Scene Clustering Based Pseudo-labeling Strategy for Multi-modal Aerial View Object Classification

Jun Yu[1] , Hao Chang[*1], Keda Lu[12], Liwen Zhang[1], Shenshen Du[1], Zhong Zhang[3]

[1]University of Science and Technology of China,[2] Ping An Technology Co., Ltd,
[3]Hefei ZhanDa Intelligence Technology Co., Ltd

harryjun@ustc.edu.cn, changhaoustc@mail.ustc.edu.cn, wujiekd666@gmail.com,
zlw1113@mail.ustc.edu.cn, nibility163@163.com, zhangzhong@zalend.com

Abstract

Multi-modal aerial view object classification (MAVOC) in Automatic target recognition (ATR), although an important and challenging problem, has been under studied. This paper firstly finds that fine-grained data, class imbalance and various shooting conditions preclude the representational ability of general image classification. Moreover, the MAVOC dataset has scene aggregation characteristics. By exploiting these properties, we propose Scene Clustering Based Pseudo-labeling Strategy (SCP-Label), a simple yet effective method to employ in post-processing. The SCP-Label brings greater accuracy by assigning the same label to objects within the same scene while also mitigating bias and confusion with model ensembles. Its performance surpasses the official baseline by a large margin of +20.57% Accuracy on Track 1 (SAR), and +31.86% Accuracy on Track 2 (SAR+EO), demonstrating the potential of SCP-Label as post-processing. Finally, we **win the championship both on Track1 and Track2** *in the CVPR 2022 Perception Beyond the Visible Spectrum (PBVS) Workshop MAVOC Challenge [35]. Our code is available at* https: //github.com/HowieChangchn/SCP-Label.

1. Introduction

With the advent of research on computer vision, the performance of ATR has witnessed incredible progress. Synthetic aperture radar (SAR) and electro-optical (EO), two popular modalities for modern remote sensing (RS) systems, have the advantage of capturing visible light features and ignoring climate effects, respectively. Although each individual data modality has advantages and disadvantages, traditional RS systems typically leverage only a single modality. Therefore, ATR algorithms that utilize mul-
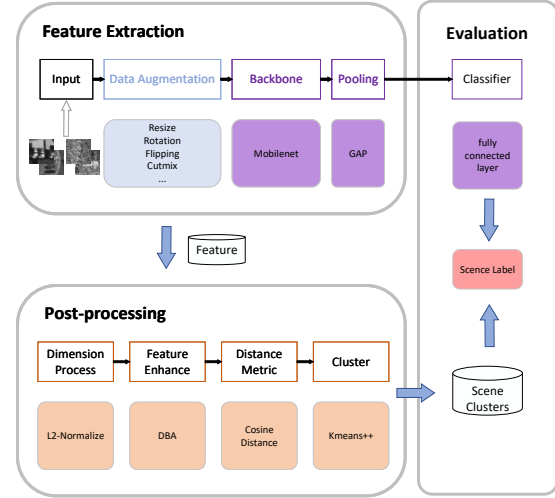
---

[*]Corresponding author.

Figure 1. **SCP-Label for MAVOC.** We utilize a convergent model in the supervised training step for feature extraction, followed by a pseudo-label strategy based on scene clustering for post-processing.

tiple data modalities may be able to mitigate the disadvantages associated with each sensor type.

A lot of research has been done on ATR for both EO and SAR images. In terms of EO-ATR, a large number of related research results have been achieved due to the use of early traditional visual methods [3] and current convolutional neural networks (CNNs) [34] to extract features. In terms of SAR-ATR, the acquisition of SAR image data is difficult, and related research is lagging behind. Recently, there are also some CNNs structures designed for SAR images, such as fuzzy logic and dynamic neural networks [37], and end-to-end CNNs [9] for SAR images. But most research directions focus on their respective fields, i.e. ATR development for SAR and EO are disjoint.

In contrast, ATR on multiple data modalities has been under studied. In fact, the area for aerial view images poses further challenges in ATR with lower target resolution, different target texture and light reflectance, high inter-class similarity, etc. In the general ATR algorithm, the image is extracted by CNNs and classified by a classifier. However, if the identified objects are fine-grained and affected by the factors mentioned above, and what is worse, the dataset has a long-tailed distribution, the classification results will be confused and biased towards the majority class. Both class imbalance and fine-grained multi-modal data has not been thoroughly evaluated in most existing ATR algorithms.

