Occlusion Scene Classification via Cascade Supervised Contrastive Learning

jianming xu

1Sun Yat-Sen University

October 30, 2023

Abstract

This article proposes a contrastive learning-based scene classification method to extract features under occlusion circumstance. Novel occlusion features analysis methods are proposed also.
Abstract—Recent occlusion scene classification focuses on extracting occlusion-invariant features by considering class separability, unsupervised learning, and generalization ability, separately. Considering them together is profitable to address real-world occlusion problem. In this work, we propose cascade supervised contrastive learning to accommodate those considerations simultaneously. First, the pretext task uses rectangles to occlude the original images. Second, we use the siamese residual network to extract the features of original images, and pretext task augmented images. Third, compute the cascade supervised contrastive loss of the intermediate layers and final layer of the network. By optimizing the cascade supervised contrastive loss, the network can learn pretext-task invariant features. Besides, the pretext task-invariant features can adapt to the downstream occlusion-invariant feature extraction task due to the outstanding generalization ability of contrastive learning. Finally, a multi-layer perceptron is used to perform scene classification on the extracted features. The proposed method is validated on two simulated datasets and tends to achieve performance improvement compared with typical unsupervised occlusion scene classification methods. The first application of contrastive learning for remote sensing image occlusion scene classification is the innovative contribution of this paper. This method can achieve accurate remote sensing scene classification with any degree and type of occlusion.

Index Terms—remote sensing, scene classification, constrastive learning, occlusion handling

I. INTRODUCTION

Scene classification [1] is the process of extracting semantic features and classifying a scene. It is crucial to remote sensing deep learning-based recognition tasks such as scene understanding [2], object detection [3], and segmentation [4]. Scene classification networks often serve as feature extractors in recognition tasks. Hence, the recognition performance is highly dependent on the feature extraction capacity of scene classification networks. Recently developed scene classification networks have superior feature extraction performance due to their inborn spatial modeling capability and deep hierarchical nature [5]. By employing them as feature extractors, downstream recognition have witnessed tremendous progress [6]. However, the widely used CNN-based scene classification models are not robust towards occlusion [7]. As shown in Fig. 1, occlusion is usually inevitable in remote sensing. It leads to severe performance degradation of recognition tasks when objects are partially occluded [8]. Therefore, endowing CNNs with occlusion handling ability has already been noticed. And related studies can often be categorized into occlusion recovery-based and invariant feature extraction methods [9].

Occlusion recovery-based methods mimic the human neural system information completion process in the occlusion scenario [11], which retrieves the missing information in image [12] or feature space [13]. With the rapid development of generative models like Auto-Encoder(AE) [14]–[16] and Generative Adversarial Network(GAN) [17]–[19], occlusion recovery-based methods achieve significant improvement. However, image space recovery (image inpainting) [12], [20] is computationally expensive because of using deep generative models. Meanwhile, feature space recovery commonly requires extra information [21], while fusion is already a complex problem. Thus, occlusion-invariant feature extraction methods are more common in scene classification. Occlusion-invariant feature extraction methods aim to search for a space insensitive to occlusion change. Triplet [23]–[25] and low-rank representation [26]–[28] are utilized to relieve low feature separability caused by occlusion. Houshman et al. [31] and Aly et al. [32] used transfer learning to improve the model generalization to occlusion patterns. Qiu et al. [29] and Li et al. [30] employed semi / unsupervised learning to alleviate the dependency of occlusion annotations. Existing occlusion-invariant feature extraction methods improve the performance significantly, but they view occlusion handling from an isolated perspective. However, real-world occlusion is complex, which hinders collecting representative supervised datasets and demands networks with strong feature separability and generalization ability. Considering feature separability, semi / unsupervised learning, and generalization ability simultaneously is beneficial for real-world occlusion handling.

The goal of this study is to build a network could si-
multaneously accommodate feature separability, generalization ability and unsupervised manner. Contrastive learning has the potential to achieve the goal. First, contrastive learning effectively improves feature separability. It aims to learn a metric space in which positive sample pairs are closer and negative sample pairs are further apart. The inter-class similarity will be effectively reduced by defining occluded and non-occluded images as positive sample pairs and the rest as negative sample pairs. Second, the features learned by contrastive learning are generalizable. Recently, contrastive learning has been widely used for model pretraining. The pretrained model will adapt to downstream applications by fine-tuning with a small amount of data. Contrastive learning models such as SimCLR [33], [34], MoCo [35]–[37], and BYOL [38] outperform traditional supervised pretraining. It shows that contrastive learning usually learns features with better generalization. Third, contrastive learning could eliminate the dependence on occlusion annotation. Often, real tasks are complex and require manual annotations. Contrastive learning first simplifies the real task to form a pretext task [39]. For example, the complex real occlusions could be simplified as randomly positioned rectangle occlusions. Then, features invariant to randomly positioned rectangle occlusion are learned. Due to the generalization ability of contrastive learning, the pretext task-invariant features could adapt to real occlusion, forming occlusion-invariant features. However, directly using contrastive learning on remote sensing occlusion scene classification is inadequate, problems of which can be summarized as follows:

1) The improvement of feature separability is incomplete. Occlusion affects feature separability by causing low intra-class similarity and inter-class diversity [22]. Contrastive learning ignores category information, only source image and occlusion augmented image are treated as positive. In such case, same class images in the input batch are defined as negative, and the feature will be pulled away. There are up to 1000 categories in computer vision datasets [40], and the ratio of the same class samples in the batch is low, so the problem is not significant. However, in remote sensing datasets, the number of categories is usually 10-20 [1], [10], [41], which leads to the problem of loose intra-class distribution.

