A TWO-STAGE CONTRASTIVE LEARNING FRAMEWORK FOR IMBALANCED AERIAL SCENE RECOGNITION

Lexing Huang, Senlin Cai, Yihong Zhuang, Changxing Jing, Yue Huang, Xiaotong Tu, Xinghao Ding⋆

[1]School of Informatics, Xiamen University, China
[2]Institue of Artificial Intelligent, Xiamen University
⋆dxh@xmu.edu.cn

ABSTRACT

In real-world scenarios, aerial image datasets are generally class imbalanced, where the majority classes have rich samples, while the minority classes only have a few samples. Such class imbalanced datasets bring great challenges to aerial scene recognition. In this paper, we explore a novel two-stage contrastive learning framework, which aims to take care of representation learning and classifier learning, thereby boosting aerial scene recognition. Specifically, in the representation learning stage, we design a data augmentation policy to improve the potential of contrastive learning according to the characteristics of aerial images. And we employ supervised contrastive learning to learn the association between aerial images of the same scene. In the classification learning stage, we fix the encoder to maintain good representation and use the re-balancing strategy to train a less biased classifier. A variety of experimental results on the imbalanced aerial image datasets show the advantages of the proposed two-stage contrastive learning framework for the imbalanced aerial scene recognition.

Index Terms— Aerial image, class imbalanced, scene recognition, contrastive learning, two-stage framework

1. INTRODUCTION

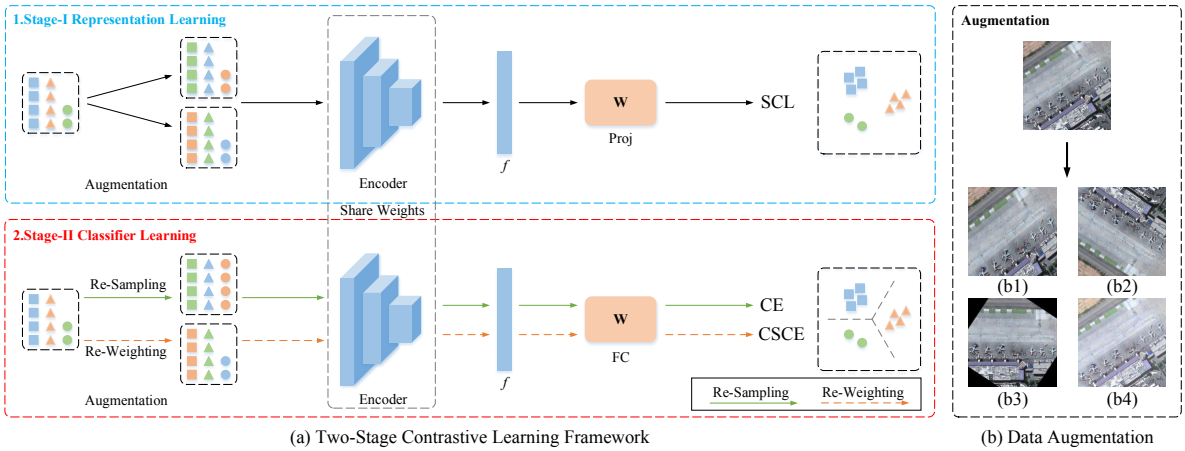
Scene recognition is an essential work for aerial image analysis, which plays a key role in many applications, such as natural disaster monitoring, urban planning, environmental investigation et al. [1, 2, 3]. With the development of observation technology, the amount of aerial images is growing rapidly, which justifies the need for an efficient scene recognition algorithm.

In recent years, convolutional neural networks (CNNs) have shown excellent performance in aerial scene recognition [4, 5]. However, real-world aerial image datasets are always class imbalanced. Specifically, the common categories among the dataset, such as farmlands, forests, and mountains, can be considered as the majority classes with rich samples. The rare categories in the dataset, such as harbors, airports, and railway stations, can be regarded as the minority classes with a few samples. The problem of class imbalance poses a significant challenge to aerial scene recognition, and how to improve the classification accuracy of the minority classes became important.

