Aircraft Detection in Aerial Remote Sensing Images Based on Contrast Self-supervised Learning

Yuanyuan Liu(✉)

School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China
YuanyuanLiuch@gmail.com

Abstract. UAV aerial remote sensing images, which have the advantages of high resolution, convenient acquisition, high amount of information, and simple preprocessing process, are widely used in classification and detection tasks. The difficulty in the study of aircraft target detection in remote sensing images is that it relies on a large amount of labeled data and the target is relatively small. Therefore, this paper firstly studies the existing comparative learning methods in the self-supervised field, and then proposes the ZL (a comparative learning method in the remote sensing field) method for the small sample remote sensing data set to realize the extraction of the representation of the aircraft target. The ZL method mainly modifies the data enhancement combination form in data augmentation pipeline of the training process and the fine-tuning process and modifies the activation function in the projection head. Finally, the ZL trained model is combined with Faster R-CNN to improve the accuracy of aircraft target detection in aerially remote sensing images.

Keywords: Aerial remote sensing images · Small sample data set · ZL method · Target detection

1 Introduction

The emergence of remote sensing expands human's knowledge of their living environment by using non-contact, long-range detection technology to achieve large-range, multi spectral imaging for object detection. Aerial remote sensing has the advantages of mature technology, large imaging scale, high ground resolution, suitable for large-area terrain mapping and small-area detailed investigation, etc. It provides a new way for grounding target detection. Therefore, the research on target detection in aerially remote sensing images is of great significance. At the same time, an aircraft plays an important role in transportation, and play an important role in civil and military. In this context, aircraft small target detection is one of the hot research directions in the field of remote sensing environment detection.

X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 284–296, 2022.
https://doi.org/10.1007/978-3-031-04409-0_26

Target detection using convolutional neural networks is one of the most challenging projects in the field of computer vision, which contains two main sub-tasks, target localization and target classification. Convolutional neural network based target detection algorithms can be divided into two categories, one is region based target detection algorithms, represented by R-CNN as a two-stage algorithm. One category is regression-based target detection algorithms, forming a single-stage algorithm represented by You Only Look Once (YOLO) [9–11] and SSD (Single Shot multibox Detector). Gong et al. [2] put forward a context-aware convolutional neural network (CA-CNN) method to improve the performance of object detection. Liu et al. [3] proposed a multi-layer abstraction salience model for airport detection in synthetic aperture radar (SAR) images. Wei et al. [13] came up with an HR ship detection network (HR-SDNet) to perform precise and robust ship detection in SAR images. Those methods are more suitable for large-scale targets with high contrast in natural scenes, but the accuracy of detection results is lowed in the case of complex background and small target detection. Meanwhile, in order to achieve the fit of neural network models for application problems, large-scale data sets are required to be trained upfront, and the degree of models fit are affected by the accuracy of the labels in training set. The production of labels for large-scale data sets is time-consuming and expensive, and problems such as model generalization errors, false associations and adversarial attacks are encountered.

Based on the above problems, the main contributions of this article are show below: Firstly, this paper proposes a small target detection framework for aerially remote sensing images following the process of "ZL (A comparative learning method in remote sensing) self-supervised representation learning+Faster R-CNN target detection". Secondly, we design a comparative learning method of ZL in the field of remote sensing. It is a more suitable for small-sample aerial remote sensing images, and it is applicable to the characterization learning method of target detection with complex background, diverse targets and small size. Most importantly, the ZL contrast learning method is first to train model based on OPT Aircraft v1.0 data set to achieve higher accuracy in classifying aircraft types. We extract some of these weight parameters in this model, further base on RSOD data set and Faster R-CNN to achieve aircraft small target detection for aerial images.

2 Related Work

2.1 Analysis of Existing Comparative Learning

Self-supervised learning is a special type of unsupervised learning method with a supervised form, where the supervision sources self-supervised tasks instead of prior knowledge. The general idea consists of two main parts: a self-supervised training part and a downstream task part specific to a certain application. In the self-supervised training phase, the learning of the neural network model for the target representation is trained using a pretext task, rather than artificial labeling, to generate pseudo-labeled information. In the downstream task, the corresponding computer vision related tasks, such as classification, detection, segmentation, future prediction, etc., are done with the help of pseudo-labeled and the trained convolutional neural network model.