As a preliminary of our work, we investigate the performance of general image classification algorithm in the context of class-imbalanced and fine-grained MAVOC dataset. Specifically, to figure out how algorithms perform, we manually divide the training set and calculate the bias and confusion of the model in each category. By conducting validation experiments, we find that resampling helps balance the classification performance of the model and observes confusing categories in the dataset. More importantly, our method is motivated by the further observation that, despite the aforementioned facts, the MAVOC dataset has scene aggregation characteristics. The model will recognize objects in the same scene as different categories due to the interference of various factors.

Therefore, in this paper, for exhaustively improving the recognition performance of aerial view images in ATR, we propose SCP-Label for post-processing, as illustrated in Fig.1. To avoid the model being biased towards the majority classes, the proposed method leverages the re-sampled data as model input. In addition, in order to improve the fine-grained learning ability of the model, we utilize a backbone with an attention mechanism. Rather than directly assigning labels to each image, we use a clustering algorithm to cluster scenes, and then assign scene labels through model integration instead.

In the experimental phase, we determine our chosen feature extraction backbone, data augmentation strategy, and under-sampling data volume through multiple experimental settings, and set up multiple ablation experiments to determine our key parameters. Both on Track 1 (SAR) and Track 2 (SAR+EO) of the CVPR 2022 PBVS Workshop MAVOC Challenge [35], our Top-1 accuracy rate in the final test phase reach **36.44%** and **51.09%**, respectively. Experimental results show that our method **performs the best among all participants**.

Our contributions can be summarized as follows:

- We address the challenge of MAVOC by proposing a new pseudo-labeling strategy that can aggregate images of the same scene and assign pseudo-labels based on scene clustering.

- We propose a multi-modal integration scheme to generate pseudo-labels based on scene clustering. Compared with general pseudo-labeling, this design allows the model to have better generalization ability under different illumination, shooting angles and other disturbances.

- We achieve superior performance over the state-of-the-art approaches, and win first place on two tracks of the MAVOC challenge.

2. Related work

EO is the most popular sensor type in modern RS systems. It can efficiently capture images in the visible spectrum that are human-interpretable, and data collection for EO is fairly easy (for example, aerial images can be collected in the EO domain using the Google Earth [14] platform). In the past, ATR has been a lot of research work in the field of EO [28, 31, 40]. For the classification of EO images, traditional visual methods [3] were used to extract texture features to classify ships at sea. Katherine [34] used a VGG-16-based CNNs for detection and classification of EO images of ships at sea.

On the other hand, SAR data is also one of the sensor patterns for other commonly used ATR algorithms. Although less well understood, SAR has the advantage of operating at night and in different weather conditions, which gives it a distinct advantage over EO sensors in some applications. However, the acquisition process of SAR data is much more complex and costly, which means that the breadth of SAR-ATR studies lags behind EO-ATR studies in several ways [10, 24–26].

With the increasing application of deep learning in SAR, terrain surface classification [33], resolution inversion [38], speckle removal [39], specific methods in interferometric SAR (InSAR) applications [1], SAR-optical data fusion [23], etc. have attracted extensive attention. Tzeng et al. [37] proposed a method for classifying SAR images using fuzzy logic and dynamic neural networks, which was an early attempt to classify SAR images using neural networks. Popular deep learning algorithms used end-to-end CNNs as feature extractors to generate discriminative features that can work with subsequent classifiers automatically [9]. Geng et al. [16] proposed a single polarimetric/supervised SAR image classification system to improve the problems of noise and speckle in the data that are not easy to characterize effectively.

There have been many outstanding studies of class imbalanced data in recent years. Important adjustment methods are mainly divided into re-sampling [5, 6] and re-weighting [7, 12]. Re-sampling includes majority class [15, 30] under-sampling or minority class [8, 17] over-sampling, which aims to rebalance the distribution of data. How-

ever, this approach may lead to overfit to the minority class while losing some valuable training samples, thereby deteriorating the modeling quality on large-scale datasets. Re-sampling weights, assigning higher weights to tails [12] or hard samples [7] in the loss function, makes the expectation of each class closer to the test distribution. However, when applying re-weighting to large-scale real-world scenarios, it is often difficult to optimize. In addition, there are some other ways to mitigate the impact of model representation on long-tailed data, such as FocalLoss [29].

