A Two-Stage Shake-Shake Network for Long-Tailed Recognition of SAR Aerial View Objects

Gongzhe Li[*], Linpeng Pan[*],Linwei Qiu[*], Zhiwen Tan, Fengying Xie, Haopeng Zhang[†]
Beihang University
{gzli20,linpengpan,qiulinwei,tanzhiwen,xfy_73,zhanghaopeng}@buaa.edu.cn

Abstract

Synthetic Aperture Radar (SAR) has received more attention due to its complementary superiority on capturing significant information in the remote sensing area. However, for an Aerial View Object Classification (AVOC) task, SAR images still suffer from the long-tailed distribution of the aerial view objects. This disparity limit the performance of classification methods, especially for the data-sensitive deep learning models. In this paper, we propose a two-stage shake-shake network to tackle the long-tailed learning problem. Specifically, it decouples the learning procedure into the representation learning stage and the classification learning stage. Moreover, we apply the test time augmentation (TTA) and the classification with alternating normalization (CAN) to improve the accuracy. In the PBVS [1] 2022 Multi-modal Aerial View Object Classification Challenge Track 1, our method achieves 21.82% *and* 27.97% *accuracy in the development phase and testing phase respectively, which wins the top-tier among all the participants.*

1. Introduction

Synthetic aperture radar (SAR) is an active earth observation system, which can be installed on aircraft, satellites, spacecraft and other flight platforms [5]. It can generate high-resolution radar frequency (RF) images under low visibility and various scenarios [20]. Comparing to electro-optical (EO) sensors, SAR can effectively identify camouflage and penetrate shelter during the whole day. Therefore, the usage of image dataset obtained by SAR received progressive attention [33], such as parameter estimation [16], object detection [3, 13], classification [17, 21, 27], *etc.* The motivition of our work is to investigate a more effective and efficient method using SAR images to improve the classification accuracy for AVOC(Aerial View Object Classification). This task requires predicting the class label of an aerial low-resolution image based on a number of prior examples of images and their class labels.

There are two major problems need to be addressed in this AVOC competition, especially when using the given SAR dataset. Firstly, the class distribution of the dataset is typically long-tailed.It means that samples of head-class occupy the vast majority of whole dataset, while the samples of tail-class is negligible compared to the head-class. (Table 1). Consequently, neural networks trained on this imbalanced dataset face overfitting problems, tending to classify all test samples as head-classes and ignore tail-classes. The second obstacle of this challenge is low image quality and high noise interference, as shown in Fig. 1. This perturbation increases the difficulty of feature extraction by neural network and causes the experimental result sensitive to the pre-processing method.

To alleviate overfitting problem, we adopt three essential methods which greatly improve the accuracy of classification results. Firstly, a novel two-branch ResNet [11] with shake-shake regularization [7] is selected to be baseline architecture. This simple ResNet variant is specifically proposed to overcome overfitting problem by two random coefficients. Second, two-stage training strategy [15] is performed to decouple the learning procedure into the representation learning stage and the classification learning stage for data of the long-tailed distribution. In the representation learning stage, neural network is trained on given long-tailed dataset as usual. For the classification learning stage, the model is retrained on a class-balanced dataset by freezing parameters of backbone and only updating parameters of classifier. This training procedure greatly mitigate the phenomenon that classification results are dominant by head-class. Finally, after obtaining test results, the test time augmentation (TTA [23]) is applied to counteract low quality and high noise interference problem. By rotating, flip-

[*] The first three authors contributed equally.

[†] Corresponding Author.

[1] https://codalab.lisn.upsaclay.fr/competitions/1388.

This work supported in part by the National Key Research and Development Program of China (Grant No. 2019YFC1510905) and the Fundamental Research Funds for the Central Universities.

