW , which can comprehensively consider intraclass differences and interclass differences. The larger the J , the more significant the interclass difference, and the more compact the intraclass distribution; consequently, the better the separability.

B. Experimental Setups

We conduct experiments using six occlusion level unsupervised/semisupervised networks, namely, ResNet-50, ResNet-50 with occlusion augmentation, MoCo, SimCLR, SupCon, and Cascade-SupCon. ResNet-50 experiments demonstrate that conventional CNNs are not robust to RS occlusion and reveal the effects of occlusion on representations learned by CNNs. To conduct ablation experiments, we use ResNet-50 with occlusion augmentation since contrastive learning methods (i.e., SimCLR,

TABLE I
UAVDT CLASSIFICATION ACCURACIES

	ResNet-50	ResNet-50 + aug.	MoCo	SimCLR	SupCon	Cascade-SupCon
Unoccluded	0.990	0.983	0.826	0.878	0.986	0.981
Occluded	0.506	0.576	0.579	0.583	0.650	0.670

The bolded entities indicates the best indices.

MoCo, SupCon, and Cascade-SupCon) rely on data augmentation. We analyze the SimCLR and MoCo classification performance and network behavior to decide our baseline contrastive learning model and identify the gap between contrastive learning and occlusion handling. Furthermore, we reference SupCon to add category constraints to contrastive learning, which alleviates the loose intraclass distribution problem of contrastive learning in occlusion handling. Finally, the experiments of Cascade-SupCon further certify the effectiveness of the proposed method, specifically category constraints and cascade strategies.

The general network setups are as follows: input images are resized to 224×224 , and all networks are trained for 20 epochs using SGD as the optimizer, with a learning rate of 0.001 and a momentum of 0.9. We set the learning rate empirically without any specific considerations. Setting the learning rate to 0.001 has been used in many existing works, such as [51], [52], and [53] and deep learning API [54]. ExponentialLR is used with the parameter γ set to 0.6 to decrease the learning rate after each epoch. The experiments are performed on an NVIDIA RTX-2070 Super with 8-GB RAM, and the batch size is set to 4. For Cascade-SupCon, we set the temperature parameter τ to 0.07, which is popular in contrastive learning. Regarding the cascade contributions a_i , we set them to 1 when $i = 4, 5$ and 0 when $i = 1, 2, 3, 4$. This choice is based on the consideration that low-level representations are not necessarily occlusion invariant.

C. Experiment 1—Classification Performance

We conduct separate scene classification experiments on the UAVDT and simulation datasets. The following sections present and discuss the respective experimental results. The UAVDT experiments confirm that the proposed Cascade-SupCon approach is capable of satisfactorily handling real occlusion. Furthermore, the simulation dataset experiments provide additional evidence of the advantages of the Cascade-SupCon method.

1) *UAVDT Classification Accuracy Comparison*: Table I showcases the classification accuracies on the UAVDT dataset. The first column demonstrates that ResNet-50 can accurately classify unoccluded scenes; however, the accuracy drops by 48.4% for occluded scenes. This lack of robustness highlights the need for developing occlusion-invariant methods for applying CNNs to unconstrained RS images. SimCLR shows a 7.7% improvement in accuracy for occluded scenes compared with ResNet-50. However, the 11.2% drop in accuracy for unoccluded scenes indicates that it is insufficient for occlusion handling when compared with ResNet-50. Regarding MoCo, the occlusion classification accuracy is unsatisfactory, and the unoccluded classification accuracy even shows severe degradation. This highlights a discrepancy between the potential of trendy contrastive learning for learning occlusion-invariant representation

and its current limitations in handling occlusion effectively. We evaluate the performance of SupCon and find that it can classify both unoccluded and occluded scenes accurately, indirectly indicating that the gap between contrastive learning and occlusion handling stems from the lack of category constraints. Finally, we present the classification performance of Cascade-SupCon in the sixth column. Cascade-SupCon improves the occlusion classification performance by 16.4% compared with ResNet-50 while retaining the ability to classify unoccluded scenes. These results demonstrate that the proposed Cascade-SupCon approach is suitable for real RS applications in occlusion environments.

2) *Simulation Dataset Classification Accuracy Comparison*: We present the classification accuracies on simulation datasets in Table II. In two simulation datasets, the average accuracies perform similarly to the UAVDT experiments. The effects of occlusion on CNNs' classification accuracy are more specific in this result. We can see that ResNet-50 can classify L0 data well. However, the accuracy drops sharply as the occlusion degree increases. In DIOR-Occ, the classification accuracy degradation even reaches 58.1% when the objects are severely occluded. It demonstrates that applying CNNs in occlusion RS images is unreliable, significantly when the occlusion degrees are high. The comparison between SimCLR and MoCo reveals that SimCLR performs better in all occlusion levels. Therefore, we take SimCLR as the baseline contrastive learning model in the following experiments. Comparing the results of SupCon and Cascade-SupCon, we can conclude that the cascade strategy improves the classification ability in every occlusion level. The advantages of the cascade strategy are more prominent when the occlusion degree is higher.