3. How CNNs perform on MAVOC

In this section, we attempt to investigate the misclassified and biased behavior of CNNs on MAVOC dataset. Many existing state-of-the-art image classification algorithms are trained on balanced and coarse-grained high-quality benchmarks, such as ImageNet [13], Cifar [27], etc. Furthermore, these benchmarks are all visible light images and the recognized objects generally have clean backgrounds, which is profitable for the model to capture features such as color and texture. In contrast, aerial view images present a unique set of challenges due to the fine-grained image features and long-tailed data distribution. In addition, other specific issues in aerial view images also pose further challenges such as the nature of the target, the ratio of the target size to the background of the image, lower resolution of the target, target texture, light reflectance, etc. Since the image properties of the two data domains are significantly different, we propose a conjecture on MAVOC dataset, the image classification model will be confused and biased due to high interclass similarity, skewed class distribution, which may further result in severe performance on validation and test set.

Instead of extending the protocol, which utilizes various class-imbalanced ratios to produce long-tailed versions of benchmark datasets, such as Cifar, MAVOC dataset has been split into train, validation and test set. In order to justify our conjecture, we observe and analyze the performance of official baseline, implemented with Resnet50 [19]. Concretely, we manually divide the training set and analyze the classification of the baseline on the artificial validation set.

Firstly, we manually divide the training set. Specifically, we sample 60 images from each category as the artificial validation set to keep the division result similar to the official validation set, and utilize the official baseline for training. As illustrated in Fig.2. we can find that the long-tailed distribution of the training set causes the model to be biased towards the majority class both on SAR and EO set. Therefore, in order to solve the serious data imbalance, we re-sample 3000 images in each category for training. The accuracy rates before and after resampling show that resampling can effectively alleviate the data imbalance.

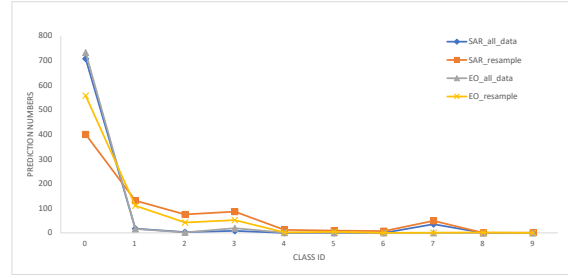Crucially, we find that the MAVOC dataset has scene ag-

Figure 2. The baseline is biased towards the majority class. Resampling will alleviate the class imbalance problem.

gregation characteristics, that is, the same target in the same scene exists multiple images with different sensors, shooting angles, blur degrees, and illumination. Due to the influence of the above multiple distractions, objects in the same scene are identified as different categories. As shown in Fig.3, in a scene we selected, obviously, the results of all the images should be the same, whereas the performance of the classifier is not the case. Through observation, we can find that under the action of different factors, the model's focus is not the same. Therefore, if we can identify targets in the same scene as the same category through corresponding data augments or model integration methods, the classification accuracy will improve over baseline by a large margin.

Besides, the MAVOC challenge is not a multi-label classification task, although in some of the images, multiple targets exist at the same time. This finding also poses certain challenges for us, and we have enhanced the model's ability to learn to the corresponding labeled targets by means of data augments such as Cutmix, which mitigate the effect of multiple targets to some extent.

In order to solve the above problems, we introduce pseudo-labeling strategy based on scene clustering (SCP-Label) for post-processing.

4. Approach

4.1. Overall framework

As shown in Fig.4, our SCP-Label consists of two training steps, supervised training step and generating pseudo-labels step.

4.1.1 Supervised training step

In this step, the model contains three main components. Concretely, we use MobileNet as our backbone, introduce

id	115892	159537	246449	331412	332518
EO					
Pre_label	2	2	2	0	0

id	107094	110148	265529	365772	463195
SAR					
Pre_label	3	2	3	3	6

Figure 3. Scene aggregation properties in EO and SAR images.

SeNet as our attention module, and utilize global average pooling (GAP) for feature aggregation.

MobileNet [21] as a general backbone usually ignores the subtle but discriminative features in fine-grained recognition, hence a network structure with an attention mechanism is necessary. In order to satisfy the needs of fine-grained recognition, SeNet [22] can effectively capture the fine-grained features rather than original structure.

Considering the labeled set, let x denote a training sample with its corresponding label $y \in 1, 2, \cdots, C$ for a C-class recognition task. For class-imbalanced distribution, we separately apply under-sampling and over-sampling for the majority and minority classes, respectively.