Most existing work solves the class imbalanced problem by re-balancing strategies, including data re-sampling or loss re-weighting [6, 7]. These methods are expected to increase the weight of the minority classes to reduce the bias of the majority classes. Although re-balancing strategies can improve the performance of the minority classes, they can destroy the original data distribution and damage the learning of representation ability [8]. Concretely, re-sampling usually over-samples the minority classes, thus increasing the risk of over-fitting to the minority classes. For re-weighting, usually change each class weight, which can distort the original data distribution. Therefore, the improvement effect of these methods is limited.

In this work, we propose a two-stage contrastive learning framework, which focuses on both representation learning and classification learning to improve the recognition performance of imbalanced aerial scene images. As shown in Fig. 1 (a), the framework includes the representation learning stage and classification learning stage. The first stage is proposed to learn discriminative feature representations from class imbalanced data. We employ supervised contrastive learning to train the encoder in the stage. In addition, different from natural images, the aerial images are acquired based on imaging platforms, such as satellites, airplanes, etc., we design a data augmentation policy for aerial images to further improve the potential of contrastive learning effectively. Then, to retain the strong representational ability and get a less biased classifier, we keep the encoder fixed and re-sampling or re-weighting is used for the second stage. Abundant experimental results show that the proposed framework outperforms existing state-of-the-art methods.

The study is supported partly by the National Natural Science Foundation of China under Grants 82172033, U19B2031, 61971369, 52105126, China Postdoctoral Science Foundation (No. 2021M702726), Science and Technology Key Project of Fujian Province(No. 2019HZ020009) and Fundamental Research Funds for the Central Universities 20720200003.



(a) Two-Stage Contrastive Learning Framework (b) Data Augmentation

Fig. 1. (a) The framework includes two stages: representation learning and classifier learning. In the representation learning stage, the encoder is trained by contrastive learning. In the classifier learning stage, keep the encoder fixed and re-sampling or re-weighting is used to train a classifier. (b) The augmentation policy consists of random crop and resize (b1), random flip (b2), random rotation (b3), and random color distortion (b4).

2. PROPOSED METHOD

2.1. Overall framework

Fig. 1 (a) provides the two-stage contrastive learning framework for imbalanced aerial scene recognition. Specifically, the framework includes the representation learning stage and classification learning stage. The two stages share the same encoder network to extract features.

2.2. Proposed representation learning stage

The primary goal of the representation learning stage is to learn a feature space reflecting the original information. Given the advantages of contrastive learning in representation learning, we introduce contrastive learning into the representation learning stage. To realize the idea that better features ease classifier learning, we do not introduce any re-sampling or re-weighting methods in the first stage, and only use the original data and ordinary contrastive learning loss function to train the encoder network. This training strategy can be beneficial to learning better representation.

Supervised Contrastive Learning. Compared with self-supervised contrastive learning, supervised contrastive learning (SCL) introduces label information in the training process [9, 10]. Hence, in the representation learning stage, we use the SCL to make the encoder learn the correlation between aerial images of the same scene.

Assume that N samples are picked up randomly in each training iteration, for each sample x_i, we randomly generate two samples through twice data augmentation. As a result, we can obtain $2N$ samples. Then, $2N$ samples are respectively fed into the base encoder network (i.e., ResNet50), and the encoder outputs corresponding feature vector $f_i = ResNet(x_i)$ of each sample. After that, the projection network is used to map the feature f_i into a vector representation $z_i = Proj(x_i) = W^{(2)}\sigma(W^{(1)}f_i)$, where σ is the ReLU function and W is a single linear layer. Concretely, the projection network is implemented by a 2-layer perceptron with one hidden layer. Next, we normalize z_i by using the ℓ_2 norm so that the inner product can be used as distance measurements. Finally, these normalized vectors are used to calculate the supervised contrastive loss, which is defined as:

$$L_{SCL}(z_i) = \frac{-1}{|S(i)|}\sum_{s\in S(i)} log\frac{\exp(z_i \cdot z_s/\tau)}{\sum_{k=1,k\neq i}^{2N}\exp(z_i \cdot z_k/\tau)} \quad (1)$$

where $S(i)$ denotes a set, including the indices of samples of the same label as z_i, $|S(i)|$ is the size of this set, and τ is a temperature parameter.