D. Experiment 2—Intraclass Analysis

1) *Occlusion Effects on Intraclass Distribution*: This section explores the impact of occlusion on the intraclass representation distribution by comparing the differences in distribution between unoccluded and occluded data. We begin by obtaining representations of unoccluded objects. Specifically, we extract image patches of unoccluded objects from the original DIOR and LEVIR datasets and apply ResNet-50 for representation learning. We then use TSNE [55] for dimensionality reduction to obtain 2-D representations, which facilitate visualization and intuitive analysis. Fig. 6(a) displays the representation distribution of unoccluded storage tanks in the DIOR dataset, which mainly form four clusters.

We investigate the impact of occlusion degrees on the intraclass representations distribution by conducting an analysis on image patches that are occluded by trees. We employ ResNet-50 and TSNE to learn representation and reduce dimensionality. We assign 0.25 to represent L0, 0.5 to represent L1, 0.75 to represent L2, and 1 to represent L3 and use the GMI to measure the existence of positive spatial autocorrelation in the occlusion levels. The results, shown in Table III, indicate that the GMI of both the DIOR and LEVIR datasets, after being occluded by trees, is significantly greater than 0. This suggests that the occlusion levels have an impact on the distribution of representations, and objects with similar occlusion levels tend to

TABLE II
CLASSIFICATION ACCURACIES ON SIMULATED DATASETS

	DIOR-Occ					LEVIR-Occ				
	L0	L1	L2	L3	Overall	L0	L1	L2	L3	Overall
ResNet-50	100	89.03	64.97	41.90	73.97	99.34	85.41	62.65	49.57	74.24
ResNet-50 + aug.	99.51	94.02	77.51	59.51	82.64	96.94	92.09	78.10	61.37	82.13
MoCo	87.81	90.27	76.94	72.11	81.78	92.13	85.55	75.27	70.70	80.90
SimCLR	98.78	92.48	85.52	73.38	87.54	97.98	92.94	81.89	71.73	86.13
SupCon	98.86	95.94	91.91	80.03	91.68	99.73	98.29	90.36	77.47	91.46
Cascade-SupCon	99.84	99.08	94.50	83.16	94.14	99.95	98.34	94.14	81.28	93.43

The bolded entities indicates the best indices.

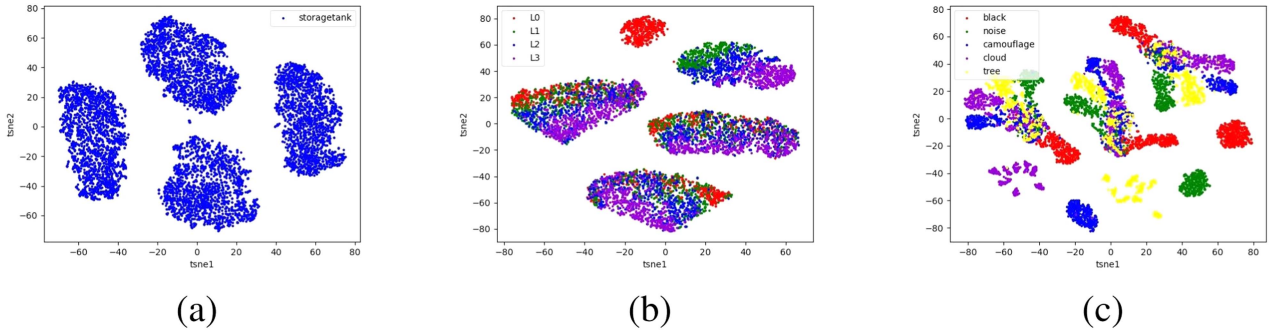


Fig. 6. Effects of occlusion levels and occlusion types to representation distribution. (a) Scatterplot after ResNet-50 extracts the unoccluded storage tank representations in the DIOR dataset after TSNE dimensionality reduction. (b) After using trees with different occlusion levels to occlude the storage tank scene in the DIOR dataset, we show the scatterplots of TSNE dimensionality reduction representations extracted by ResNet-50. Specifically, we use red for L0 level data, green for L1 level data, blue for L2 level data, and purple for L3 level data. (c) After the storage tank scene in the DIOR dataset is occluded using five types of occluder at the L3 level, we show the scatterplots of TSNE dimensionality reduction representations extracted by ResNet-50. Specifically, we use red for black rectangle occlusion data, green for camouflage occlusion data, blue for cloud occlusion data, and purple for noise rectangle occlusion data.

TABLE III
GMI OF DIOR AND LEVIR OCCLUDED BY TREE DIFFERENT OCCLUSION LEVELS

	airplane	ship	storage tank	windmill	average
DIOR(Occluded)	0.704	0.542	0.715	0.667	0.657
LEVIR(Occluded)	0.654	0.383	0.622	—	0.553

cluster together. In addition, Fig. 6(b) shows the representation distribution of occluded storage tank image patches in the DIOR dataset, where the representations are grouped into five clusters and the spatial distribution of representation points significantly correlates with the occlusion level. This finding suggests that the occlusion levels amplify the differences in the representations of samples from the same category.