After sampling, two samples (x_s, y_s) and (x_e, y_e) are obtained as the input data, where (x_s, y_s) is from the SAR domain and (x_e, y_e) is from the EO domain. Then, the sampled data from the two domains are loaded into backbone respectively, and by GAP the feature vectors $\mathbf{f}_s \in \mathbb{R}^D$ and $\mathbf{f}_e \in \mathbb{R}^D$ can be acquired.

During the training phase of the classifier, the predicted output $z \in \mathbb{R}^C$ is illustrated as

$$z = \mathbf{W}_s^\top \mathbf{f}_s \quad or \quad \mathbf{W}_e^\top \mathbf{f}_e \tag{1}$$

Then softmax function will utilize each component in z, i.e., $[z_1, z_2, \cdots, z_C]^\top$, to calculate the probability for each category $i \in \{1, 2, \cdots, C\}$ by

$$\hat{\mathbf{p}}_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_i}} \tag{2}$$

Generally, we denote the output probability distribution as $\hat{\mathbf{p}} = [\hat{p_1}, \hat{p_2}, \cdots, \hat{p_C}]^\top$, $E(\cdot, \cdot)$ as the cross-entropy loss function. Therefore, our supervised training step generates a weighted cross-entropy recognition loss, which is formulated as

$$L = E(\hat{\mathbf{p}}, y_s) \quad or \quad E(\hat{\mathbf{p}}, y_e) \tag{3}$$

4.1.2 Generating pseudo-labels step

As observed in Sec.3, the same target in the same scene will be identified as different categories due to various factors. Therefore, instead of per image strategy of pseudo label, we propose pseudo-labeling strategy based on scene clustering.

The converged model in the supervised training step will serve as a pre-trained model for generating pseudo-labels step. The feature vector after feature extraction by backbone and GAP will undergo L2-normalize and feature enhancement. As a commonly used feature enhancement strategy in retrieval, DataBase-side Augmentation (DBA) [2] particularly helps improve feature quality, where every feature in the database is replaced with a weighted sum of the point's own value and those of its top k nearest neighbors (k-NN). Moreover, feature enhancement can further strengthen the scene aggregation characteristics and alleviate the intra-cluster distance.

As an improved algorithm of Kmeans [18] clustering, Kmeans++ [4] has the advantages in speed and accuracy, which is used as the clustering algorithm for our scene clustering. By specifying the initial number of clusters, we cluster the enhanced feature vectors and calculate the sum of squared distances within each cluster, discarding the clusters with poor quality.

Then, we assign cluster labels using a multi-model ensemble approach. The single model exploits the mode of the predicted labels of all images in the cluster as the cluster label. However, the model is prone to be affected by model bias and inter-class confusion, so we utilize one-by-one comparison for multi-model ensemble. Specifically, if the cluster labels of model_1 and model_2 are consistent, the label is used as the final cluster label, and the model ensemble is terminated. Otherwise it continues to compare model_2 and model_3. If the model ensemble does not terminate early, the comparison is made until the last model, and the cluster label of the last model is used as the final cluster label.

SCP-Label, as pseudo-label, will be applied directly to the validation set or test set as a post-processing method, so the black arrow "pseudo label" connects the top and the bottom parts. Since our pseudo label is only done as a post-processing and not as a semi-supervised learning, the label of each cluster will be directly output as a category without any subsequent task.