Contrast learning is a differentiated approach, a self-supervised task in the self-supervised training phase, for generating pseudo-labels of images. In principle some measure of similarity is used to bring similar samples closer together and different samples away from each other. Most of the early research in this area combined some form of instance-level classification methods [1, 7, 15] with contrast learning, and in recent years, many contrast learning approaches have been proposed, such as AMDIM [5], MoCo [14], SimCLR [8], BYOL [4], and SwAV [6]. The classification accuracy trained using these contrast learning approaches on the ImageNet [11] data set is comparable to that of supervised learning methods.

William Falcon et al. proposed a conceptual framework for CSL methods [12], which describes CSL methods in terms of five aspects: data augmentation pipeline, encoder selection, representation extraction, similarity metric, and loss function. Based on this framework, we analyze the existing comparative learning methods of AMDIM, MoCo, SimCLR, BYOL and SwAV (Table 1 Comparison of AMDIM, MOCO, SimCLR, BYOL, and SWAV methods).

Table 1. Comparison of AMDIM, MoCo, SimCLR, BYOL, and SWAV methods.

Author	Method	Data Augmentation Pipeline	Encoder	Representation extraction	Loss Function
Philip Bachman et al. (2019.7.8)	AMDIM	Random resizing of cropping Random color jitter Random grayscale	Self Encoders for ResNet	maximizing mutual information between arbitrary features extracted from multiple views of a shared context	Negative Contrast Estimation (NCE) loss
Kaiming He et al. (2020.3.23)	MoCo	Random resizing cropping Random color jitter Random horizontal flipping Random grayscale	Standard encoders for ResNet	Discrete dictionaries Momentum update	InfoNCE
Ting Chen et al. (2020.7.1)	SimCLR	Random crop Resizing to original size Random color jitter Random Gaussian blur	Standard encoders for ResNet	Final feature map Projection head	Negative Contrast Estimation (NCE) loss
Jean-Bastien Grill et al. (2020.9.10)	BYOL	Random crop Resizing to original size Random color jitter Random Gaussian blur	ResNet with deeper (50,101,152 200 layers) and wider (from 1× 4×)	The online network The target network	Positive sample loss estimation L_{θ}^{BYOL}
Mathilde Caron et al. (2021.1.8)	SWAV	Multi-crop strategy to increase the number of samples	Standard encoders for ResNet	Online clustering approach Projected head A "swap" prediction mechanism	$l(z_s, q_t)$

AMDIM method is mainly proposed by increasing multi-scale DIM information (between global and local features), maximizing negative samples of each image independently augmenting each image, maximizing mutual information between multiple feature scales, and a powerful encoder architecture designs to achieve technical innovation. The use of self-encoder design increases the complexity and poor practical results, while the use of feature map cross-validation increases the computational effort. The MOCO method is a mechanism for building dynamic dictionaries for comparison learning, maintaining dictionaries as queues of data samples: the coded representation of the current small batch enters the queue and the oldest leaves the queue. The queue decouples the dictionary size from the small batch size, making it larger. However, it uses dynamic refreshing of the coding area, which leads to inconsistency between old and new candidate codes, while occupying larger memory. SimCLR greatly improves the quality of the learned representation mainly by introducing a learnable nonlinear transformation between representation and contrast loss, and representation learning to use contrast cross-entropy loss benefits from normalized embedding and properly tuned temperature parameters. The nonlinear projection approach exacerbates the opacity of feature extraction and compresses the feature map to the implicit space, reducing the impact of the data augmentation pipeline. However, interpret ability of the data is reduced and the impact of unknown dimensions is unknown. BYOL relies on two neural networks, called the online network and the target network, which interact with and learn from each other. Starting from the enhanced view of an image, we train the online network to predict the target network representation of the same image under different augmentation pipeline. Meanwhile, we update the target network with a slow moving average of the online network. The BYOL method creatively combines GAN network ideas in MOCO and SimCLR, using only positive samples for training, for negative samples are not involved in training, and the small variability between samples is not considered. SWAV increases the number of views of an image without computation or memory overhead by introducing a multi-cropping strategy. Utilizes a scalable online clustering loss those works in both large and small batch settings without large memory banks or momentum encoders.