We explore the influence of occlusion types on representations by using black rectangles, noise rectangles, camouflages, clouds, and trees with specific occlusion levels to occlude original image patches. We let 0.2 represent black occlusion, 0.4 represent noise occlusion, 0.6 represent camouflage occlusion, 0.8 represent cloud occlusion, and 1 represent tree occlusion. ResNet-50 and

TABLE IV
GMI OF DIOR AND LEVIR OCCLUDED BY DIFFERENT OCCLUSION TYPES

	DIOR			LEVIR		
	L1	L2	L3	L1	L2	L3
airplane	0.373	0.660	0.873	0.426	0.650	0.820
ship	0.301	0.587	0.849	0.283	0.495	0.711
storage tank	0.367	0.619	0.882	0.402	0.557	0.782
windmill	0.382	0.661	0.86	—	—	—
average	0.356	0.632	0.866	0.370	0.567	0.771

TSNE are used to obtain 2-D representations. Then, we use GMI to measure whether there is spatial autocorrelation between representations of occlusion types. As shown in Table IV, when the occlusion level is L1, the mean values of the GMI of the DIOR and LEVIR datasets are 0.356 and 0.370, respectively, which are significantly greater than 0. This shows that even when the degree of occlusion is relatively small, the representation

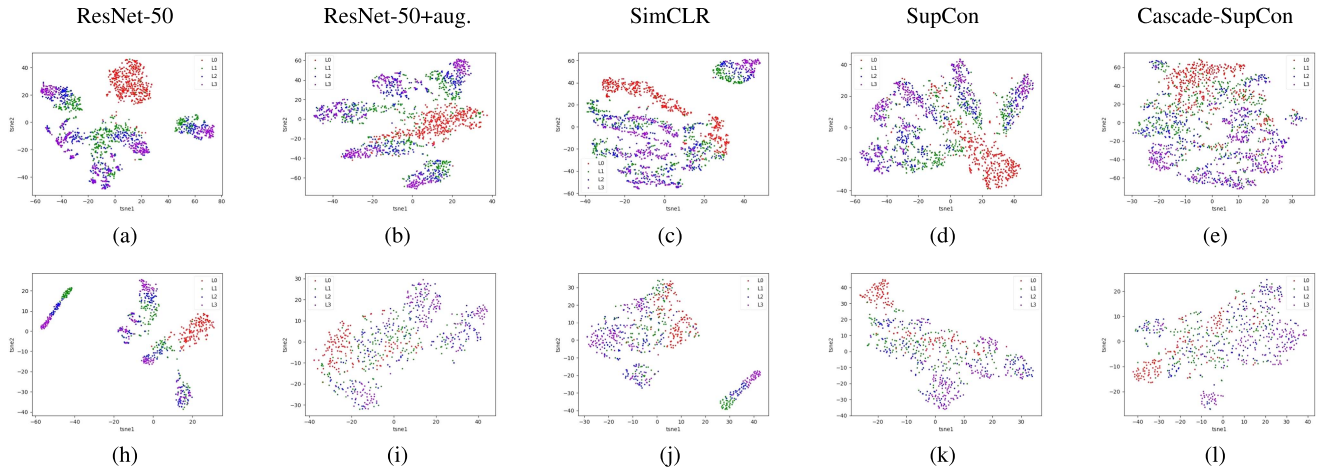


Fig. 7. (a)–(l) TSNE dimensionality reduction scatterplot after extracting representations from storage tank images using each model. Use red for L0 level data, green for L1 level data, blue for L2 level data, and purple for L3 level data. (a)–(e) are visualizations of the DIOR-Occ dataset. (h)–(l) are visualizations of the LEVIR-Occ dataset.

distribution still has a significant positive spatial autocorrelation of occlusion types. That is, representations with similar occlusion types tend to cluster together. When the occlusion level is L2, the mean GMI values of the two datasets reach 0.632 and 0.567. When the occlusion level is L3, the mean values of GMI of the two datasets reach 0.866 and 0.771, respectively. This shows that with the increase of occlusion levels, the influence of occlusion type on the representation distribution increases sharply. Especially, for the DIOR dataset, the GMI even reaches 0.866, which indicates that the representation is almost distributed according to the occlusion types. Fig. 6(c) shows the representation distribution of storage tank images with L3 occlusion. It can be seen that the representations form clusters by occlusion types, and the representation distribution is severely fragmented. This shows that occlusion types cause significant differences in terms of intraclass representations, and the richer the type of occlusion, the more significant the difference.

In summary, the distribution of representations exhibits positive spatial autocorrelation on both the occlusion level and the occlusion type. The occlusion level and the occlusion type affect the spatial distribution of representations, and representations with similar occlusion levels and occlusion types form clusters. This phenomenon leads to the fragmentation of intraclass representation distribution, resulting in increased intraclass variability. In order to reduce the impact of occlusion, it is necessary to reduce the intraclass variability caused by occlusion level and occlusion type.

2) *Network Intraclass Distribution Comparison*: As shown in Fig. 7, TSNE is used to reduce the dimension of the storage tank scenes in the two datasets, and the representations are colored according to the occlusion level. In ResNet-50, the scattered points are severely fragmented, and the representations form clusters according to the occlusion level. As shown in Fig. 7(a)–(d), pretext-task-augmented ResNet-50, SimCLR, and SupCon effectively reduce the distance between clusters for the DIOR-Occ dataset, the clusters are still obvious, and there is a significant positive spatial autocorrelation. As shown in Fig. 7(e), for the DIOR-Occ dataset, Cascade-SupCon not only

effectively reduces the distance of each cluster but also mixes representations of different occlusion levels more evenly, effectively reducing the positive spatial autocorrelation. As shown in Fig. 7(h)–(l), for the LEVIR-Occ dataset, compared with ResNet-50, pretext-task-augmented ResNet-50, SimCLR, and SupCon, Cascade-SupCon could better reduce the clustering distance and reduce the spatial positive autocorrelation. As shown in Table V, it can be seen that the GMI of ResNet-50 in the two datasets is 0.839 and 0.805, respectively, showing a strong positive spatial autocorrelation. The GMI of Cascade-SupCon in the two datasets is 0.676 and 0.638, respectively, and the positive spatial autocorrelation is the lowest, which can best eliminate the influence of the occlusion level on the distribution of representation classes.