Data Augmentation. The first step in contrastive learning is data augmentation, which can improve the diversity of data and help SCL to learn generalizable features. We first follow the widely used augmentation policy: random crop and resize [11, 12]. Different from the natural images, aerial images are captured from the bird's-eye view perspective [13] so that the targets in images have different orientations. Rotation invariance is a unique property of aerial images. Therefore, random flip and random rotation are applied to our strategy. And aerial images are easily affected by weather and day-night factors. It is appropriate to simulate this situation with random color distortion [12, 14]. Each augmentation method is depicted in Fig. 1 (b).

2.3. Proposed classifier learning stage

To obtain the classification ability, we introduce the classifier learning stage. Based on the encoder from the previous stage, a fully connected (FC) layer is applied to predict class-wise logits and calculate the cross-entropy loss. Although re-weighting or re-sampling methods may damage the representative ability, they can reduce the bias for the majority classes and generate better decision boundary. Therefore, we keep the encoder fixed, then use the re-weighting or re-sampling methods to further improve the performance for the minority classes. Here, the commonly used re-sampling or re-weighting methods are selected as the training strategy of the classifier.

Class-balanced sampling (CB). All classes are selected with equal probability so that the sample size of each class in the training set is the same [15]. Specifically, we choose a class with equal probability, and then randomly select a sample from this class as the training sample. Assume the number of the class is C, the probability of the i-th sample being selected is as follows:

$$P_{CB}(i) = \frac{1}{C} \times \frac{1}{n_{y_i}} \tag{2}$$

where y_i is the label of i-th sample , n_{y_i} denotes the sample size of y_i-th class.

Cost-sensitive cross-entropy loss (CSCE). As an extension of the cross-entropy loss (CE), the CSCE expects the minority classes to have greater weight [16]. For each sample, the loss is added with an adjusting factor, whose size is inversely proportional to the sample size of the corresponding class. The loss function of i-th sample is defined as:

$$L_{CSCE}(i) = \frac{n_{min}}{n_{y_i}} L_{CE} \tag{3}$$

where $n_{min} = min\{n_i \,|\, 1 \leq i \leq C\}$.

3. EXPERIMENTS

3.1. Dataset

We conduct experiments on artificially created aerial image datasets, including class imbalanced AID (CI-AID) [17] and NWPURESISC45 (CI-NWPURESISC45) [1]. The class imbalanced datasets are generated following the protocol mentioned in [18]. Here, s is the training sample size of each class, $|C_{maj}|$ and $|C_{min}|$ are the number of categories among the majority and minority classes, separately.

CI-AID. The AID dataset contains 30 classes, about from 220 to 420 each class, a total of 10,000 images, whose each pixel size is about 600×600. Here, we divide them into two main classes. The first 20 classes are selected as the majority classes and the remaining 10 classes as the minority classes (i.e., $|C_{maj}| = 20$ and $|C_{min}| = 10$). The training set is set as follows: for each majority class, 50% samples are randomly selected (i.e., $s > 100$), and for each minority class, only choose a few samples (i.e., $s = 5$ or $s = 10$). Then 100 samples from the rest of each class are selected to construct a balanced testing set.

CI-NWPURESISC45. The NWPURESISC45 dataset has a total of 31,500 images, covering 45 classes, each class has 700 images with 256×256 pixels. For this dataset, the first 30 classes are selected as the majority classes and the remaining 15 classes as the minority classes (i.e., $|C_{maj}| = 30$ and $|C_{min}| = 15$). The training set is constructed as follows: for each majority class, randomly select 140 samples (i.e., $s = 140$), and for each minority class, only a few samples are chosen (i.e., $s = 5$ or $s = 10$). Then 400 samples from the rest of each class are selected as the testing set.