2.2 ZL Comparative Learning Method

The purpose of the data augmentation pipeline is to generate anchors, positive and negative features for comparative learning. It is experimentally demonstrated that the extraction strategy of the representations depends on the different data augmentation methods in the early stage. Different data augmentation pipeline affects not only the computational complexity, but also the effectiveness of representation extraction in the later stage. For example, the random flipping and random color grayscale processing strategies used by AMDIM have negligible effects on the results, but increase the computational complexity.

The resolution of aerially remote sensing images is higher than that of satellite remote sensing images, which has the advantages of high clarity, large scale and small area. The accuracy of aerial images depends on the aerial altitude under the same resolution of camera lens; and aerial photography can freely choose the weather and time. Aerially images mainly contain the top view information of ground objects, and the same type

of targets are prone to rotational variability. Scale variability between the same target in the information collected by sensors of different heights and impulse noise and gaussian noise are common in aerially remote sensing images. Those issues need attention in processing aerial remote sensing images.

Based on the characteristics of the remote sensing aerial images and the technology of existing data augmentation pipeline, the ZL method data enhancement methods (Fig. 1 ZL's positive sample data augmentation pipeline, Fig. 2 ZL's negative sample data augmentation pipeline) are mainly median filter processing, random rotation, random cropping resizing to original size, random flip, random color distortion, random grayscale transformation, and random Gaussian blur. The random rotation processing of positive and negative samples are necessary to improve the accuracy of top-1 linear evaluation by 1% in the image classification task of a 100 epoch trained ResNet-50 model with small batch processing. Secondly, this method combines the same image enhancement set of SimCLR for positive samples with random cropping followed by fifty percent probability level (left-to-right) flip, random color dithering, Gaussian blur, etc.

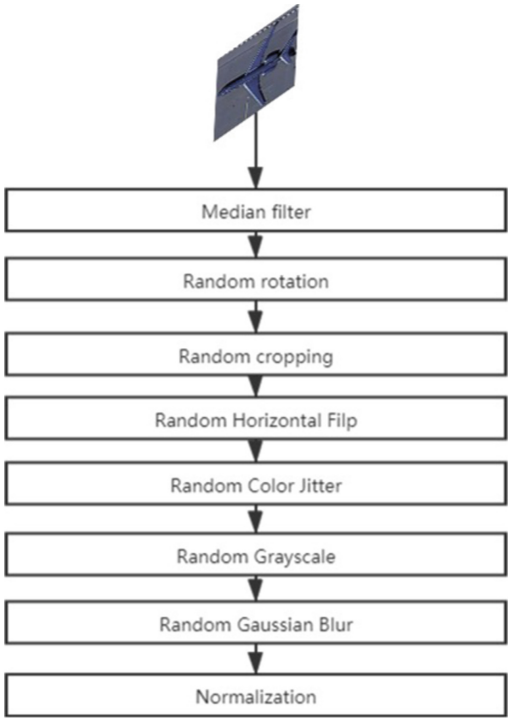

Fig. 1. Positive sample data augmentation pipeline of ZL.

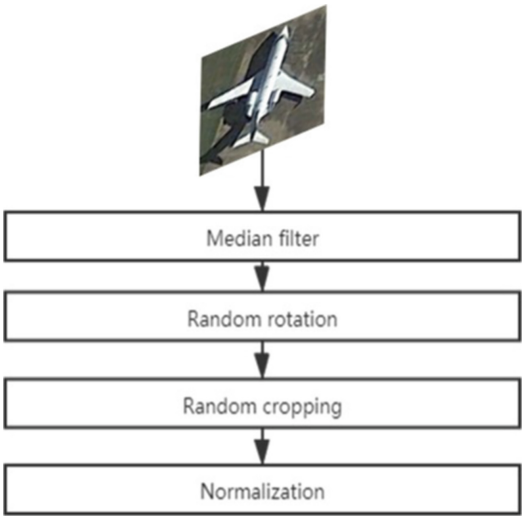

Fig. 2. Negative sample data augmentation pipeline of ZL.