2) The constraint of contrastive learning is insufficient. CNNs extract features hierarchically. As the depth increases, the extracted features are more high-level [42]. However, recent contrastive learning only constrains the final layer of CNN. The intermediate layers also contain semantic information, which should also be occlusion-invariant. Lacking the constraints on intermediate layers could weaken the occlusion scene classification performance.

To fill the gap between contrastive learning and remote sensing occlusion scene classification, we propose a network called cascade supervised contrastive learning (Cascade-SupCon). First, this network introduces supervised contrastive learning to handle the feature separability problem. The model defines positive and negative samples more accurately than self-supervised contrastive learning by introducing category information. Images with the same class are defined as positive and the others as negative. The inter-class similarity reduces after enlarging the distance of the negative features. The intra-class similarity increases after diminishing the distance of the positive features. Second, cascade strategy adds contrastive loss constraint to intermediate layers and the last layer. The reason is not only the last layer contains the semantics, but the intermediate layers of CNN also contain semantics. It is insufficient to only require the last layer to fit contrastive loss constraint. The main contributions of this paper can be identified as follows:

1) We propose a new contrastive learning-based occlusion scene classification method. The existing scene classification network are not robust enough to occlusion. The proposed Cascade-SupCon promotes the CNN occlusion handling ability through increasing the feature separability and generalization in an occlusion-level unsupervised manner. To the best of our knowledge, this is the first work that introduces contrastive learning into the remote sensing occlusion scene classification.

2) We improve the current contrastive learning strategy. The proposed network adopts a cascade strategy that requires the intermediate and last layer to satisfy the contrastive constraint. Through cascade strategy, occlusion-invariant features are learned hierarchically in CNN. Therefore, occlusion scene classification performance is boosted.

3) We introduce new evaluation metrics for occlusion scene classification. Usually, occlusion handling performance is only evaluated by overall accuracy. To evaluate the occlusion handling capability more comprehensively, we additionally introduce metrics to measure the spatial autocorrelation between occlusion and feature distribution, inter-class discriminative, intra-class similarity, and feature separability.

The remainder of this paper is organized as follows. Section II describes the architecture of Cascade-SupCon in detail. Section III present the experiment results and analysis of Cascade-SupCon. In Section IV conclusion and brief discussion of future works will be presented.

II. PROPOSED METHOD

Section II-A analyzes the effects of occlusion on feature space and explains our motivation. As shown in Fig. 2, the proposed Cascade-SupCon consists of the following components: 1) pretext task for training model in occlusion-level unsupervised manner; 2) siamese residual network for feature extraction; 3) cascade supervised contrastive learning for constraining features be occlusion-invariant; 4) classifier for category prediction. Then, the components are described in detail in section II-B.

A. PROBLEM DESCRIPTION

1) Occlusion Effects on Feature Space: In CNNs, the learned features are normalized. Therefore, features are dis-
Contrastive learning introduces a certain degree of occlusion-invariant by decreasing the distance between the anchor image and the occlusion augmented image. As shown in Fig. 3(b), features of augmented images and source images (e.g., airplanes represented by red and pink dot) clustered together. However, the same class images in input batches are defined as negative samples due to contrastive learning ignoring category information. It cause samples of same category exclude each other. As shown in Fig. 3(b), the intra-class diversity is even more extensive by using contrastive learning. To alleviate occlusion effects, the method should satisfy the following characteristics:

1) The feature is occlusion augmentation-invariant. The feature distances between the anchor image and its occlusion augmented image should be reduced.
2) Compact the feature distribution of same class images by decreasing the distances of same class image features.
3) Enlarge the inter-class diversity by increasing the distances of different class image features.

Therefore, supervised contrastive learning is introduced. First, supervised contrastive learning is occlusion augmentation-invariant because it defines occluded augmented image as positive like contrastive learning. Second, samples of the same category are considered positive and their distances will be decreased. As shown in Fig. 3(c), airplanes features cluster tightly and form compact intra-class distribution. Third, samples of the different category are considered negative. The inter-class diversity will increases by enlarging the negative samples distance.

### B. Cascade-SupCon Implementation

1) **Pretext Tasks**: Pretext tasks [39] are simplified downstream tasks that enable artificially generating pseudo-labels.

2) **Motivation** Intuitively, contrastive learning is a suitable technique to handle occlusion effects. Given an anchor image \( x_i \), its positive samples set is \( \{ \text{Occ}(x_i) \} \) in traditional contrastive learning.
In order to simulate occlusion, the pretext task is pasting green rectangles of arbitrary size and location. Contrastive learning models are proposed to address tasks like scene classification and object detection. Their training process are class-level unsupervised, and the pseudo-label is the data itself. However, only occlusion-level unsupervised is required in occlusion scene classification. Besides, ignoring category information will cause increased intra-class diversity, according to the analysis in section II-A. The pseudo-label is the category of the image in this work. As shown in Fig.2, the input \( B = \{I_i, y_i\}_{i=1,2,...,N} \) denotes a minibatch of \( N \) samples where each sample consists of an optical image \( I_i \in \mathbb{R}^{3 \times H \times W} \) and corresponding one-hot encoded label \( y_i \). Let the occlusion augmentation be \( \text{Occ}(\cdot) \), the output of the pretext task module is \( A = \{(I_1, y_1), ..., (I_N, y_N), (\text{Occ}(I_1), y_1), ..., (\text{Occ}(I_N), y_N)\} \).