3.2. Setup

In this section, we present some key implementation details for experiments on the CI-AID and CI-NWPURESISC45.

Encoder Setup. We use ResNet-50 [19] as the encoder network, and a 2-layer MLP as the projection network. We use SGD [20] with a momentum of 0.9 and weight decay of 1×10^{-4} as optimizer to train the encoder with batch size of 128. For each minority class, when s is 5 or 10, the corresponding training epochs are 500 or 600. The initial learning rate is 0.5, which is adjusted by cosine decay [21]. The temperature τ is set to be 0.1 for the SCL loss functions.

Classifier Setup. We employ an FC layer as the classifier. The classifier is trained for 100 epochs using SGD with the momentum of 0.9 and weight decay of 5×10^{-5}, batch size of 128, the learning rate of 0.01 with cosine decay while keeping the encoder network fixed.

3.3. Main results

We compare the performance of our proposed two-stage contrastive learning framework with other state-of-the-art methods used to address the imbalanced aerial scene recognition. The experimental comparison to some existing work on the CI-AID and CI-NWPU-RESISC45 is reported in Table 1, where SCL-CSCE and SCL-CB denote CSCE or CB adopted in the classifier learning, respectively. The difference is that their proposed methods are Wide ResNet-50 (WResNet50) [22] pre-trained on ImageNet, while ours is based on plain ResNet-50. As can be seen from Table 1, our method outperforms other methods in the minority classes and overall performance. What's more, SCL-CB performs better than SCL-CSCE in total.

3.4. Ablation studies

In this section, we conduct three ablation experiments to present a detailed analysis of our proposed method.

Ablation studies on data augmentation. To explore the effects of data augmentation policy on contrastive learning,

Table 1. Top-1 accuracy rates on the CI-AID and CI-NWPU-RESISC45.

Methods	CI-AID: $\|C_{maj}\| = 20, \|C_{min}\| = 10$						CI-NWPU-RESISC45: $\|C_{maj}\| = 30, \|C_{min}\| = 15$					
	Majority $s > 100$	Minority $s = 10$	Overall -	Majority $s > 100$	Minority $s = 5$	Overall -	Majority $s = 140$	Minority $s = 10$	Overall -	Majority $s = 140$	Minority $s = 5$	Overall -
Plain [22]	94.75	48.60	79.37	94.85	37.00	75.57	89.46	50.75	76.56	90.96	30.38	70.77
Re-Weighting [6]	93.35	56.70	81.13	91.75	47.80	77.10	89.07	55.62	77.92	86.41	39.10	70.64
Re-Sampling [7]	93.50	50.60	79.20	93.10	40.30	75.50	90.04	50.73	76.94	88.67	32.52	69.95
RF-MML-Proto [18]	92.15	63.20	82.50	90.80	54.50	78.67	87.52	61.49	78.73	86.13	50.24	74.11
RF-MML-SVM [18]	92.15	62.70	82.33	91.80	53.09	78.99	88.31	63.86	80.80	87.87	50.99	75.51
SCL-CSCE (ours)	94.10	69.90	86.03	94.10	55.70	81.29	88.66	63.21	80.18	89.56	51.00	76.71
SCL-CB (ours)	94.00	**70.50**	86.16	93.65	**57.90**	81.73	88.67	**64.15**	80.50	89.07	**53.28**	77.14

Table 2. Ablation studies of different data augmentation policies on the CI-NWPURESISC45 ($s = 10$).

Data Augmentation Policy	Majority	Minority	Overall
RCR (Baseline)	78.93	45.23	67.69
RCR + RF	82.40	52.66	72.48
RCR + RF + RR	85.71	58.63	76.68
RCR + RF + RR + RCD	**88.67**	**64.15**	**80.50**

Table 3. Ablation studies of different methods for the encoder on the CI-AID ($s = 10$).