4.2. Usage of SCP-Label

SCP-Label can be used in two situations. On the one hand, it can be employed as a post-processing method for the test set, through which erroneous class labels within clusters can be corrected. On the other hand, it can be used as a pseudo-label strategy under the semi-supervised learning framework to achieve end-to-end learning, such as FixMatch [36].
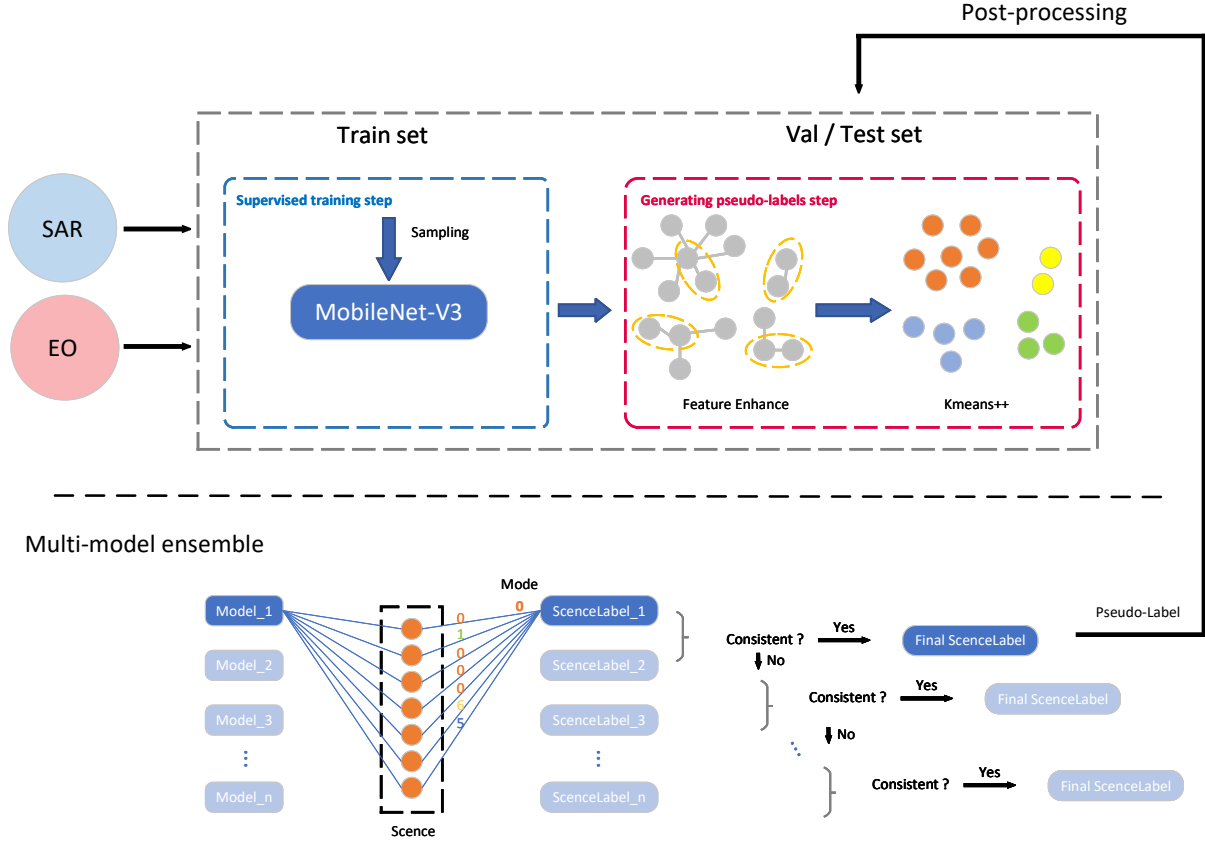
4

Figure 4. **Overall framework for SCP-Label.** The overall framework is divided into two steps, namely the supervised training step and the generating pseudo-labels stage. By extracting features with the pre-trained model, we utilize the DBA feature-enhanced features for scene clustering. And our model ensemble method leverages a strategy of comparing votes one by one.

5. Experiments

In this section, we first introduce the dataset and implementation details, and then quantitatively evaluate the performance of our method on the dataset. Finally, some ablation studies are performed to demonstrate the effectiveness of each component of the network. On both Track 1 (SAR) and Track 2 (SAR+EO) of CVPR 2022 PBVS Workshop MAVOC Challenge, our Top-1 Accuracy ranked 1st place in the final leaderboard.

5.1. Dataset

The data for this challenge consists of two types of small window regions (chips) generated from large images captured by several aircraft-mounted EO and SAR sensors. An EO chip is a 31×31 px image. A SAR chip covers the same approximate field of view as the corresponding EO image and has a finer resolution than the EO image. Chips vary in pixel size due to SAR processing, but are typically around 55×55 px. Fig. 5 provides the samples of EO and SAR chips. The target belongs to a list of 10 classes that correspond to a training set with a non-uniform distribution of the number of samples per class, while the validation and test sets are based on a small number of uniformly distributed samples per class.

The dataset is divided into:

- Training set: The data in this set is seriously imbalanced. (i.e. some classes have more samples than others)

- Validation set: This set is a uniformly distributed among all classes with ≤ 100 samples per class.

- Test set: This split resembles the validation test with a uniform distribution of testing images among the classes.