2) Siamese Residue Network: Siamese network consists of two sub-networks, which separately extract the features of two sets of input. In the proposed method, we use the convolutional layer of the ResNet-50 [45] as the sub-network to construct the siamese network. As shown in Fig. 4, Siamese residue network is composed of stem layer and stage layer. The stem layer is constructed by the convolution layer containing 64 \( 7 \times 7 \) convolution kernels and a \( 3 \times 3 \) max-pooling layer. Stage layer is defined by a tuple \((N, C)\) which means it is constructed by \( N \) cascade-connected bottleneck that are defined by the number of channels \( C \). The bottleneck first uses \( C \) convolution kernels of size 1 to reduce the input dimension. Secondly, it uses \( C \) convolution kernels of size 3 to extract features. Then, \( 4 \times C \) convolution kernels of size 1 are adopted to raise the dimension of the features. Finally, input and features are sums to form the output.

Similar to self-supervised contrastive learning, the siamese residue network is weight-shared in order to reduce the amount of parameters. Fig. 4 shows that instead of building two identical sub-networks, an equivalent implementation that integrates two sets of data and processes them with the same network is adopted. The input \( x = \{I_1, ..., I_N, \text{Occ}(I_1), ..., \text{Occ}(I_N)\} \) integrates \( N \) occlusion-free, and \( N \) occluded \( W \times H \) three-channel optical images into a mini-batch and therefore can be represented as a \((2N, 3, H, W)\) tensor. The input first get through a stem layer and then be processed by four stage layer in sequence. Finally, an adaptive pooling layer is used to flatten the tensor to feed the classifier. In this work, we output the features of the adaptive pooling layer and the 4th stage layer to take part in cascade supervised contrastive learning.

3) Cascade Supervised Contrastive Learning: Given a set of original images features \( u = (u_1, ..., u_N) \), augmented images features \( v = (v_1, ..., v_N) \) and corresponding labels \( y = (y_1, ..., y_N) \). Considering the deep representation is high dimensional, adapt cosine similarity to measure the distance of features. Eq. (1) illustrates the cosine similarity computation between \( u_i \) and \( v_j \) where \( i = 1, ..., N \) and \( j = 1, ..., N \).

\[
d(u_i, v_j) = \frac{u_i^T v_j}{||u_i|| ||v_j||}
\]

Because the residue siamese network uses a single network to process integrated two sets of data, the input of cascade supervised contrastive learning is in the form of \( x = (x_1, ..., x_{2N}) = (u_1, ..., u_N, v_1, ..., v_N) \). Assume \( x_i \) and \( x_j \) where \( i \in (1, ..., 2N) \) and \( j \in (1, ..., 2N) \) is positive pair, the loss of positive pairs is defined as:

\[
l_{i,j} = -\log \frac{\exp (d(x_i, x_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp (d(x_i, x_k)/\tau)}
\]

where \( \mathbb{1}_{k \neq i} \) denotes an indicator function output 1 if \( k \neq i \) and \( \tau \) refers to temperature parameter. Hard negative sample refers to a negative sample whose features are similar to the anchor image. Temperature parameter \( \tau \) decide the model sensitivity towards hard negative samples. When \( \tau \) is higher, the loss of hard negative cases is greater, and the model is more sensitive towards hard negative samples.

Let \( P(x_i) \) be positive sample set. The loss of \( x_i \) can be defined as:

\[
l_i = \frac{1}{|P(x_i)|} \sum_{x_p \in P(x_i)} l_{i,p}
\]

\[
l_i = -\frac{1}{|P(x_i)|} \sum_{x_p \in P(x_i)} \log \frac{\exp (d(x_i, x_p)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp (d(x_i, x_k)/\tau)}
\]

where \(|P(x_i)|\) is the cardinal number of \( P(x_i) \). When minimizing Eq.(3), the numerator ensures the distances of positive pairs will be decreased. Meanwhile, the denominator enlarges the distances of negative pairs. In self-supervised contrastive learning, the positive sample set of \( x_i \) is \( P^+(x_i) = \{x_i\} \). It introduces pretext task-invariant into the model. In Cascade-SupCon, the positive sample set of \( x_i \) is \( P^+(x_i) = \{x_i|y_i = y_j\} \). First, \( P^+(x_i) \subseteq P(x_i) \). Cascade-SupCon is also pretext-task invariant and help enlarge the inter-class discriminative. Second, same class features are defined as positive. This
where \( y \) is the probability of the occlusion-inviant features. Cross entropy loss is selected as loss. Cascade supervised contrastive loss conduce to extracting class prediction. As Eq.7 loss function can be defined as the generating occlusion-invariant features and producing accurate connected layer is adopted to be classifier and the computation of the adaptive pooling layer and 4th stage layer of residule siamese network is respectively.