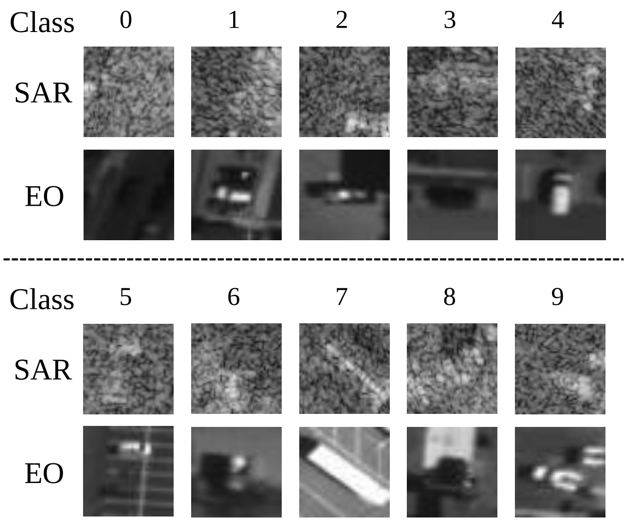

Figure 1. Examples of SAR and EO images with ten class indexes from the given dataset.

ping and scaling, a more abundant dataset is generated to enrich information and reduce specific pattern bias to some extent. It should be noticed that image sharpening, edge enhancement and other image enhancement algorithms are not employed to prevent the introduction of new noise. Moreover, a post-processing approach called the classification with alternating normalization (CAN) [14] is also employed to improve the classification performance by re-adjusting prediction results with prior of the dataset.

Based on these improvement methods mentioned above, we test our model on the Track 1 SAR imagery dataset. We achieve 21.82% in the validation set and obtain 27.97% on the final test data. Adequate experiments are performed to show the effectiveness of our approach. In summary, the main contributions of this paper are summarized as follows,

- A two-stage training strategy is employed on a lightweight shake-shake network to mitigate the overfitting problem of the long-tailed dataset.
- The test time augmentation and a post-processing approach are applied to balance results with prior of the dataset.
- We have achieved the top-tier accuracy both in the development phase and testing phase among all the teams.

The rest is planned as follows. In Section 2, related works are presented. In Section 3, our proposed method is introduced in detail. In Section 4, we evaluate our method on the Track 1 SAR imagery dataset of the PBVS 2022 Challenge. Finally, we conclude our study about this competition in Section 5. Our code is available at https://github.com/LinpengPan/PBVS2022-Multi-modal-AVOC-Challenge-Track1.

2. Related Works

Recently, long-tailed recognition has attracted lots of attention in the field. We briefly review previous methods on long-tailed recognition. These methods can be divided into three categories [31]: data distribution re-balancing, transfer learning, and decoupled learning.

2.1. Data Distribution Re-balancing

Data distribution re-balancing consists of re-sampling and re-weighting. Re-sampling methods are to make the class distribution more balanced. It includes oversampling [2, 9, 24] for minority class and undersampling [6, 10] for majority class or learning to sample [22]. Re-weighting approaches are to re-weight the loss functions [19]. These series of methods [1, 4] assign minority category instances more costs which are always misclassified or not confident. However, all of these methods sacrifice the accuracy of the head to compensate for the tail.

2.2. Transfer Leaning from Head to Tail class

To transfer knowledge from head to tail class is another branch [18, 26]. Transfer-learning based methods address the issue of imbalanced training data by transferring features learned from head classes with abundant training instances to under-represented tail classes. Recent some works include transferring the intra-class variance [28] and transferring semantic deep features [18]. However, it is usually a non-trivial task to design a specific model for feature transfer.

2.3. Decoupled Learning

Recently some works [15, 32] show that the distribution of datasets have no impact on representation learning of networks. Therefore, decoupling the representation and classifier learning improves the performance on long-tailed datasets significantly. We also make use of this core into our model.

3. Methods

3.1. Overall Framework

The distribution of this PBVS 2022 Multi-modal Aerial View Object Classification Challenge Datasets is long-tailed and the percent of class "0" is close to 80% (described in Section 4.1). To tackle this severely long-tailed dataset, special attention is paid on designing the proposed deep learning strategy.

The overall framework of our proposed method can be separated into three components, the shake-shake regularization, two-stage training strategy and testing strategy.

Since SAR images are of low-resolution and complex architectures tend to be overfitting, we introduce a lightweight ResNet backbone with shake-shake regularization to alleviate the over-fitting problem. To mitigate the problem of long-tailed distribution, a two-stage training strategy is used to decouple the learning procedure into the representation learning stage and the classification learning stage. At the test phase, we use the test time augmentation (TTA [23]) and a post-processing approach to improve the accuracy.