Encoders play an essential role in any self-supervised learning pipeline, as they are responsible for mapping input samples into a latent space, enabling the transformation of an input sequence of indefinite length into a variable of fixed length. ZL method uses a dual channel encoder based on the Resnet network and end-to-end updates via backpropagation in training. At the same time, non-linear changes are added after the encoder to remove the information related to the enhancement channel of the image, and the number of negative samples is increased using a batch operation. This method is based on the nonlinear projection and implicit layer of SimCLR, and modifies the activation function in the projection header to a Sigmoid function. Based on a small sample size data set, the train and val training sets use the same augmentation pipeline, combine with normalization processing, and fine-tuning of label information using 10% of the data. The result achieves about 9% improvement in accuracy relative to the SimCLR method top-1.

A representation is a collection of unique characteristics that allows a system as well as humans to understand how something differs from other objects. This method uses the final feature map for comparison and reduces the corresponding computational effort. The excuse tasks used is a simple target classification task to extract images representation.

The measure of similarity is to measure the closeness between two sample embedding, and this method uses the cosine similarity, which the cosine similarity of two variables (vectors) is the cosine of the angle between them, which is defined as follows formula 1:

$$\cos\_\mathrm{sim}(A, B) = \frac{A \cdot B}{\backslash A \backslash \backslash B \backslash} \tag{1}$$

The loss function uses contrasting positive-negative samples to represent learning ability and is defined as a combination of positive and negative scores reflecting learning

progress. The minimization loss function corresponds to maximizing the positive scores and minimizing the negative scores. The loss functioned used in ZL method is the negative contrast estimate (NCE) loss, which is defined as follows formula 2.

$$L_{NCE} = -\log \frac{\exp(\text{sim}(q, k_+)/\tau)}{\exp\big(\text{sim}(q, k)_+/\tau\big) + \exp(\text{sim}(q, k)/\tau)_-} \tag{2}$$

where q is the original sample, k_+ is the positive sample, and k_- is the negative sample. τ is the hyperparameter used in most recent methods and is known as the temperature coefficient. The *sim*() function can be any similarity function, but generally uses the cosine logistic regression to distinguish between observed data and some artificially generated noise.

In summary, the process at the core of the ZL comparison methodology is as follows: Where N samples from the same batch are extracted to produce $2N$ sample data representation representations through the image enhancement method, encoder, and projection head processing, and negative samples are 2(N 1) data points from the same batch. The positive and negative sample representation point data are normalized using cosine similarity, and the encoder and projection head parameters are updated in reverse by minimizing the NCE loss function.

3 Experiment

3.1 Processing of the Data Set

In order to verify the effectiveness of the proposed ZL method based on remote sensing data, and to validate the accuracy of target detection based on aerial aircraft, extensive experiments are conducted in the paper with the existing data sets OPT Aircraft v1.0 and aircraft data in RSOD. OPT Aircraft v1.0 data set from Chinese Academy of Sciences based on public data set DIOR, UCAS AOD, NWPU VHR-10, DOTA and Google Earth images extracted from some valid images, total 3594 images, size 96 96 pix. RSOD aircraft data is labeled by Wuhan University team, total 446 images, size 1044 915, spatial resolution 0.3–3 m. Based on the above data, we eliminate some unreasonable data and rationalize the images to make them uniform in specification.