In Fig. 8, TSNE is used to reduce the dimensionality of the storage tanks scenes in both datasets for visualization purposes. The resulting plot is colored according to the type of occlusion. The plot in Fig. 8(a) and (h) shows that ResNet-50 produces clusters formed by occlusion types, and the representation distribution is fragmented. Pretext-task-augmented ResNet-50 and SimCLR reduce the distance between clusters and spatial autocorrelation to varying degrees, as shown in Fig. 8(b), (c) and (i), (j). However, the results are unsatisfactory. SupCon reduces the distance between clusters, but the positive spatial autocorrelation still persists, as illustrated in Fig. 8(d) and (e). On the other hand, Cascade-SupCon mixes the representations evenly, effectively reducing the positive spatial autocorrelation, as shown in Fig. 8(f) and (g). Both SupCon and Cascade-SupCon effectively reduce the positive spatial autocorrelation for the LEVIR-Occ dataset, as depicted in Fig. 8(k) and (l). The GMI of Cascade-SupCon is 0.701 on the DIOR-Occ dataset, which most effectively eliminates the influence of occlusion types, as indicated in Table VI. On the other hand, for the LEVIR-Occ dataset, the GMI of SupCon and Cascade-SupCon is 0.622 and 0.627, respectively, indicating that they almost equally eliminate the effect of occlusion type.

Based on the aforementioned analysis, it can be concluded that among the five unsupervised occlusion scene classification

TABLE V
GMI FOR OCCLUSION LEVELS OF REPRESENTATIONS EXTRACTED ON DIOR-OCC AND LEVIR-OCC TEST SETS

	DIOR-Occ					LEVIR-Occ			
	airplane	ship	storage tank	windmill	average	airplane	ship	storage tank	average
ResNet-50	0.856	0.792	0.862	0.848	0.839	0.843	0.709	0.864	0.805
ResNet-50 + aug.	0.830	0.735	0.791	0.837	0.798	0.715	0.574	0.670	0.653
SimCLR	0.713	0.600	0.786	0.829	0.732	0.627	0.619	0.705	0.650
SupCon	0.807	0.565	0.761	0.756	0.722	0.758	0.550	0.669	0.659
Cascade-SupCon	0.775	0.483	0.708	0.738	0.676	0.780	0.600	0.535	0.638

The bolded entities indicates the best indices.

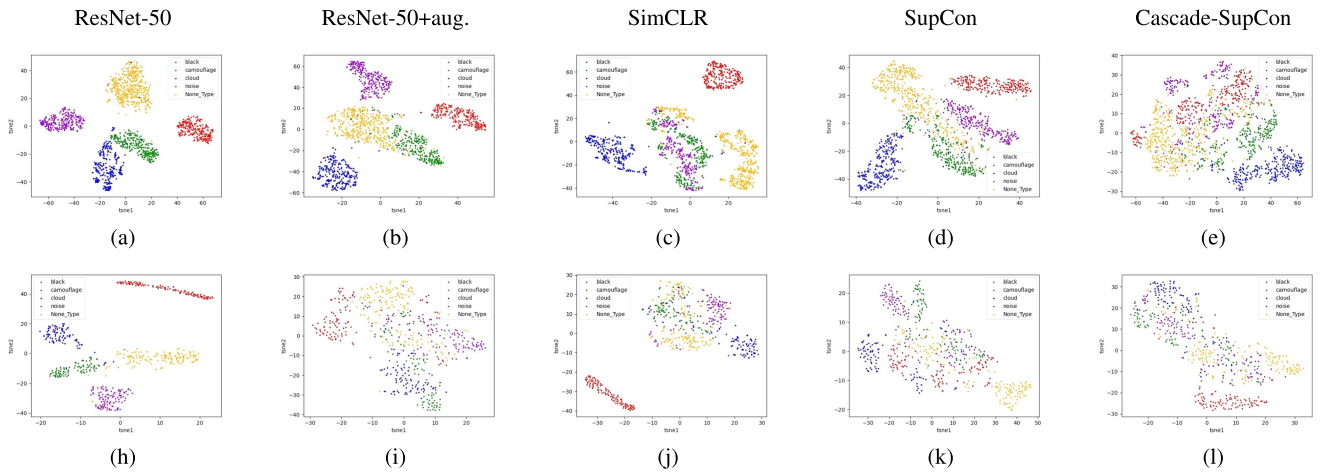


Fig. 8. (a)–(l) TSNE dimensionality reduction scatterplot after extracting representations from storage tank images using each model. We use red for black rectangle occlusion data, green for camouflage occlusion data, blue for cloud occlusion data, purple for noise rectangle occlusion data, and yellow for unoccluded data. (a)–(e) are visualizations of the DIOR-Occ dataset. (h)–(l) are visualizations of the LEVIR-Occ dataset.