4) Classifier: The output of adaptive pooling layer of residue siamese network is \( x^{pool} = \{x_1^{pool}, ..., x_{2N}^{pool}\} \) where \( x_i^{pool} \in \mathbb{R}^{2048}, i = 1, ..., 2N \). In \( x^{pool} \), the occlusion-free features set is \( x^{ori} = \{x_i^{pool}\}_{i=1,...,N} \) and the augmented features set is \( x^{aug} = \{x_i^{pool}\}_{i=N+1,...,2N} \). Considering the occlusion augmented samples are benefits to occlusion scene classification, take \( x^{aug} \) as the input of classifier.

Given the class number \( M \) and input \( x^{aug} \), one layer of fully connected layer is adopted to be classifier and the computation can be defined as:

\[
y' = x^{aug}W + b \tag{6}
\]

where \( y' \) is the classification result, \( W \) and \( b \) is the weight matrix and bias of fully connected layer respectively.

Finally, the objective of the network is simultaneously generating occlusion-invariant features and producing accurate class prediction. As Eq.7 loss function can be defined as the sum of cascade supervised contrastive loss and classification loss. Cascade supervised contrastive loss conduce to extracting occlusion-inviant features. Cross entropy loss is selected as classification loss as Eq.8 and contribute to precise classification.

\[
L = L^{con} + L^{cls} \tag{7}
\]

\[
L^{cls} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{i,j} \log(y'_{i,j}) \tag{8}
\]

where \( y_{i,j} \in [0,1] \) denotes if the \( i \)th image in mini-batch belongs to the \( j \)th class and \( y'_{i,j} \in [0,1] \) refers to the probability of the \( i \)th image in mini-batch belongs to the \( j \)th class.

III. EXPERIMENTS

A. Experiment Configuration

1) Simulation Data: The experiments are performed on the simulated occlusion scene classification datasets DIOR-Occ and LEVIR-Occ, as shown in Fig. 5. They are manufactured based on the DIOR [10] and LEVIR [41] object detection datasets. Although the proposed method is occlusion-level unsupervised, a test set with occlusion information is required to measure the occlusion robustness. To the best of our knowledge, there are no remote sensing scene classification datasets annotated with occlusion information, so we use simulation occlusion datasets. The simulation data is generated refers to datasets [46], [47] widely used in other scopes and has following characters:

- Accurate occlusion annotation. The occlusion of the generated data must be consistent with the given occlusion parameters. The scheme of cropping the foreground area of the target detection dataset and generating occlusion is adopted. It can make the occlusion of the generated data closer to the parameters. If the scene classification dataset is used, the position and size of the target are unknown, and there is a big difference between the generated occlusion and the parameters. For example, only the environment is occluded in the generated data, and the target is not, resulting in inaccurate labeling.

- Multiple occluder sizes. The differences in the size of occluders are common and will affect the classification accuracy to varying degrees. Hence, there should be a large number of data with different occlusion degrees in the simulated data. In the simulated data set, there are four occlusion levels of data, namely: L0 (no occlusion), L1 (20%-40% of the area is occluded), L2 (40%-60% of the area is occluded) and L3 (60%-80% of the area is occluded).

- Diverse occlusion type. In order to simulate the spectral diversity of occluders in the real world, different spectral distributions are introduced into the simulation data. For rectangular occlusion, there are black rectangles and random noise rectangles. For tree occlusion, there are green-leaf trees and yellow-leaf trees. For cloud cover, there are white clouds and dark clouds. For camouflage occlusion, different types of camouflage are set for different objects.

- Multifarious occluder properties. In order to simulate the diversity of occluder properties, the same type of occluders with different shapes are introduced into the simulation data. Generate rectangle and camouflage occlusion by random aspect ratio. For tree and cloud occlusion, introduce occluders of different shapes, such as sparse trees, dense trees, sparse clouds, and thick clouds.

2) Dataset Configuration: 70% of the data was used as a training set, 10% was used as a validation set, and 20% was used as a test set. After the simulated dataset is generated:

a) DIOR-Occ has 4 types of objects (airplanes, ships, storage tank, and windmills), with 17635 samples in the training set, 2518 samples in the validation set, and 5042 samples in the test set.

b) LEVIR-Occ consists of 3 types of objects (airplanes, ships, and storage tank), with 7718 samples in the training set,
1101 samples in the validation set, and 2209 samples in the test set. The proposed algorithm is occlusion level unsupervised, so there is only L0 unoccluded data in the training set. In the dataset, there are four occlusion categories (rectangle, tree, cloud, and camouflage) and four occlusion levels (L0, L1, L2, L3). The occlusion categories and occlusion degrees are uniformly distributed in both test and validation sets.

3) Evaluating Metrics: Five quantitative indicators are used to analyze the model as follows:

- The classification accuracy is used to measure 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 performance of the model using the overall classification accuracy.

- In order to analyze the response of intra-class features to occlusion, the Global Morans Index (GMI) [48] is used to measure the spatial autocorrelation of features on occlusion levels and occlusion types. Positive spatial autocorrelation means that objects with similar properties will gather together, while negative spatial autocorrelation means that objects with similar properties will repel each other. When the spatial autocorrelation is weak, the data is randomly distributed. The value of GMI ranges from -1 to 1. When the data is closer to -1, it shows negative spatial autocorrelation, the closer the data is to 0, it shows the random distribution, and the closer the data is to 1, it shows positive spatial autocorrelation.