3.2 Learning of Aircraft Features Based on ZL Method

The types of aircraft at airports can be divided into swept-back aircraft, swept-back aircraft, swept-forward aircraft with trailing edge, deltawing aircraft, flat-wing aircraft, propeller aircraft, helicopters and so on mainly based on the color of the aircraft and the color of the engine. The similarity between different types is high, and it is difficult to learn sample features without classification labels, while the contrast learning method can better explore the differences between similar samples, so as learning features with more sample differentiation, which is more conducive to downstream task such as detection and classification of aircraft targets within the scene.

To verify the effectiveness of this method for aircraft feature learning, a linear evaluation protocol is used to train a linear classifier frozen on top of the underlying network for the aircraft classification task to test the accuracy validity of the characterization extraction on the OPT Aircraft v1.0 data set.

3.3 Faster R-CNN Based Aircraft Target Detection

In the aircraft target detection phase, the Faster R-CNN model, which has performed well on target detection tasks for generic scenarios, is used. The model consists of a backbone (base network) for feature extraction, an RPN for generating regions of interest (ROI), and the ROI-Head for generating the final detection results, where the ROI-Head includes two branches of classification and localization.

Backbone consists of a deep convolutional neural network that maps the input image into a deep feature map. The backbone parts of the model in the paper is initialized with the feature extraction network parameters obtained from the self-supervised aircraft feature learning to train in phase 1 to provide the target detection model with as much a prior knowledge of the airfield environment and aircraft-related knowledge as possible, and then the backbone parameters are finetuned while the rest of the model is trained. Figure 3 shows the part of the ZL learning method based on the ResNet50 network structure combined with the Faster R-CNN network structure, replacing part of the structure in the BackBone in Faster R-CNN.

The RPN network uses the feature map to generate a series of candidate boxes by outputting the offset of the predicted boxes relative to the anchor points, using the predefined anchor points as a reference. By integrating the feature map with the candidate frame information, the features of each region of interest can be obtained. In the paper, ROI-Align (ROI alignment) method is used instead of ROI-Pooling (ROI pooling) method to implement the process. In addition, the cross-first loss function and the Smooth L1 loss function are used for the classification task and the localization task, respectively, during the model training process.

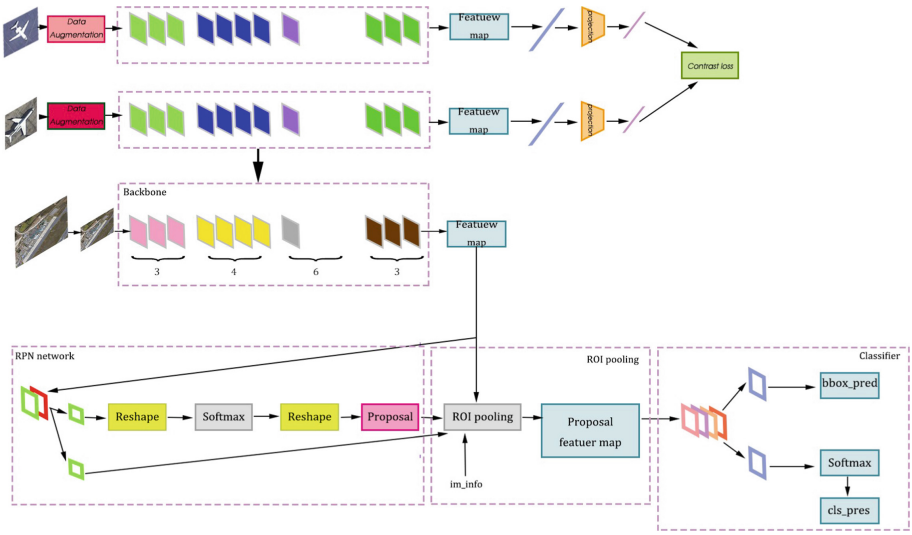

Fig. 3. ZL learning method combined with Faster R-CNN network structure.

At the same time, to better verify the effectiveness of the self-supervised pre-training method proposed in the paper, a type-independent aircraft target detection is performed, which all targets in the images are classified as “targets”.