TABLE VI
GMI FOR OCCLUSION TYPES OF REPRESENTATIONS EXTRACTED ON THE DIOR-OCC AND LEVIR-OCC TEST SETS

	DIOR-Occ					LEVIR-Occ			
	airplane	ship	storage tank	windmill	average	airplane	ship	storage tank	average
ResNet-50	0.964	0.914	0.948	0.962	0.947	0.952	0.860	0.959	0.923
ResNet-50 + aug.	0.957	0.778	0.933	0.984	0.913	0.670	0.756	0.577	0.667
SimCLR	0.838	0.717	0.874	0.883	0.828	0.660	0.770	0.711	0.713
SupCon	0.858	0.732	0.952	0.921	0.865	0.647	0.586	0.634	0.622
Cascade-SupCon	0.737	0.529	0.792	0.748	0.701	0.722	0.638	0.522	0.627

The bolded entities indicates the best indices.

algorithms, Cascade-SupCon is the most effective in reducing the positive correlation between representations and occlusion. It achieves this by effectively reducing the distance between clusters and minimizing the differences in representation distribution. Therefore, Cascade-SupCon can effectively mitigate the impact of occlusion on the distribution of intraclass representations.

E. Experiment 3—Interclass Analysis

1) *Occlusion Effects on Interclass Distribution:* We investigate the impact of occlusion on interclass distribution by comparing the differences in representations between unoccluded and occluded images. We perform representation learning using ResNet-50 on both the original DIOR and LEVIR datasets, as

TABLE VII
SEPARABILITY INDEX COMPARISON BETWEEN OCCLUDED AND UNOCCLUDED DIOR AND LEVIR DATASET

	DIOR			LEVIR		
	TSB	TSW	J	TSB	TSW	J
original	51.072	34.998	1.460	75.481	245.383	0.307
Occ	18.671	55.127	0.338	38.311	2612.051	0.015

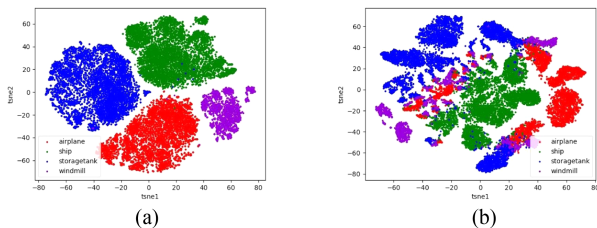


Fig. 9. Effects of occlusion on interclass representations. (a) TSNE dimensionality reduction scatterplot after using ResNet-50 to extract representations from the unoccluded DIOR dataset. (b) TSNE dimensionality reduction scatterplot showing the use of ResNet-50 to extract representations of the DIOR dataset after occlusion. Both of them use red for planes, green for boats, blue for storage tank, and purple for windmills.

well as their respective occluded versions. Subsequently, we compute the TSB, TSW, and J, as shown in Table VII. In the DIOR dataset, the unoccluded TSB of 51.072 and TSW of 34.998 indicate significant interclass distance and a compact intraclass distribution, with a J value of 1.460, indicating good separability. However, upon the addition of occlusion, TSB drops to 18.671, TSW increases to 55.127, and J decreases sharply to 0.338, indicating significantly reduced class separability due to the decrease in interclass distance and the loosening of intraclass distribution. A similar trend is observed in the LEVIR dataset. The scatterplots in Fig. 9 of the original DIOR and DIOR-Occ datasets reveal that the overlap between different classes is minimal prior to occlusion, with a relatively concentrated distribution of intraclass representations. However, after the introduction of occlusion, the overlap increases considerably, the distribution of intraclass representations becomes very loose, and the separability decreases significantly. These findings demonstrate that occlusion can decrease interclass distance and loosen intraclass distribution, leading to misclassification.

2) *Network Interclass Distribution Comparison*: Five algorithms are used for representation learning on the test sets of DIOR-Occ and LEVIR-Occ, and TSB, TSW, and J are calculated as shown in Table VIII. TSB values of ResNet-50 on both datasets are 18.522 and 30.477, respectively, indicating that the distances between various representations are small in the representations learned by ResNet-50. However, its TSW values are 55.043 and 2517.045, respectively, indicating a very loose intraclass distribution that could result in serious misclassification due to overlapping representations among various categories. Consequently, its J value is only 0.336 and 0.012, indicating poor separability. Augmentation significantly improves the intraclass distribution while maintaining a similar distance between classes, leading to a TSB of 20.430 in the DIOR-Occ dataset

TABLE VIII
SEPARABILITY INDEX OF REPRESENTATIONS EXTRACTED IN DIOR-OCC AND LEVIR-OCC TEST SETS

	DIOR-Occ			LEVIR-Occ		
	TSB	TSW	J	TSB	TSW	J
ResNet-50	18.522	55.043	0.336	30.477	2517.045	0.012
ResNet-50 + aug.	20.430	57.611	0.373	29.539	79.152	0.355
SimCLR	71.017	642.189	0.111	70.978	2146.392	0.033
SupCon	84.277	72.901	1.156	70.665	71.568	0.987
Cascade-SupCon	73.241	27.636	2.650	52.950	25.368	2.087

The bolded entities indicates the best indices.