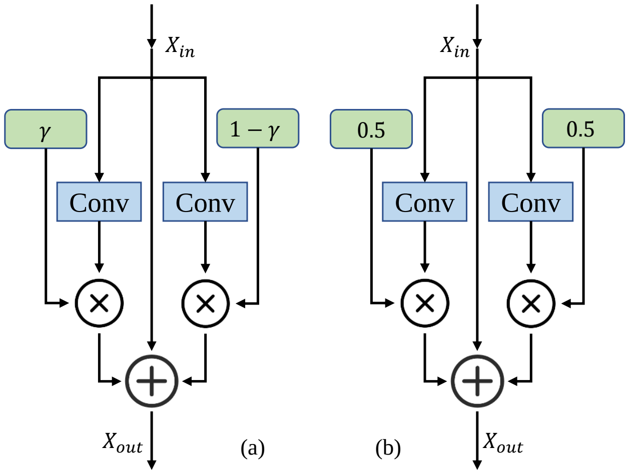

Figure 2. Shake-shake regularization. **(a):** Training pass. **(b):** At test time.

3.2. Shake-Shake Regularization

Inspired by data augmentation, shake-shake regularization [7] was proposed to augment the internal representations. We use X_{in} and X_{out} to denote the inputs of a residual block and the output of one residual block respectively. Let $f(\boldsymbol{\theta}_1)$ and $f(\boldsymbol{\theta}_2)$ represent two convolution units where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the weights. A residual block with two branches can be described as the following

$$X_{out} = X_{in} + f(X_{in}; \boldsymbol{\theta}_1) + f(X_{in}; \boldsymbol{\theta}_2). \quad (1)$$

Let γ denote a random variable between 0 and 1 which is sampled from a uniform distribution. The residual block with shake-shake regularization can be described by

$$X_{out} = X_{in} + \gamma f(X_{in}; \boldsymbol{\theta}_1) + (1-\gamma) f(X_{in}; \boldsymbol{\theta}_2). \quad (2)$$

The training procedure is shown in Fig. 2. In the forward training pass, γ is sampled to obtain the output of the residual block. Then in the backward training pass, another random number denoted as η is sampled to calculate the gradients, which can be seen as a form of gradient augmentation. Finally, at the test phase, the scaling coefficient value is set to 0.5 following the same logic as Dropout [25].

3.3. Two-Stage Training Strategy

The long-tailed distribution of the SAR images causes a great challenge to the classification methods based on deep learning. However, there still are some interesting findings [15]:

1) High quality representation can be learned with the long-tailed datasets;

2) Strong long-tailed recognition ability can be obtained by adjusting only the classifier with the class-balanced datasets.

Inspired by this, we introduce a two-stage training strategy which decouple the learning procedure into representation learning stage and the classification learning stage, as shown in Fig. 3. For the training phase, we train the shake-shake model with the complete dataset to learn the feature representation, then we freeze the parameters of the feature extractor and only train the classifier with the class-balanced dataset. We construct this class-balanced dataset from the given data without no extra images, which will be discussed in Section 4.4.

3.4. Testing Strategy

Data Augmentation is the process of randomly applying some operations (*e.g.* rotation, crop, flips) to the input data. By this mean, a model cannot see the same example twice and has to learn more general features about the classes it has to recognize. Test Time Augmentation [23] is to perform random modifications to the test images. For TTA, instead of showing the regular, “clean” images, only once to the trained model, we input it the augmented images several times. We then average the predictions of each corresponding image and take it as our final guess.

Specifically in this paper, given a test sample, we can get m different samples by augmenting it. Then we can get m probability distributions with the trained shake-shake model and obtain the final probability distribution by averaging them. More details can be seen in Section 4.5.

3.5. Classification with Alternating Normalization

Applying data augmentation at test time can achieve better performance in general. However, due to long-tail constraint, the model tends to predict the sample as the head class. In this case, the effect of TTA is not satisfactory. To solve this problem, we introduce a non-parametric post-processing approach, called the classification with alternating normalization (CAN [14]).