Module	Methods	Majority	Minority	Overall
Encoder	Baseline	94.00	**70.50**	**86.16**
	CS-SCL	93.30	66.40	84.33
	CB	**94.90**	64.20	84.66

we design several groups and conduct a series of comparative experiments based on SCL-CB. Data augmentation policies include random crop and resize (RCR), random flip (RF), random rotation (RR), and random color distort (RCD). Table 2 reports the classification accuracy after the encoder is trained by different data augmentation policies. As can be seen from Table 2, the last group achieves the best recognition performance. In summary, for aerial images, RF, RR, and RCD can help SCL learn generalizable features.

Ablation studies on the encoder. For a better understanding of the influence of re-balancing strategies in representation learning, we adopt various methods for the encoder training. And CB is applied in classifier learning. In Table 3, we give the model recognition accuracy under different methods. CS-SCL is an SCL re-weighting we design, similar to CSCE (i.e., $(n_{min}/n_{yi})L_{SCL}$). As shown in Table 3, the minority classes and overall performance of Baseline are much better than CS-SCL or CB. It shows that good features make good classifications. Specifically, the encoder trained from original data can obtain better representation, then improve

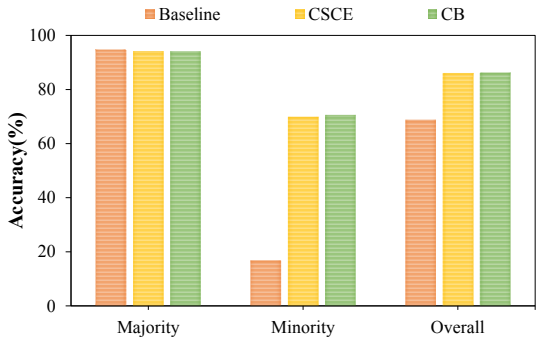


Fig. 2. Ablation studies of different methods for the classifier on the CI-AID ($s = 10$).

the classification performance of the model.

Ablation studies on the classifier. To discuss the performance of re-balancing strategies in classifier learning, several methods are used to train the classifier based on the trained encoder. We present the accuracy of different methods adopted at the classifier learning stage in Fig. 2. As shown in Fig. 2, the model can achieve better performance based on CSCE or CB. The results show that re-balancing strategies can help to shift the classifier decision boundary and greatly boost the model's recognition accuracy.

4. CONCLUSION

In this work, we propose a two-stage contrastive learning framework to solve the problem of imbalanced aerial scene recognition. We explore how re-balancing strategies influence representation learning and classifier learning, and reveal that they can promote classifier learning significantly but also to some extent damage representation learning. Moreover, we systematically study the data augmentation policy of aerial images and show the effects of different design choices. Extensive experiments proved that our framework can achieve the most advanced classification performance.

5. REFERENCES

[1] Gong Cheng, Junwei Han, and Xiaoqiang Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[2] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340–4354, 2020.

[3] Hongguang Li, Yang Shi, Baochang Zhang, and Yufeng Wang, "Superpixel-based feature for aerial image scene recognition," *Sensors*, vol. 18, no. 1, pp. 156, 2018.

[4] Ying Li, Haokui Zhang, Xizhe Xue, Yenan Jiang, and Qiang Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, pp. e1264, 2018.

[5] Jiayun Wang, Patrick Virtue, and Stella X Yu, "Successive embedding and classification loss for aerial image classification," *arXiv preprint arXiv:1712.01511*, 2017.

[6] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[7] Li Shen, Zhouchen Lin, and Qingming Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 467–482.

[8] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9719–9728.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[13] Youtian Lin, Pengming Feng, Jian Guan, Wenwu Wang, and Jonathon Chambers, "Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection," *arXiv preprint arXiv:1912.00969*, 2019.

[14] Andrew G Howard, "Some improvements on deep convolutional neural network based image classification," *arXiv preprint arXiv:1312.5402*, 2013.

[15] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2019.

[16] Nathalie Japkowicz and Shaju Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[17] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[18] Jian Guan, Jiabei Liu, Jianguo Sun, Pengming Feng, Tong Shuai, and Wenwu Wang, "Meta metric learning for highly imbalanced aerial scene classification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4047–4051.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] Léon Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

[21] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[22] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.