4 Experimental Verification and Result Analysis

In the self-supervised aircraft feature learning phase, the coding in the model are used in the ResNet-50 network, and the model is trained using image data from the OPT Aircraft v1.0 data set (without using its corresponding labels), and the learned features are directly used for aircraft classification, which a linear classification layer is trained in a self-supervised manner with a fixed feature extraction network.

In the aircraft target detection phase, the backbone part of the model uses the −50 trained in the previous phase for model initialization, which ResNet-50+FPN is used to form the network structure of Faster R-CNN, and the training and testing of target detection are performed in the RSOD data set, which is divided into train and val as the training set, while the post test data set for the testing of the model. In the testing phase, average accuracy (AP) and average recall (AR) are selected as the scoring criteria, and the tests are conducted under the conditions of intersection over union (IOU) threshold of 0.3 and 0.5, respectively. The ZL comparative learning method is used to learn aircraft features on the OPT Aircraft v1.0 data set, and a linear classifier is trained on this basis to complete the classification task for 7 classes of aircraft. Table 2 shows the Top-1 for classification on the OPT Aircraft v1.0 data set. The experimental process is compared by training both the SimCLR method for self-supervised feature learning and the ZL method with 100 epochs of training. The data in Table 2 shows that the ZL method, compared with the SimCLR method, has a small difference in classification time consumption for each image in the test data set in small-sample aerial remote sensing data, but the improvement in classification accuracy is 20%. It proves the effectiveness of this method in achieving the target detection domain using small sample of remote sensing data, and further proves that this algorithm has certain advantages in practical use. Existing comparative learning methods based on ImageNET public data set learning of representation characteristics in classification TOP-1 accuracy comparable to supervised learning, relying on the training of a large data set in the early stage and deeper network structure as a support, based on simple, low computation network structure with small sample remote sensing data set has more obvious disadvantages.

In order to refine the optimal model trained by the ZL method, the optimal model is used for the subsequent aircraft target detection task to improve the convergence speed of the network training, while the optimizer of the ZL method selects the LARS optimizer, based on the capacity limitation of the storage of the actual device CUDA, and further analyzes the selection of the appropriate Batch size and Epoch size to improve the Top-1 classification accuracy. Table 3 shows the Top-1 accuracy for the different Batch size and Epochs training linear evaluation. Figure 4 shows the line graph of Top-1 accuracy for linear evaluation with different Epoch size and Epoch training. Based on Fig. 2, it can be intuitively seen that the optimization efficiency is better and the rate is faster under the same epoch of high batch size compared to the setting of low batch size data, but the improvement of the later data is relatively less. To sum up, in the process of

characterization learning, we should avoid raising Batch size, which causes unnecessary experimental settings, and determine the respective stable growth range through rough Batch size and Epoch settings, and reduce unnecessary iterations and excessive Batch size settings while adjusting the experimental accuracy to efficiently realize the learning rate in the training process. Rationalization of the parameter adjustment.

Table 2. Aircraft Top-1 on the OPT aircraft v1.0 dataset.

Method	Top-1	Training time	FPS
SimCLR	0.222841	70 min	22.27
ZL	0.423398	60 min	22.72

Table 3. Accuracy of Top-1 for linear evaluation with different batch size and epoch training.

Batch_size \ Epoch	100	200	300	400
64	0.3272980	0.3372987	0.3782993	0.4442896
128	0.3272980	0.4916435	0.5172930	0.5181001

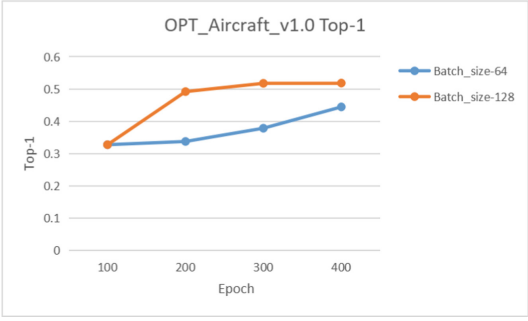

Fig. 4. Line graph of Top-1 accuracy for linear evaluation with different Batch size and Epoch training.