Ideally, the prediction category distribution should be same as the prior distribution. But the actual prediction may deviate this assumption and we can correct it. As show in Fig. 4, the principle behind CAN is to make the category distribution of the whole prediction examples closer to the

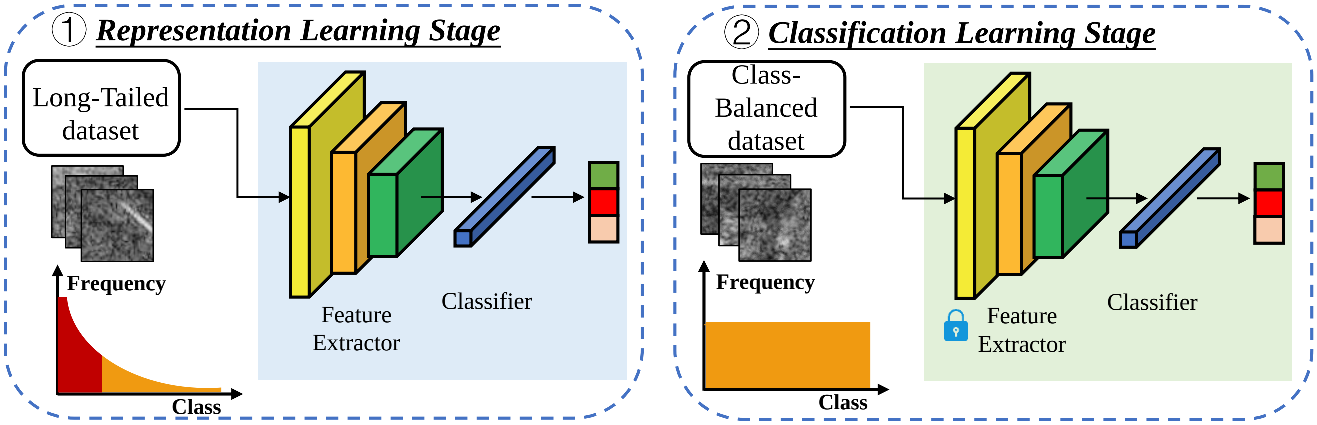

Figure 3. Two-stage training strategy. For representation learning stage, we train the whole model. For classification learning stage, we train the classifier and fix the other structures.

prior distribution by adjusting the prediction distribution of the challenging examples (low-confidence examples). CAN adjusts those low-confidence results based on the prior category distribution to improve the overall accuracy. In detail, it consists of two main steps.

1) Example Division

We need to divide the examples into high-confidence examples and low-confidence examples by computing the top-k entropy of its category probability distribution in this step.

First, the normalized probability distribution of the top-k probability values $\hat{\mathbf{p}}^i$ is

$$\hat{\mathbf{p}}^i = \frac{[p_1^i, p_2^i, \dots, p_k^i]}{\sum\limits_{i=1}^{k} p_j^i}, \tag{3}$$

where p_j^i represent the prediction probability of j-th ($j = 1, \cdots, M$) category of i-th ($i = 1, \cdots, N$) example. Then the top-k entropy $H_{top-k}(\hat{\mathbf{p}}^i)$ can be calculated by following

$$H_{top-k}(\hat{\mathbf{p}}^i) = -\sum_{j=1}^{k} \hat{p}_j^i \log(\hat{p}_j^i). \tag{4}$$

Finally, we take $H_{top-k}(\hat{\mathbf{p}}^i)/\log k$ as the final metric, which has been normalized to $[0, 1]$. We can set threshold τ to divide the samples by the final metric. Obviously, an instance with high entropy is hard to recognize and has low-confidence. As seen in Fig. 4, sample 1 is a high-confidence example while sample 2 is an instance with low-confidence.

2) Alternating Normalization In this step, the class probability distribution of the non-confidence examples is adjusted by alternating normalization.

Ideally, the predicted category distribution is equal to the prior category distribution $\tilde{\mathbf{p}}$ (assume we have known the prior category distribution)

$$\tilde{\mathbf{p}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{p}^{(i)}. \tag{5}$$

Without loss of generality, assume there are n samples $\{1, 2, ..., n\}$ with high-confidence, and then the remaining $\{n+1, n+2, ..., N\}$ belong to low-confidence examples. We consider high-confidence examples to be more reliable and use them to correct low-confidence examples.