and a TSW of 57.611, which is better than ResNet-50. Similarly, in the LEVIR-Occ dataset, the TSB remains similar, while the TSW is greatly improved, demonstrating the effectiveness of augmentation in compacting the intraclass distribution. SimCLR shows substantial improvement over ResNet-50 and agent task enhancement with TSB values of 71.017 and 70.978 in the DIOR and LEVIR datasets, respectively. However, SimCLR performs poorly on TSW, with the DIOR data's TSW even expanding to 642.189 and a very loose intraclass distribution. SupCon has the highest TSB on both datasets and effectively expands the interclass distance. However, the TSW is enlarged in the DIOR dataset, and on the LEVIR-Occ dataset, the TSW is even larger than the augmentation strategy. This implies that SupCon can expand the interclass distance but is inadequate in reducing the intraclass distribution. J values of SupCon on the two datasets are 1.156 and 0.987, respectively, indicating a significant improvement in representation separability. Cascade-SupCon has a significantly more compact intraclass distribution than other models, with TSW values of 27.636 and 25.368 on the two datasets. TSB values of Cascade-SupCon on the two datasets are 73.241 and 52.950, respectively, which are lower than SupCon. This is because Cascade-SupCon employs a stringent intraclass distribution constraint, which results in a compact intraclass representation distribution. This means that even with a lower interclass distance, TSB converges to a lower value without causing extra loss. Although Cascade-SupCon's TSB values are not the best, its J values of 2.650 and 2.087 on the two datasets are two times larger than the J of SupCon, indicating that it learns the representations with the highest separability among the comparison methods.

Fig. 10(a) and (h) indicates that ResNet exhibits heavy overlap between representations of various categories, resulting in poor separability with close distance between classes. Augmented ResNet-50, depicted in Fig. 10(b) and (j), effectively reduces the intraclass distance, but still exhibits severe overlap between categories and poor separability. Conversely, Fig. 10(c) and (k) demonstrates that SimCLR increases interclass differences but suffers from a loose intraclass distribution, with lower values of J. Despite this, SimCLR still performs better than augmented ResNet-50. SupCon, as shown in Fig. 10(d) and (i), demonstrates good separability and significant differences between classes, but the intraclass distribution remains relatively loose. Finally, Fig. 10(e) and (k) demonstrates that Cascade-SupCon

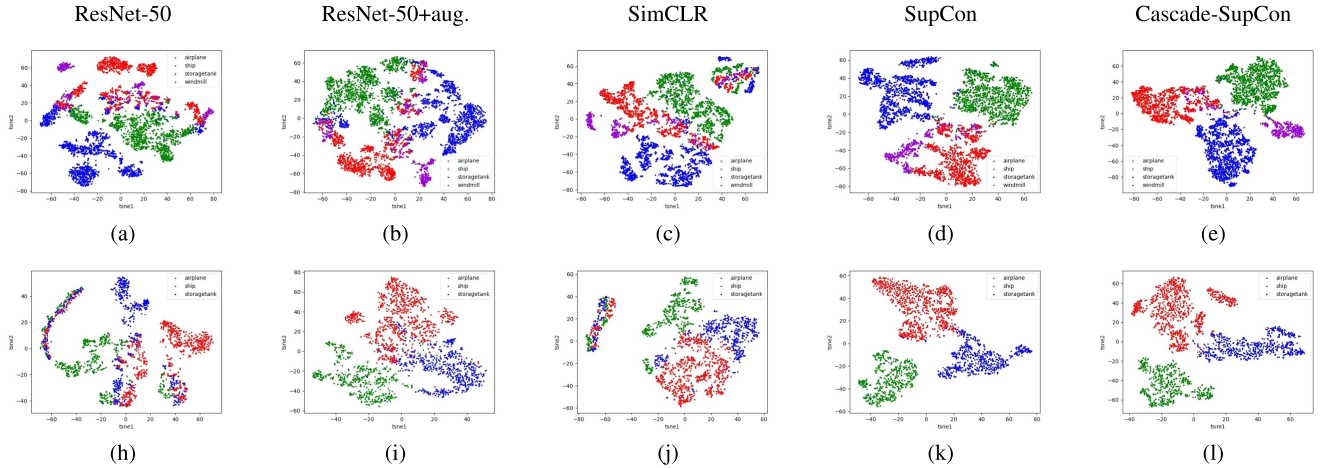


Fig. 10. (a)–(l) TSNE dimensionality reduction scatterplot after using each model to extract the representations of the datasets. We use red for airplanes, green for ships, blue for storage tanks, and purple for windmills. (a)–(e) are visualizations of the DIOR-Occ dataset. (h)–(l) are visualizations of the LEVIR-Occ dataset.

TABLE IX
OCCLUSION SCENE CLASSIFICATION ACCURACY OF SUPCON WITH DIFFERENT CASCADE LAYERS ON DIOR-OCC AND LEVIR-OCC TEST SETS

Cascaded Layers	DIOR-Occ					LEVIR-Occ				
	L0	L1	L2	L3	Overall	L0	L1	L2	L3	Overall
Pooling Layer (SupCon)	98.86	95.94	91.91	80.03	91.68	99.73	98.29	90.36	77.47	91.46
Pooling Layer + Stage 4 (Cascade-SupCon)	99.84	99.08	94.50	83.16	94.14	99.95	98.34	94.14	81.28	93.43
Pooling Layer + Stage 4-3	97.96	97.78	90.94	78.15	91.21	98.31	95.40	86.68	73.02	88.35
Pooling Layer + Stage 4-2	99.59	97.01	93.20	80.19	92.50	99.84	97.17	87.70	70.76	88.86
Pooling Layer + Stage 4-1	99.67	96.70	92.48	82.77	92.91	99.07	97.01	90.46	78.49	91.26

The bolded entities indicates the best indices.

has the best separability, displaying a very compact intra-class distribution and minimal confusion between categories, ultimately leading to more accurate classifications. We conclude that Cascade-SupCon provides an optimal solution to the problems of small interclass distance, loose intraclass distribution, and severe occlusion-induced misclassification.