Based on the above data, In the small-sample unlabeled data, the ZL contrast learning method is used for training, in terms of time, is significantly lower than the time to make a large number of labels and train them, while the better classification level for small-sample unlabeled remote sensing data can be improved by reasonable planning of epoch and batch size sizes. Thus the method proves to be effective as a feature for learning unmarked remote sensing data, thus better serves for downstream aircraft classification and aircraft target detection.

To further validate the adaptability of the ZL contrast learning method, class-neutral target detection was performed on the RSOD data set using the Faster R-CNN framework, and ResNet-50 was used for BackBone, but a different pre-training approach was adopted for them, both using a structure based on 80% of the data set as training data and 20% as test data. Table 4 shows the experimental results based on the RSOD data set. Among them, Faster R-CNN indicates the result of training in a supervised way without modifying the internal hierarchy of the model based on OPT Aircraft v1.0 data set with batch size of 128; SimCLR+Faster R-CNN indicates SimCLR based on OPT Aircraft v1.0 data set ResNet50 with batch size of 128 and epoch of 200 as the BackBone of the Faster R-CNN network structure, trained on the RSOD data set; ZL+Faster R-CNN denotes ZL trained on the OPT Aircraft v1.0 data set with batch size of 128 and epoch of 200., epoch of 200 trained ResNet50 as BackBone of Faster R-CNN network structure, training results in RSOD data set. Through the analysis in Table 4, the ZL method and Faster R-CNN have significant advantages over Faster R-CNN and SimCLR+Faster R-CNN methods in all metrics for the target detection task. When IOU = 0.5, the AR value of ZL contrast learning method does not have a significant advantage, but there is a slight advantage relative to the current SimCLR method, thus proving that it is feasible based on the use of label-free contrast learning method for the analysis aspect of remote sensing images. Figure 5 shows the image processing results in SimCLR+Fast R-CNN and ZL+Faster R-CNN at IOU = 0.3, and Fig. 6 shows the image processing results in SimCLR+Faster R-CNN and ZL+Faster R-CNN at IOU = 0.5. Analyzed by Fig. 5 and Fig. 6, the two methods are comparable in terms of the error rate of target detection, and fewer cases of missed detection occur. The ZL+Faster R-CNN method fits the actual aircraft target better than the BBox calculated by SimCLR+Faster R-CNN, and the four-point coordinate error is smaller.

Table 4. Experimental results based on the RSOD data set.

Method	*AP*30	*AR*30	*AP*50	*AR*50
Faster R-CNN	0.6378	0.7194	0.4432	0.4989
SimCLR+Faster R-CNN	0.6384	0.7013	0.4801	0.5216
ZL+Faster R-CNN	0.6908	0.7404	0.5327	0.5528

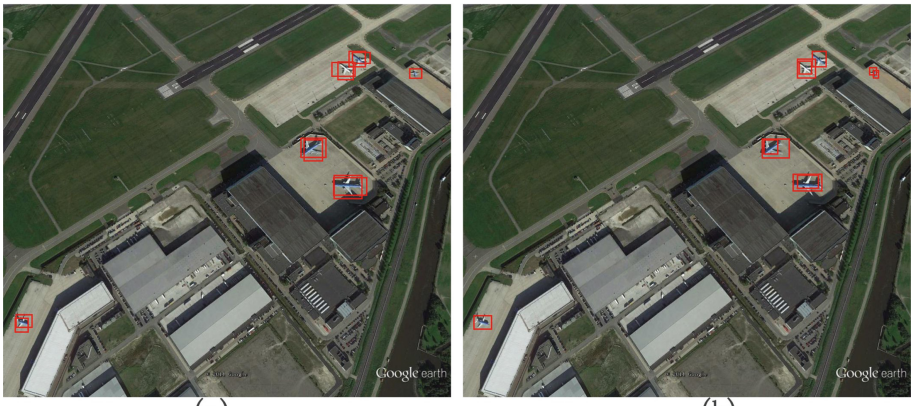

Fig. 5. Image processing results in SimCLR+Faster R-CNN and ZL+Faster R-CNN at IOU = 0.3 ((a) the result in SimCLR+Faster R-CNN (b) the result in SimCLR+Faster R-CNN).