Given any low-confidence example $s \in \{n+1, n+2, ..., N\}$, we can obtain a new set $\{1, 2, ..., n, s\}$. To make the this new set maintain the prior distribution, we perform the first normalization with $\mathbf{p}^s$ and high-confidence probability $\mathbf{p}^1, \mathbf{p}^2, ..., \mathbf{p}^n$

$$\mathbf{p}^k = \frac{\mathbf{p}^k \cdot \tilde{\mathbf{p}}}{\bar{\mathbf{p}}}, k = 1, 2, ..., n, s, \tag{6}$$

where the operators are element-wise and the mean probability $\bar{\mathbf{p}}$ is

$$\bar{\mathbf{p}} = \frac{1}{n+1}(\mathbf{p}^s + \sum_{i=1}^{n} \mathbf{p}^i). \tag{7}$$

However, after Eq. 6, the sum of $\mathbf{p}^k$ may be not equal to 1, so we need to employ the second normalization

$$\hat{\mathbf{p}}^k = \frac{[p_1^k, p_2^k, ..., p_M^k]}{\sum\limits_{j=1}^{M} p_j^k}. \tag{8}$$

Eventually, we only keep $\hat{\mathbf{p}}^s$ as the adjusted probability of sample s and discard the rest. In other word, the probability of high-confidence examples is fixed and we only update the probability of low-confidence examples. More details can be found in the original paper [14]. As shown in Fig. 4, the probability of sample 2 has been changed by the influence of sample 1 with high-confidence and the prior distribution.

4. Experiments

4.1. Datasets

For PBVS @ CVPR 2022 Multi-modal Aerial View Object Classification Challenge Track 1, there are only one type of images, which are captured by synthetic aperture radar (SAR) sensors from the aerial view. SAR images vary from 50×50 to 60×60 pixels. Since the SAR image sizes are not consistent in the dataset, we resize all samples to 56×56.

Table 1. Class Distribution of the PBVS 2022 training data. We define the first four sample-rich classes as the head classes for their domination. The rest classes are called tail classes.

Class Index	Type	Samples (#)	Percent (%)
0	sedan	234,429	79.72
1	SUV	28,089	9.56
2	pickup truck	15,301	5.21
3	van	10,655	3.63
4	box truck	1,741	0.59
5	motorcycle	852	0.29
6	flatbed truck	828	0.28
7	bus	624	0.21
8	pickup truck w/ trailer	840	0.29
9	flatbed truck w/ trailer	633	0.22

The objects in images belong to a list of 10 classes corresponding to a training set with non-uniformly distributed number of samples per class, whereas the validation set and test set is based on a small uniformly distributed number of samples per class. In other words, this is an extremely imbalanced dataset with a long-tailed distribution, as shown in Table 1. In this challenge, the ground-truth labels of the validation and test sets are not public, only the final performance scores are visible to the participants.

4.2. Network Architecture

The ResNet with shake-shake regularization is proposed to mitigate the overfitting problem which achieve the state-of-the-art performance on the CIFAR100 dataset in 2017 [7]. We evaluate the performance of several common classification models on this long-tailed dataset and the shake-shake-26 achieve the best accuracy. From the Table 2, we can see the ResNet-50 achieves the best performance in ResNet family while the some advanced and complex networks do not achieve better performance. It shows that these advanced networks may not work well in long-tailed dataset since they are inclined to overfit the head classes.

Table 2. Classification performance on different models. Shake-shake-26 model is the best choice considering both the computation and performance.

Model	Parameter	Top-1 Accuracy
ResNet-34 [11]	21.8M	14.16%
ResNet-50 [11]	25.5M	14.68%
ResNet-101 [11]	44.5M	14.29%
ResNet-152 [11]	60.2M	14.16%
WRN-28-10 [29]	36.5M	14.58%
DenseNet-100 [12]	27.2M	14.16%
PyramidNet-110 [8]	28.3M	13.77%
shake-shake-26 2x32d [7]	**2.9M**	**18.83%**
shake-shake-26 2x96d [7]	26.2M	14.68%

4.3. Implementation details

We use shake-shake-26 as our baseline model and train it from scratch. Specifically, we train 100 epochs for the representation learning stage and 10 epochs for the classification learning stage. We set mini-batch size to 128 and train our model with SGD optimizer. We set the momentum to 0.9. The cosine annealing learning strategy is applied to adjust the learning rate. The initial learning rate is set as 0.1 with a weight decay $1e-4$ for the first stage and the initial learning rate is set as 0.01 with the same weight decay for the second stage. The models are all trained by using the cross-entropy loss function with the mixup [30]. All models are built on the Pytorch framework and trained with two NVIDIA 2080Ti GPUs.