- Use the trace of the intra-class covariance matrix (TSW) to measure the difference of the intra-class features. The intra-class covariance matrix of a land class is defined as the covariance matrix of the land class features. TSW is defined as the prior probability-weighted sum of the intra-class covariance matrices of each class. The smaller the TSW, the smaller the intra-class feature differences and the better the separability.

- Use geometric separability criterion (J) to measure the overall separability of the features. J is defined as TSB/TSW, which can comprehensively consider intra-class differences and inter-class differences. The larger the J, the larger the inter-class difference, the more compact the intra-class distribution, and the larger the J, the better the separability.

4) Implementation Details: The spatial size of the input image is $224 \times 224$, which is normalized to the input network. The network is trained for a total of 20 epochs using SGD [49] as the optimizer, with a learning rate of 0.001 and a momentum of 0.9. At the same time, ExponentialLR is introduced, so that after each training Epoch, the learning rate is decreased to $\gamma$ times the learning rate of the previous Epoch. In the experiments, $\gamma$ was defined as 0.6. Our experiments are performed on an NVIDIA RTX-2070 Super with 8GB RAM and correspondingly our batch size is set to 4.

**B. Experiment Results and Comparison**

1) Classification Accuracy: Table I shows the accuracy of 5 occlusion-unsupervised scene classification methods on different occlusion levels data of the DIOR-Occ and LEVIR-Occ. According to Table 1, ResNet-50 has an accuracy of 100% in unoccluded (L0) data. However, as the degree of occlusion increases, the classification accuracy of ResNet-50 drops sharply. In the DIOR-Occ dataset, when the occlusion degree is only 20%-40% (L1), the accuracy has dropped
by 10.97%. When the occlusion degree reaches 40%-60% (L2), the accuracy drops by 35.03%, and when the occlusion degree reaches 60%-80% (L3). It even dropped by 58.1%. In the LEVIR-Occ dataset, the results are similar. This phenomenon shows that although traditional CNN can classify well-observed images well, they are not robust to occlusion.

Perform occlusion enhancement on the input image with a green rectangle of any position and scale, and use ResNet-50 to learn this pretext task. In DIOR-Occ, compared with ResNet-50, the L1 accuracy is improved by 4.99%, the L2 accuracy is improved by 12.54%, and the L3 accuracy is improved by 17.61%. In the LEVIR-Occ dataset, compared with ResNet-50, the performance of this method for occlusion scene classification is significantly improved. This proves that the proposed pretext task is practical for occluded scene classification. However, when the degree of occlusion is high, the classification accuracy of this method is unsatisfactory. In the L3 data of the DIOR-Occ dataset, the accuracy is even only 59.51%. In addition, in the LEVIR dataset, using pretext task augmentation reduces the classification accuracy on unoccluded data by 2.37% compared to ResNet-50.

The overall accuracy of SimCLR reaches 87.54% and 86.13% on DIO-Oc and LEVIR-Oc, which are much higher than 73.97% and 74.24% of ResNet-50. This shows that using unsupervised contrastive learning can help with occlusion scene classification. Compared with ResNet-50 with pretext task augmentation, the accuracy improved by 4.9% and 4%, respectively. This shows that using unsupervised contrastive learning to reduce the original sample features and the pretext task augmented sample features is better than directly learning the pretext task.

Table I: Occlusion Scene Classification Accuracies

<table>
<thead>
<tr>
<th></th>
<th>DIOR-Occ</th>
<th>LEVIR-Oc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L0  L1  L2  L3 Overall</td>
<td>L0  L1  L2  L3 Overall</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>100  89.03  64.97  41.90  73.97</td>
<td>99.34 85.41 62.65 49.57 74.24</td>
</tr>
<tr>
<td>ResNet-50 + aug.</td>
<td>99.51 94.02 77.51 59.51 82.64</td>
<td>96.94 92.09 78.10 61.37 82.13</td>
</tr>
<tr>
<td>SimCLR</td>
<td>98.78 92.48 85.52 73.38 87.54</td>
<td>97.98 92.94 81.89 71.73 86.13</td>
</tr>
<tr>
<td>SupCon</td>
<td>98.86 95.94 91.91 80.03 91.68</td>
<td>99.73 98.29 90.36 77.47 91.46</td>
</tr>
<tr>
<td>Cascade-SupCon</td>
<td>99.84 99.08 94.50 83.16 94.14</td>
<td>99.95 98.34 94.14 81.28 93.43</td>
</tr>
</tbody>
</table>

To explore the effect of occlusion levels on features, trees with evenly distributed occlusion levels are used to occlude unoccluded target images. Similarly, ResNet-50 and TSNE are used to extract features and reduce dimensionality to obtain

| GMI of DIOR and LEVIR Occluded by Tree of Different Occlusion Level |
|------------------------|----------------|----------------|----------------|----------------|
| airplane | ship | storagetank | windmill | average |
| DIOR(Occluded) | 0.704 | 0.542 | 0.715 | 0.667 | 0.657 |
| LEVIR(Occluded) | 0.654 | 0.383 | 0.622 | - | 0.553 |

The overall accuracy of SimCLR reaches 87.54% and 86.13% on DIO-Oc and LEVIR-Oc, which are much higher than 73.97% and 74.24% of ResNet-50. This shows that using unsupervised contrastive learning can help with occlusion scene classification. Compared with ResNet-50 with pretext task augmentation, the accuracy improved by 4.9% and 4%, respectively. This shows that using unsupervised contrastive learning to reduce the original sample features and the pretext task augmented sample features is better than directly learning the pretext task.