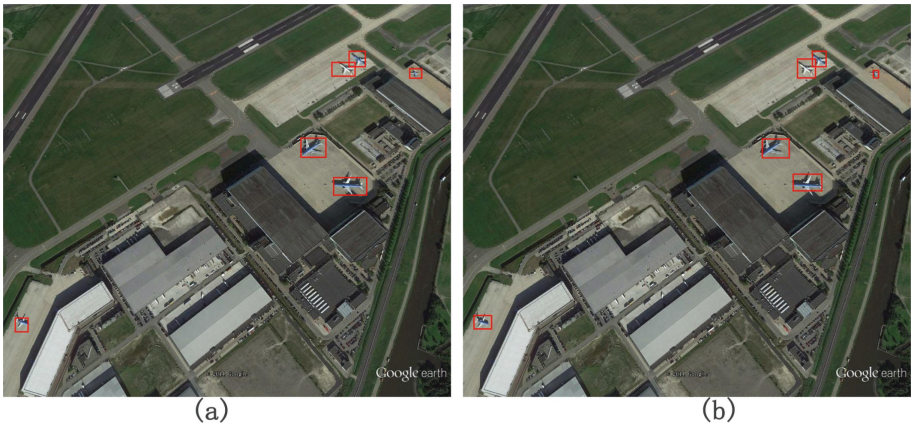

Fig. 6. Image processing results in SimCLR+Faster R-CNN and ZL+Faster R-CNN at IOU = 0.5((a) the result in SimCLR+Faster R-CNN (b) the result in SimCLR+Faster R-CNN).

5 Conclusion

In this paper, self-supervised representation learning is introduced into the field of remote sensing small target detection, based on a new ZL method for aircraft

target feature learning on smaller-scale unlabeled aerial remote sensing data, and the learned features are later used for aircraft target detection tasks. The experimental results show that the method is feasible for aircraft target detection, breaking through the dilemma of training on large-scale labeled remote sensing target data set. Remote sensing images are widely researched and used for their large range and multi-channel information collection, but in practice there is fewer images information available for a certain area, and the labeling of images requires a lot of wasted manpower and material resources, so the research application for small sample unlabeled remote sensing data

is an important research direction to deal with emergencies in the future. Furthermore, remote sensing data provides relatively little information about the side of the object, and the data is affected by deformation, so its efficient accuracy in terms of network design and feature extraction in target detection is an issue of concern specific to the field of remote sensing.

References

1. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. arXiv:1906.00910 (2019)
2. Bojanowski, P., Joulin, A.: Unsupervised learning by predicting noise. In: International Conference on Machine Learning, pp. 517–526. PMLR (2017)
3. Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1734–1747 (2016). https://doi.org/10.1109/TPAMI.2015.24961411
4. Falcon, W., Cho, K.: A framework for contrastive self-supervised learning and designing a new approach. arXiv:2009.00104 (2020)
5. Grill, J.B., et al.: Bootstrap your own latent: a new approach to self-supervised learning. arXiv:2006.07733 (2020)
6. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: International Conference on Machine Learning, pp. 4182–4192. PMLR (2020)
7. Karras, T., Laine, S., Aila, T.: A Style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
8. Oord, A.V., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International Conference on Machine Learning, pp. 1747–1756. PMLR (2016)
9. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016) 1
10. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
11. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv:1804.02767 (2018)
12. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
13. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non- parametric instance-level discrimination arXiv:1805.01978 (2018)
14. Xiao, X., Zhou, Z., Wang, B., Li, L., Miao, L.: Ship Detection under complex backgrounds based on accurate rotated anchor boxes from paired semantic segmentation. Remote Sens. **11**(21), 2506 (2019). https://doi.org/10.3390/rs112125062.1
15. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE Inter- national Conference on Computer Vision. pp. 2223–2232 (2017) 2.1