Table 3. The impact of sample size on classification learning stage. 'w/o' denotes removing TTA and CAN.

Method	# Samples	Accuracy	
		w/o	w
Baseline	-	18.83%	20.91%
Two-stage	500	20.26%	20.39%
	5000	19.87%	**21.82%**
	6000	20.39%	20.52%

4.4. Train Strategy

A two stage train strategy is parlayed to improve the performance on long-tailed dataset. We build a class-balanced dataset by random selecting from the extended dataset. The class-balanced dataset is only used to train the classifier to alleviate overfitting of the long-tailed distribution. As shown in Table 3, the classification learning stage can effectively improve the performance. When 500 samples are used on this stage, the accuracy achieves 20.26%, which is 7% higher than the baseline. However, the accuracy does not improve with more samples used. For 5000 and 6000

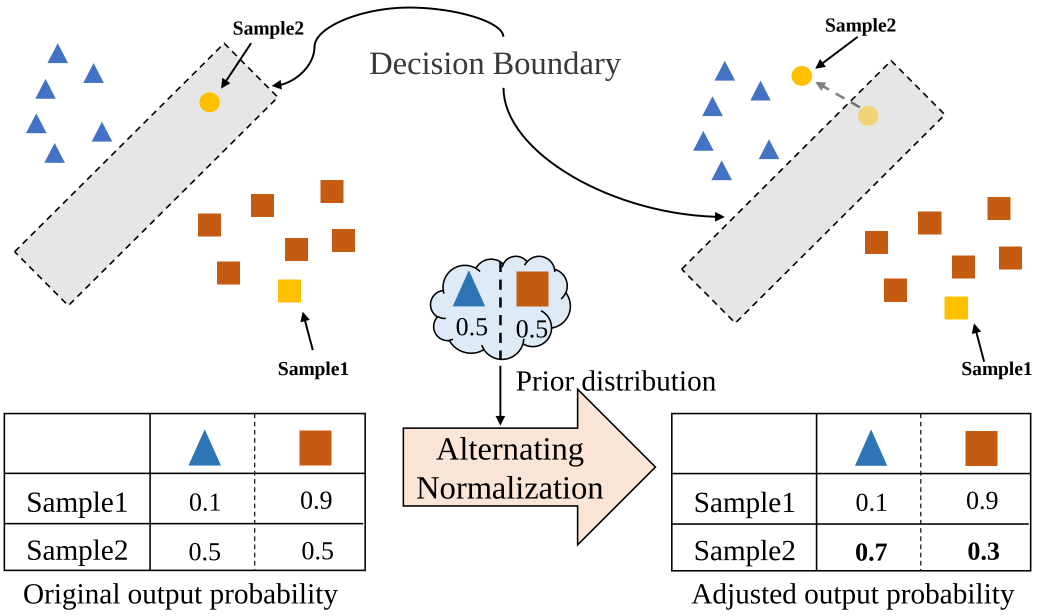

Figure 4. The illustration of Classification with Alternating Normalization (CAN). The prior distribution for two classes is $\{0.5, 0.5\}$. CAN adjusts the predict probability distribution based on prior distribution and produces a less ambiguous prediction.

samples, the accuracy is 19.87% and 20.39% respectively. Considering the additional improvement of post-processing methods, we finally choose to use 5000 samples for the classification learning stage.

Table 4. The impact of the number m of augmented samples in TTA on the accuracy. CAN contributes to the final results as well.

m of TTA	Development Phase		Test Phase
	w/o CAN	CAN	
0	19.87%	20.26%	-
4	20.65%	21.23%	-
8	**21.82%**	21.56%	26.63%
12	21.56%	**21.82%**	**27.97%**

4.5. Investigation on the parameter of TTA

We introduce two post-processing methods to further the improve classification performance. We first employ TTA to obtain the most reliable results from multiple results by random rotation, flip, brightness transformation, blur and affine transformation. Then we introduce CAN to improve the final accuracy. Table 4 shows the performance improvement in development phase and test phase, respectively. The TTA promotes the reliable predictions from several transformed version of a given input as m becomes large. However, the increasing computation entails the growth of m. Besides, CAN is always conducive to obtain more reliable results except $m = 8$. According to the performance in the test phase, we decide the integrated Testing Strategy with TTA ($m = 12$) and CAN.