Table I: Occlusion Scene Classification Accuracies

<table>
<thead>
<tr>
<th></th>
<th>DIOR-Occ</th>
<th>LEVIR-Oc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L0  L1  L2  L3 Overall</td>
<td>L0  L1  L2  L3 Overall</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>100  89.03  64.97  41.90  73.97</td>
<td>99.34 85.41 62.65 49.57 74.24</td>
</tr>
<tr>
<td>ResNet-50 + aug.</td>
<td>99.51 94.02 77.51 59.51 82.64</td>
<td>96.94 92.09 78.10 61.37 82.13</td>
</tr>
<tr>
<td>SimCLR</td>
<td>98.78 92.48 85.52 73.38 87.54</td>
<td>97.98 92.94 81.89 71.73 86.13</td>
</tr>
<tr>
<td>SupCon</td>
<td>98.86 95.94 91.91 80.03 91.68</td>
<td>99.73 98.29 90.36 77.47 91.46</td>
</tr>
<tr>
<td>Cascade-SupCon</td>
<td>99.84 99.08 94.50 83.16 94.14</td>
<td>99.95 98.34 94.14 81.28 93.43</td>
</tr>
</tbody>
</table>

2) Intra-Class Analysis: We first explore the effect of occlusion on the intra-class distribution of features. Obtain the unoccluded object region images in the original DIOR and LEVIR datasets, use ResNet-50 for feature extraction, and then use TSNE [50] to reduce the dimensionality of the features to obtain the two-dimensional representation of the features. The feature distribution of unoccluded storage tanks in the DOR dataset is shown in Fig. 6(a), which mainly forms four clusters.

Table II: GMI of DIOR and LEVIR Occluded by Tree of Different Occlusion Level

| GMI of DIOR and LEVIR Occluded by Tree of Different Occlusion Level |
|------------------------|----------------|----------------|----------------|----------------|
| airplane | ship | storagetank | windmill | average |
| DIOR(Occluded) | 0.704 | 0.542 | 0.715 | 0.667 | 0.657 |
| LEVIR(Occluded) | 0.654 | 0.383 | 0.622 | - | 0.553 |
two-dimensional scatter points of features. A mapping of 0.25 for L0 data, 0.5 for L1 data, 0.75 for L2 data, and 1 for L3 data is established, and GMI is used to measure whether there is spatial autocorrelation on the occlusion level of the feature two-dimensional scatter. It can be seen from the Table II that the GMI of the DIOR dataset and the LEVIR dataset after being occluded by trees is significantly greater than 0, showing a strong positive spatial correlation. This proves that the occlusion level affects the feature distribution, and objects with similar occlusion levels will cluster together. The feature distribution of the occluded storage tanks images in the DIOR dataset is shown in Fig. 6(b). It can be seen that the features are aggregated into five clusters, and the spatial distribution of feature points is significantly related to the occlusion level. This shows that the occlusion level will enlarge the difference of the feature distribution of the same samples.

In order to explore the influence of occlusion types on features, black rectangles, noise rectangles, camouflages, clouds, and trees with specific occlusion levels were used to occlude unobstructed target images, and ResNet-50 and TSNNE were used to obtain feature two-dimensional scatter points. Establish a mapping with black occlusion of 0.2, noise occlusion of 0.4, camouflage occlusion of 0.6, cloud occlusion of 0.8, and tree occlusion of 1. The GMI is used to measure whether there is spatial autocorrelation between feature two-dimensional scatter points on occlusion types. As shown in Table III, 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 features have a positive spatial correlation in the occlusion type, that is, the features with similar occlusion types tend to cluster together. When the occlusion level is L2, the mean values of GMI 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. This shows that with the increase of occlusion level, the influence of occlusion type on the feature distribution increases sharply. Especially for the DIOR dataset, the GMI has a mean value of 0.866, which indicates that the features are almost distributed according to the occlusion type. When the occlusion level is L3, the GMI of the storage tank is 0.882. Fig. 6(c) shows the distribution of features in this case. It can be seen that the features form clusters by occlusion types, and the feature distribution is severely fragmented. This shows that the type of occlusion will cause great differences in intra-class features, and the richer the type of occlusion, the greater the difference.

In summary, the distribution of features exhibits positive spatial autocorrelation on both occlusion level and occlusion type. The occlusion level and occlusion type will affect the spatial distribution of features, and features with similar occlusion levels and similar occlusion types will form clusters. This phenomenon leads to fragmentation of intra-class feature distributions, leading to increased intra-class variability. To reduce the impact of occlusion, it is necessary to reduce the intra-class variability caused by occlusion level and occlusion type.