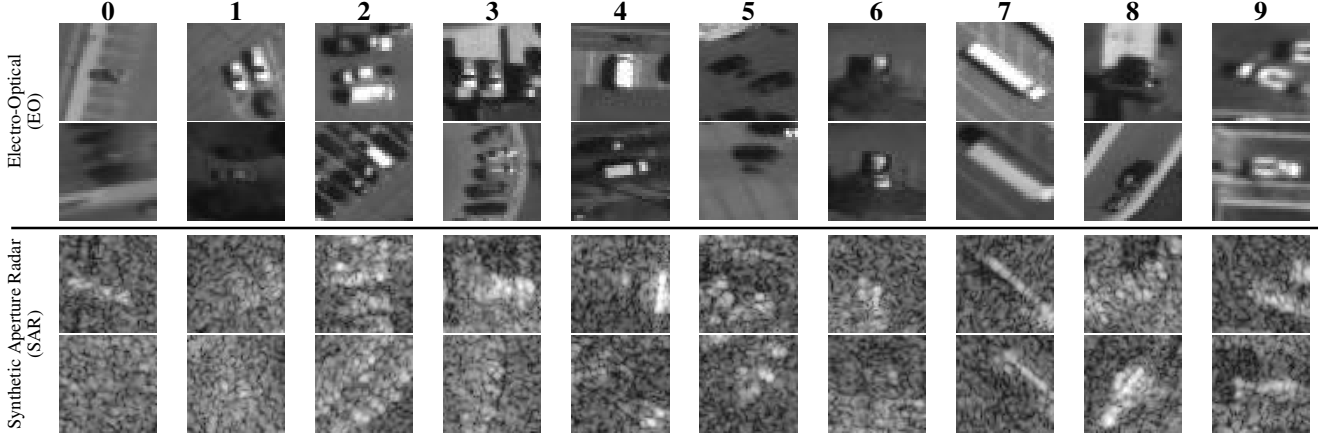These images belong to one of ten categories: 0 to 9.

Figure 5. Two sample pairs of EO and SAR chips from each of the 10 classes in the Unicorn Dataset.

Table 1. Details of the Unicorn Dataset used in this challenge (counts represent the number of (EO, SAR) pairs).

Class	Class Name	Train	Val	Test
0	sedan	234,209	-	-
1	suv	28,089	-	-
2	pickup truck	15,301	-	-
3	van	10,655	-	-
4	box truck	1,741	-	-
5	motorcycle	852	-	-
6	flatbed truck	828	-	-
7	bus	624	-	-
8	pickup truck with trailer	840	-	-
9	flatbed truck with trailer	633	-	-
	Total	293,772	770	826

The train data contains SAR and EO images and class labels. Validation / test data consists of SAR images from Track 1 (SAR) as well as SAR and EO images from Track 2 (SAR+EO) of the challenge. The goal is to use the provided (SAR+EO) training images to maximize classification accuracy when the input is only SAR images or SAR and EO images. Tab. 1 shows a breakdown of images for each category. Note that now it's not a class name, but simply labeled 0-9.

5.2. Implementation details

For the methods we proposed, we under-sample the classes with data volume greater than 1741 to 1741. We use the stratified sampling to divide the dataset on development phase, randomly selecting 70% of each category as the training set and 30% as the validation set, so as to adjust training strategies and parameters offline. But on testing phase, we no longer partition data sets. In the stage of data preprocessing, we resize all EO image to 104×104. While

Table 2. Details of the 7 models we designed for Track 1 and 2. Model 1, 2 and 3 are for Track 1, and model 4, 5, 6 and 7 are for Track 2.

Models	Type	Input classes	Epochs	Top-1 Acc (%)
1	SAR	[0,1,2,3,4,5,6,7,8,9]	3	29.78
2	SAR	[4,5,6,7,8,9]	20	-
3	SAR	[6,7,8,9]	30	-
4	EO	[0,1,2,3,4,5,6,7,8,9]	80	35.47
5	EO	[1,2,3,4,5,6,7,8,9]	70	-
6	EO	[4,5,6,7,8,9]	70	-
7	SAR	[4,6,8,9]	20	-

Table 3. The effect of EO images of different sizes on the accuracy on validation set.

Size	Top-1 Accuracy	Size	Top-1 Accuracy (%)
32×32	15.19	104×104	**23.77**
56×56	19.61	112×112	17.53
64×64	22.34	120×120	18.05
96×96	22.21	128×128	17.27

the resolution of SAR image remain unchanged, we use 0 to fill its edge to 64×64. During the training, we randomly crop all SAR image to 56×56, augment the data by Rotation and Flipping and use Cutmix [41] to reduce noise from the input image.