4.6. Ablation Study

We mainly compare the effectiveness of each method in Table 5. It proves the benefits of adding different method to the network. Every improvement we introduce improves the final accuracy and after comprehensive evaluation we achieve the 21.82% top-1 accuracy by combining these methods.

Table 5. Ablation study on the development phase for Track 1.

Method					
Baseline	✓	✓	✓	✓	✓
Two-stage		✓	✓	✓	✓
TTA			✓		✓
CAN				✓	✓
Accuracy	18.83%	19.87%	21.56%	20.26%	**21.82%**

4.7. Competition Results

The top-5 teams from preliminary results in the test phase and corresponding development results are listed in terms of top-1 accuracy in Table 6. Note that the only difference between development and test phase is the change of provided data for competitions. We achieved the best during the development phase. The top ranking team in test phase behaved badly in the development phase but ranked the top in the final test phase. This may indicate that there is

Table 6. Preliminary results in test phase and corresponding in development phase.

Team	Top-1 Accuracy	
	Development Phase	Test Phase
TeamA	-	**36.44%**
TeamB	16.62%	31.23%
TeamC	13.51%	28.09%
Ours	**21.82%**	27.97%
TeamD	-	26.76%

a large gap between the validation set and the test set, which makes some methods poorly generalized.

5. Conclusion

In this paper, a two-stage shake-shake network is designed towards the long-tailed distribution dataset of PBVS 2022 Multi-modal Aerial View Object Classification Challenge Track 1. For the training phase, we train the model with the complete dataset to learn the feature representation, then we freeze the parameters of the feature extractor and only train the classifier with the class-balanced dataset. Combining with the test time augmentation (TTA) and a post-processing approach CAN, our proposed method achieves the top-tier ccuracy in the development phase and behaves well in the testing phase.

References

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[3] Sizhe Chen and Haipeng Wang. Sar target recognition based on deep learning. *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pages 541–547, 2014. 1

[4] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2

[5] Ian G Cumming and Frank H Wong. Digital processing of synthetic aperture radar data. *Artech house*, 1(3):111–231, 2005. 1

[6] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8, 2003. 2

[7] Xavier Gastaldi. Shake-shake regularization of 3-branch residual networks. In *International Conference on Learning Representations*, 2017. 1, 3, 5

[8] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5927–5935, 2017. 5

[9] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 2

[10] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 2

[11] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 5

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[13] Zhongling Huang, Zongxu Pan, and Bin Lei. What, where, and how to transfer in sar target recognition based on deep cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 58:2324–2336, 2020. 1

[14] Menglin Jia, Austin Reiter, Ser Nam Lim, Yoav Artzi, and Claire Cardie. When in doubt: Improving classification performance with alternating normalization. In *EMNLP*, 2021. 2, 3, 4

[15] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2019. 1, 2, 3

[16] Juha Karvonen. Baltic sea ice concentration estimation from c-band dual-polarized sar imagery by image segmentation and convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 1

[17] Jerrick Liu, Nathan Inkawhich, Oliver Nina, and Radu Timofte. Ntire 2021 multi-modal aerial view object classification challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2021. 1

[18] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2

[19] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. 2

[20] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P. Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine*, 1:6–43, 2013. 1

[21] Hemani Parikh, Samir B. Patel, and Vibha Patel. Classification of sar and polsar images using deep learning: a review.

International Journal of Image and Data Fusion, 11:1 – 32, 2019. 1

[22] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020. 2

[23] Divya Shanmugam, Davis W. Blalock, Guha Balakrishnan, and John V. Guttag. When and why test-time augmentation works. *ArXiv*, abs/2011.11156, 2020. 1, 3

[24] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. 2

[25] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014. 3

[26] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[27] Cheng-Yen Yang, Hung-Min Hsu, Jiarui Cai, and Jenq-Neng Hwang. Long-tailed recognition of sar aerial view objects by cascading and paralleling experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 142–148, 2021. 1

[28] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019. 2

[29] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5

[30] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018. 5

[31] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *ArXiv*, abs/2110.04596, 2021. 2

[32] 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*, pages 9719–9728, 2020. 2

[33] Xiaoxiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liang pei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5:8–36, 2017. 1