Five models are used to extract features for each category of scenes in the DIOR-Occ and LEVIR-Occ test sets. In Fig. 7, TSNNE is used to reduce the dimension of the storage tanks scenes in the two datasets, and the feature points are colored according to the occlusion level. It can be seen that in ResNet-50, the scattered points are severely fragmented, and the features 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 correlation. 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 feature points of different occlusion levels more evenly, effectively reducing the positive spatial correlation. As shown in Fig. 7(h)-(l), for the LEVIR-Occ dataset, compared to ResNet-50, pretext task augmented ResNet-50, SimCLR, and SupCon. Cascade-SupCon can better reduce the clustering distance and reduce the spatial positive related. As shown in Table IV, it can be seen that the GMI of ResNet-50 in the two datasets is 0.839 and 0.805, respectively, showing a positive spatial correlation. The GMI of Cascade-SupCon in the two datasets is 0.676 and 0.638, respectively, and the spatial positive correlation is the lowest, which can best eliminate the influence of occlusion level on the distribution of feature classes.

As shown in Fig. 8, TSNE is used to reduce the dimensionality of the storage tanks scenes in the two datasets, and it is colored according to the type of occlusion. As shown in Fig. 8(a) and (h), there are clusters formed by occlusion types in ResNet-50, and the feature distribution is fragmented. As shown in Fig. 8(b)-(c) and Fig. 8(i)-(j), pretext task augmented ResNet-50 and SimCLR reduce the distance between clusters and reduce spatial autocorrelation to varying degrees, but the results are unsatisfactory. As shown in Fig. 8(d)-(e), though SupCon reduces the distance between clusters, the positive spatial correlation is still serious. On the other hand, Cascade-SupCon blends the features evenly. As shown in Fig. 8(k)-(l), both SupCon and Cascade-SupCon effectively reduce the positive spatial correlation and reduce the intra-class distance for the LEVIR-Occ dataset. As shown in Table V, it can be seen that on the DIOR-Occ dataset, the GMI of Cascade-SupCon is 0.701, which most effectively eliminates the influence of occlusion types. On the LEVIR-Occ dataset, the GMI of SupCon and Cascade-SupCon are 0.622 and 0.627, respectively, which eliminate the effect of occlusion type almost equally well.

According to the above analysis, among the five unsupervised occlusion scene classification algorithms, Cascade-SupCon can best reduce the positive correlation between features and occlusion, effectively reduce the distance between clusters and reduce the difference in feature distribution. Therefore, Cascade-SupCon can effectively reduce the impact of occlusion on the distribution of intra-class features.
Fig. 8. Graphical illustration of the TSNE dimensionality reduction scatterplot of extracting features from storage tank images using each model. 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. Fig. (a)-(e) are visualizations of the DIOR-Occ dataset. Fig. (h)-(l) are visualizations of the LEVIR-Occ dataset.

Table V

<table>
<thead>
<tr>
<th></th>
<th>DIOR-Occ</th>
<th>LEVIR-Occ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>airplane</td>
<td>ship</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>0.964</td>
<td>0.914</td>
</tr>
<tr>
<td>ResNet-50 + aug.</td>
<td>0.957</td>
<td>0.778</td>
</tr>
<tr>
<td>SimCLR</td>
<td>0.838</td>
<td>0.717</td>
</tr>
<tr>
<td>SupCon</td>
<td>0.858</td>
<td>0.732</td>
</tr>
<tr>
<td>Cascade-SupCon</td>
<td>0.737</td>
<td>0.529</td>
</tr>
</tbody>
</table>

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

3) Inter-Class Analysis: To explore the influence of occlusion on inter-class features distribution, we use ResNet-50 for feature extraction on the original unoccluded DIOR and LEVIR datasets and the occluded DIOR and LEVIR datasets. Calculate the TSW, TSB, and J as shown in Table VI. It can be seen that in the DIOR dataset, the unoccluded TSB is 51.072, the distance between classes is large, the TSW is 34.998, the intra-class distribution is compact, the J is 1.460, and the separability is good. After adding occlusion, TSB is 18.671, TSW is 55.127, the inter-class distance is greatly reduced, the intra-class distribution is loose, J is only 0.338, and the class separability is sharply reduced. A similar situation can also be observed in the LEVIR dataset. Fig. 9 shows the feature scatter plots after TSNE dimensionality reduction before and after occlusion of the DIOR dataset. It can be seen that the overlap of different classes is small before the occlusion, and the distribution of the intra-class features is relatively concentrated. After occlusion, there is a large amount of overlap, the intra-class feature distribution is very loose, and the separability is poor. This shows that occlusion will reduce the inter-class distance and loosen the intra-class distribution, causing confusion.