As shown in Tab. 2, for the serious class imbalance problem, we design a total of 7 models, including three for Track 1 (1, 2, 3) and four for Track 2 (4, 5, 6, 7). Their differences are mainly reflected in training data and training epoch. We adopt MobileNetV3-large [20] as feature extractor. In addition, we use an SGD optimizer with 0.9 momentum and $1e^{-3}$ weight decay, as well as fix learning rate $1e^{-3}$. Our model is implemented using PyTorch with 1 NVIDIA A100 GPU. Experiments are also conducted using MindSpore.

Table 4. Validation results for the CVPR 2022 PBVS Workshop MAVOC Challenge Track 1 with different setting.

Method	Backbone	Under-sample	Augmentation Method					Top-1 Accuracy (%)
			Rotation	Flipping	RA	TA	Cutmix	
Single model	Resnet50	all data	✓	✓	-	-	-	16.10
			-	-	✓	-	-	15.84
			-	-	-	✓	-	16.10
			✓	✓	-	-	✓	**17.10**
	Efficientnet-b1	all data	✓	✓	-	-	✓	15.88
	Swin-Transformer	all data	✓	✓	-	-	✓	16.88
	DenseNet161	all data	✓	✓	-	-	✓	18.05
	MobileNetV3-large	all data	✓	✓	-	-	✓	**18.18**
		4000	✓	✓	-	-	✓	19.35
		2000	✓	✓	-	-	✓	19.74
		1741	✓	✓	-	-	✓	**21.30**

Table 5. Latest update of the test results for the PBVS 2022 Multi-modal Aerial View Object Classification Challenge Track 1 (SAR) and Track 2 (SAR+EO).

Rank	Top-1 Acc of Track 1 (%)	Top-1 Acc of Track 2 (%)
1st	**36.44** (Ours)	**51.09** (Ours)
2nd	31.23	46.85
3rd	28.09	41.77
4th	27.97	37.65
5th	27.48	34.26

Table 6. The effect of different K values on the Kmeans clustering effect.

K	Inertias	Silhouette	Calinski-Harabaz Index
20	227	0.25	48
40	167	0.23	37
60	136	0.25	32
80	113	0.26	30
100	98	0.26	27
120	84	0.26	27
140	74	0.27	26

In terms of feature enhancement, the k1 coefficient of DBA is 1. In the Kmeans++ algorithm for scene clustering, we choose the K value to be 80.

5.3. Main results

Tab. 4 shows the experimental results on validation set of Track 1 (SAR), including the results from different backbone and data augmentation strategies. It can be seen that when MobileNetV3-large is used to extract features and the number of under-samples is 1741, the data augmentation effect is the best when when using Rotation, Flipping and Cutmix. In addition, we also introduce mainstream data augmentation methods, such as RandAugment(RA) [11] and TrivialAugment(TA) [32], but they do not perform well in this dataset, and we think that the use of too many augmentation operations cause the image to lose a lot of original information.

Since EO and SAR images are matched and the category distribution is consistent, we also apply the same backbone, under-sampling data amount and data augmentation to the EO dataset. In addition, since the size of EO images will affect the feature effect of CNNs extraction, we conduct tests on EO with different sizes, as shown in Tab. 3. When the input size of the EO image is 104×104, the test effect on the validation set is the best, which is nearly 9% higher than the original input size of 32×32.

The highest score of our single model (the result of testing on the test set) is shown in Tab. 2, where a single model Top-1 Accuracy using only SAR image data can reach 29.78%, while using only EO image data The Top-1 Accuracy of the first model can reach 35.47%, and the two Top-1 Accuracies can achieve the top 5 results in Track 1 (SAR) and Track 2 (SAR+EO) respectively. Compared with the best results of a single model, after using our multi-model ensemble SCP-Label method, we improve the Top-1 Accuracy by 6.66% and 15.62% on the two tracks, respectively, and **achieve the state-of-the-art results on both tracks**, as shown in Tab. 5.

The top 5 teams from the latest update of the test results for the CVPR 2022 PBVS Workshop MAVOC Challenge Track 1 (SAR) and Track 2 (SAR+EO) are listed in terms of the top-1 accuracy in the testing phase in Tab. 5. It can be seen that we won the first place on both tracks. On both tracks, our accuracy is 36.44% and 51.09% respectively, far ahead of the second place.