F. Experiment 4—Cascade Strategy Analysis

To investigate the impact of the number of cascade layers on occlusion scene classification, we present the classification accuracy on simulated datasets based on the number of cascade layers in Table IX. While SupCon only constrains the pooling layer and is unable to accurately classify highly occluded scenes, Cascade-SupCon employs SupCon on both the pooling layer and Stage-4 layers, resulting in more comprehensive semantics constraints. Compared to SupCon, Cascade-SupCon effectively enhances the classification accuracy for each occlusion level. However, adopting more cascaded layers does not always yield better results. The introduction of shallow representations leads to a decrease in the classification accuracy of occluded scenes to varying degrees. For instance, when using the DIOR-Occ dataset, the introduction of Stage-3 representations results in a classification accuracy of 91.21%, which is lower than the

91.68% achieved by SupCon. In the case of the LEVIR-Occ dataset, the situation is even more pronounced when considering the additional introduction of Stage-3, Stage-2, and Stage-1 representations. The classification accuracy is significantly lower than that of Cascade-SupCon and even lower than that of SupCon. This is because the shallow layers of CNNs extract detailed representations, and it is unreasonable to assume that the details of occluded and unoccluded images would remain unchanged. Based on these experimental results, we propose to limit the constraints to the pooling layer and Stage-4 in Cascade-SupCon. In future studies, we aim to develop an occlusion scene classification model that is insensitive to the number of cascaded layers.

V. CONCLUSION

In this article, we proposed a novel scene classification network capable of working in partially occluded scenarios, called Cascade-SupCon. The network is based on contrastive learning, which is used to classify occluded RS images for the first time. By employing contrastive learning, interclass similarity induced by occlusion decreases without the need for occlusion annotation. In addition, the method can increase intraclass similarity by introducing category information. Furthermore, it generates

occlusion-invariant representation hierarchically using a cascade strategy, which improves its occlusion handling capability. Our network is evaluated on UAVDT and two simulation datasets, and the results demonstrate that it can adapt to severe and unseen occlusion. In our work, we also introduced additional evaluation metrics to comprehensively assess occlusion scene classification performance. Experimental results suggest the excellent performance of our method. In future work, this method could be extended to object detection and semantic segmentation applications.