Five algorithms are used for feature extraction on the test
sets of DIOR-Occ and LEVIR-Occ, respectively, and TSB, TSW, and J are calculated as Table VII. The TSB of ResNet-50 on the two datasets are 18.522 and 30.477, which are both small, which indicates that the distances of various features are small in the features extracted by ResNet-50. At the same time, its TSW are 55.043 and 2517.045 respectively, and the intra-class distribution is very loose and larger than TSB. This will lead to a lot of confusion among various categories, resulting in serious misclassification. Therefore, J is only 0.336 and 0.012, and the separability is poor. After using the pretext task, in the DIOR-Occ dataset, TSB is 20.430 comparable to ResNet-50, and TSW is 57.611 better than ResNet-50. That is, the pretext task reduces the intra-class distance while maintaining a similar distance between classes, and the separability is improved to a certain extent. In the LEVIR dataset, the situation is similar, keeping the TSB similar, while the TSW is greatly improved. This shows that the pretext task can effectively compact the intra-class distribution. SimCLR has a TSB of 71.017 and 70.978 in the DIOR and LEVIR datasets, respectively, which is a big improvement compared to ResNet-50 and agent task enhancement. However, SimCLR performs poorly on TSW, and in DIOR data, TSW even expands to 642.189, and the intra-class distribution is very loose. SupCon has the highest TSB on both datasets and has the best effect on expanding the inter-class distance. However, on the DIOR dataset, the TSW is enlarged, and on the LEVIR-Occ dataset, the TSW is larger than the surrogate task augmentation. This shows that SupCon can effectively expand the inter-class distance, but its ability to narrow the intra-class distribution is unsatisfactory. On the two datasets, the J of SupCon is 1.156 and 0.987, respectively. Compared with ResNet-50, the pretext task augmented ResNet-50 and SimCLR are both effective at reducing the intra-class distance, but the overlap is still serious and the separability is poor. As shown in Fig. 10(c) and (k), Cascade-SupCon has the best separability, its intra-class distribution is very compact, and the confusion of various types is relatively small, which is conducive to correct classification. Therefore, Cascade-SupCon can optimally solve the problems of small inter-class distance, loose intra-class distribution, and serious confusion caused by occlusion.

4) Cascade Layer Selection: In order to obtain the ideal classification effect of occlusion scenes, the number of cas-
Table VIII: Occlusion Scene Classification Accuracy of Supervised Contrastive Learning with Different Cascade Layers on Dior-Occ and Levir-Occ Test Sets

<table>
<thead>
<tr>
<th>Cascaded Layers</th>
<th>Dior-Occ</th>
<th>Levir-Occ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L0</td>
<td>L1</td>
</tr>
<tr>
<td>Pooling Layer (SupCon)</td>
<td>98.86</td>
<td>95.94</td>
</tr>
<tr>
<td>Pooling Layer + Stage 4 (Cascade-SupCon)</td>
<td>99.84</td>
<td>99.08</td>
</tr>
<tr>
<td>Pooling Layer + Stage 4-3</td>
<td>97.96</td>
<td>97.78</td>
</tr>
<tr>
<td>Pooling Layer + Stage 4-2</td>
<td>99.59</td>
<td>97.01</td>
</tr>
<tr>
<td>Pooling Layer + Stage 4-1</td>
<td>99.67</td>
<td>96.70</td>
</tr>
</tbody>
</table>

Cascaded layers needs to be carefully considered. Supervised contrastive learning maps specific-level features into an occlusion-robust space based on the assumption that unoccluded and occluded image semantics should be consistent. The number of cascade levels controls the feature level that is mapped. In order to explore the influence of the number of cascade layers on the classification of occlusion scenes, the occlusion scene classification accuracy on the two datasets by the number of cascade layers is shown in Table VIII. SupCon constrains the features of the pooling layer and achieves better classification accuracy of occluded scenes, but the classification performance of scenes with high occlusion degrees is not good. Cascade-SupCon performs supervised contrastive learning on pooling layer features and Stage-4 features and imposes more comprehensive constraints on semantic features. Compared with supervised contrastive learning, Cascade-SupCon effectively improves the classification accuracy of each occlusion level. However, adopting more cascaded layers is not always better. When shallow features are introduced, the classification accuracy of occluded scenes decreases to varying degrees.

In the Dior-Occ dataset, Stage-3 features are additionally introduced for supervised contrastive learning, and its classification accuracy is 91.21%, which is lower than 91.68% of supervised contrastive learning. In the Levir-Occ dataset, the phenomenon is more serious, and the features of Stage-3, Stage2, and Stage-1 are additionally introduced for supervised comparative learning, and the accuracies are 88.35%, 88.86%, and 91.26%, respectively. The classification accuracy is not only much lower than 93.43% of Cascade-SupCon but also lower than 91.46% of SupCon. This may be because the shallow layers of the convolutional neural network extract detailed features, and it is unreasonable to require that the details of the occluded image and the unoccluded image remain unchanged. According to the experimental results, we choose to perform supervised contrastive learning on the pooling layer and Stage-4 features to form a Cascade-SupCon model. In the follow-up work, we will build an occlusion scene classification model that is insensitive to the number of cascaded layers.

IV. Conclusion

We propose a novel scene classification network work under object partially occluded circumstances. The proposed Cascade-SupCon is based on contrastive learning, which is used to address remote sensing occlusion scene classification for the first time. By employing contrastive learning, the occlusion causes high inter-class similarity decreases without the need for occlusion annotation. In addition, the method can also increase the intra-class similarity by introducing category information. Cascade-SupCon generates occlusion-invariant features hierarchically using cascade strategy, which improves the occlusion handling capability. We evaluate the model in two simulated datasets. The result shows that our model can adapt to severe and unseen occlusion. We further introduce more evaluation metrics to assess occlusion scene classification performance comprehensively. We investigate that occluder effects features distribution massively due to their positive spatial autocorrelated. It directly leads to high inter-class similarity and intra-class diversity. Cascade-SupCon can effectively reduce spatial autocorrelation and increase feature separability. The future work will expand this model to object detection and semantic segmentation scope.

References