REFERENCES

- [1] W. Han, A. Kuerban, Y. Yang, Z. Huang, B. Liu, and J. Gao, "Multi-vision network for accurate and real-time small object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 6001205.
- [2] C. Zhang, C. Yang, K. Cheng, N. Guan, H. Dong, and B. Deng, "MSIF: Multisize inference fusion-based false alarm elimination for ship detection in large-scale SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224811.
- [3] H. Chen et al., "Stacked spectral feature space patch: An advanced spectral representation for precise crop classification based on convolutional neural network," *Crop J.*, vol. 10, pp. 1460–1469, 2022.
- [4] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 117, pp. 11–28, 2016.
- [5] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [6] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5614914.
- [7] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4062–4076, Jun. 2019.
- [8] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
- [9] J. Kang, R. Fernandez-Beltran, Z. Wang, X. Sun, J. Ni, and A. Plaza, "Rotation-invariant deep embedding for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5509713.
- [10] R. Jiang, S. Mei, M. Ma, and S. Zhang, "Rotation-invariant feature learning in VHR optical remote sensing images via nested siamese structure with double center loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3326–3337, Apr. 2020.
- [11] J. Parmar, S. S. Chouhan, V. Raychoudhury, and S. S. Rathore, "Open-world machine learning: Applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, pp. 1–37, 2021.
- [12] T. Ahmad et al., "Variable few shot class incremental and open world learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3688–3699.
- [13] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised scene de-occlusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3784–3792.
- [14] A. Kortylewski, J. He, Q. Liu, and A. L. Yuille, "Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8940–8949.
- [15] S. Qiu, G. Wen, and Y. Fan, "Occluded object detection in high-resolution remote sensing images using partial configuration object model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1909–1925, May 2017.
- [16] Y. Cui, B. Hou, Q. Wu, B. Ren, S. Wang, and L. Jiao, "Remote sensing object tracking with deep reinforcement learning under occlusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5605213.
- [17] X. Yang, S. Li, S. Sun, and J. Yan, "Anti-occlusion infrared aerial target recognition with multi-semantic graph skeleton model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629813.
- [18] J. Zhang and H. Huang, "Occlusion-aware UAV path planning for reconnaissance and surveillance," *Drones*, vol. 5, no. 3, 2021, Art. no. 98.
- [19] K. Scott, R. Dai, and M. Kumar, "Occlusion-aware coverage for efficient visual sensing in unmanned aerial vehicle networks," in *Proc. IEEE Glob. Commun. Conf.*, 2016, pp. 1–6.
- [20] D. Zeng, R. Veldhuis, and L. Spreeuwiers, "A survey of face recognition techniques under occlusion," *IET Biometrics*, vol. 10, no. 6, pp. 581–606, 2021.
- [21] J. Chen, T. Zhou, H. Yang, and F. Fang, "Cortical dynamics underlying face completion in human visual system," *J. Neurosci.*, vol. 30, no. 49, pp. 16692–16698, 2010.
- [22] M. Shao, C. Wang, W. Zuo, and D. Meng, "Efficient pyramidal GAN for versatile missing data reconstruction in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626014.
- [23] T.-Y. Ji, N. Yokoya, X. X. Zhu, and T.-Z. Huang, "Nonlocal tensor completion for multitemporal remotely sensed images inpainting," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3047–3061, Jun. 2018.
- [24] L. Sun, Y. Zhang, X. Chang, Y. Wang, and J. Xu, "Cloud-aware generative network: Removing cloud from optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 691–695, Apr. 2020.
- [25] J. Dong, R. Yin, X. Sun, Q. Li, Y. Yang, and X. Qin, "Inpainting of remote sensing SST images with deep convolutional generative adversarial network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 173–177, Feb. 2019.
- [26] D. Feng, X. Shen, Y. Xie, Y. Liu, and J. Wang, "Efficient occluded road extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 4974.
- [27] C. Yang, V. Ablavsky, K. Wang, Q. Feng, and M. Betke, "Learning to separate: Detecting heavily-occluded objects in urban scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 530–546.
- [28] X. Zhang, Y. Liu, C. Huo, N. Xu, L. Wang, and C. Pan, "PSNet: Perspective-sensitive convolutional network for object detection," *Neurocomputing*, vol. 468, pp. 384–395, 2022.
- [29] F. Yang, R. Wang, and X. Chen, "Semantic guided latent parts embedding for few-shot learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5447–5457.
- [30] S. Nazir, M. H. Yousaf, and S. A. Velastin, "Inter and intra class correlation analysis (IICCA) for human action recognition in realistic scenarios," in *Proc. IEEE 8th Int. Conf. Pattern Recognit. Syst.*, 2017, pp. 1–6.
- [31] J. Wang, A. Ma, Y. Zhong, Z. Zheng, and L. Zhang, "Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery," *Remote Sens. Environ.*, vol. 277, 2022, Art. no. 113058.
- [32] Y. He, M. Li, J. Zhang, and J. Yao, "Infrared target tracking based on robust low-rank sparse learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 232–236, Feb. 2016.
- [33] Y. Wang et al., "Detecting occluded and dense trees in urban terrestrial views with a high-quality tree detection dataset," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707312.
- [34] Y. Ren, C. Zhu, and S. Xiao, "Deformable faster R-CNN with aggregating multi-layer features for partially occluded object detection in optical remote sensing images," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1470.
- [35] S. Qiu, G. Wen, Z. Deng, Y. Fan, and B. Hui, "Automatic and fast PCM generation for occluded object detection in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1730–1734, Oct. 2017.
- [36] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.
- [37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
- [38] 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.*, 2020, pp. 22243–22255.
- [39] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.
- [40] J.-B. Grill et al., "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.
- [41] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9640–9649.

- [42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
- [43] I. Misra and L. V. D. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6707–6717.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [45] P. Khosla et al., "Supervised contrastive learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.
- [46] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.
- [47] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond pascal: A benchmark for 3D object detection in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 75–82.
- [48] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [49] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1100–1111, Mar. 2018.
- [50] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, nos. 1/2, pp. 17–23, 1950.
- [51] D. R. Wilson and T. R. Martinez, "The need for small learning rates on large problems," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2001, pp. 115–119.
- [52] R. Zhu et al., "ScratchDet: Training single-shot object detectors from scratch," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2268–2277.
- [53] W. Li and C.-Y. Hsu, "Automated terrain feature identification from remote sensing imagery: A deep learning approach," *Int. J. Geograph. Inf. Sci.*, vol. 34, no. 4, pp. 637–660, 2020.
- [54] N. Ketkar, "Introduction to Keras," in *Deep Learning With Python*. Berlin, Germany: Springer, 2017, pp. 97–111.
- [55] L. Van derMaaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.



Jianming Xu received the B.E. degree in remote sensing from the China University of Geosciences, Wuhan, China, in 2020. He is currently working toward the M.E. degree in cartography and geography information system with Sun Yat-sen University, Guangzhou, China.

His research interests include deep learning of remote sensing images, representation learning, and multimodal information fusion.



Yunfei Li (Graduate Student Member, IEEE) received the B.S. degree in remote sensing from Jilin University, Changchun, China, in 2018, and the M.S. degree in remote sensing in 2021 from Sun Yat-sen University, Guangzhou, China, where he is currently working toward the Ph.D. degree.

His research interests include fusion of multisource remote sensing data.



Qian Shi (Senior Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, in 2015.

She is currently an Associate Professor with the School of Geography and Planning, Sun Yat-sen University, Guangzhou, China. Her research interests include remote sensing image classification, including

deep learning, active learning, and transfer learning.



Lin He (Member, IEEE) received the B.S. degree from the Xi'an Institute of Technology, China, in 1995, the M.S. degree from Chongqing University, Chongqing, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xi'an, China, in 2007.

Since 2007, he has been with the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China, where he is currently a Full Professor. His research interests include hyperspectral image processing, UAV image processing, spectral-spatial-temporal information fusion, machine learning and high-dimensional signal processing.