5.4. Ablation studies

Kmeans++ is one of the most commonly used methods in clustering, which calculates the best class attribution based on the similarity of point-to-point distances. Among

Table 7. The influence of the k1 coefficient of different DBAs on the clustering effect.

k1	Inertias	Silhouette	Calinski-Harabaz Index
1	39	0.50	93
2	53	0.45	68
3	73	0.38	48

them, the choice of K value is particularly important. We use a variety of clustering-related evaluation indicators to choose the appropriate K value, including Inertias (it represents the sum of the distances from the sample to the nearest cluster center. The smaller its value, the better, indicating that the distribution of samples between classes is more concentrated) , Silhouette (the higher the value, the better, the closer the distance between samples of the same class and the farther the distance between samples of different classes), and the Calinski-Harabaz index (for K clusters, the Calinski-Harabazindex defined as the ratio of between-group dispersion to within-group dispersion, the larger the score, the better the clustering effect). As shown in Tab. 6, we record the feedback for each metric after trying Kmeans clustering with different values of K. Through the feedback of the results of the three indicators and our attempts in this challenge, we determin that the K value of Kmeans++ is 80.

DBA is the feature enhancement on the database side, which only uses the attributes of the adjacent images on the database side to expand the features of the image itself, thereby enhancing the features. After we have determined the K value of the clustering, we can further improve the clustering effect by enhancing the features. As shown in Tab. 7, we record the effect of the DBA k1 coefficient on various metrics of Kmeans++. Compared with k1 = 2 or k1 = 3, when k1 = 1, the smaller the Inertias, the larger the Silhouette and Calinski-Harabaz Index, indicating that the kmeans++ clustering effect is better. Therefore, we determine the k1 coefficient of DBA to be 1.

6. Conclusion

In this work, we propose a multi-model ensemble pseudo-label strategy based on scene clustering, named SCP-Label, to improve the Top-1 accuracy of predicting SAR and EO test set images. The motivation for this strategy is that we observe that this dataset not only has the problem of extreme imbalance of samples between classes, but also has the problem of fine grain within classes, and has scene aggregation characteristics. SCP-Label performs scene clustering on the test set, and then uses multiple models trained from different angles to label each sample in each cluster with pseudo-labels, and finally obtains the cluster label of each cluster, and cluster labels are mapped to each

sample in the cluster to obtain the final output. Our method **win champions on both Track 1 (SAR) and Track 2 (SAR+EO)** of the CVPR 2022 PBVS Workshop MAVOC Challenge.

7. Acknowledgement

References

[1] Nantheera Anantrasirichai, Juliet Biggs, Fabien Albino, Paul Hill, and David Bull. Application of machine learning to classification of volcanic deformation in routinely generated insar data. *Journal of Geophysical Research: Solid Earth*, 123(8):6592–6606, 2018. 2

[2] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2911–2918. IEEE, 2012. 4

[3] Virginia Fernandez Arguedas. Texture-based vessel classifier for electro-optical satellite imagery. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3866–3870. IEEE, 2015. 1, 2

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 4

[5] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 2

[6] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019. 2

[7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019. 2, 3

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

[9] Sizhe Chen and Haipeng Wang. Sar target recognition based on deep learning. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pages 541–547. IEEE, 2014. 1, 2

[10] Sizhe Chen, Haipeng Wang, Feng Xu, and Ya-Qiu Jin. Target classification using the deep convolutional networks for sar images. *IEEE transactions on geoscience and remote sensing*, 54(8):4806–4817, 2016. 2

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 7

[12] 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, 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[14] Google Earth. https://earth.google.com/web/. 2

[15] Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern recognition*, 46(12):3460–3471, 2013. 2

[16] Jie Geng, Hongyu Wang, Jianchao Fan, and Xiaorui Ma. Deep supervised and contractive neural network for sar image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4):2442–2459, 2017. 2

[17] 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

[18] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. 4

[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*, pages 770–778, 2016. 3

[20] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 6

[21] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[23] Lloyd H Hughes, Michael Schmitt, Lichao Mou, Yuanyuan Wang, and Xiao Xiang Zhu. Identifying corresponding patches in sar and optical images with a pseudo-siamese cnn. *IEEE Geoscience and Remote Sensing Letters*, 15(5):784–788, 2018. 2

[24] Nathan Inkawhich, Eric Davis, Uttam Majumder, Chris Capraro, and Yiran Chen. Advanced techniques for robust sar atr: Mitigating noise and phase errors. In *2020 IEEE International Radar Conference (RADAR)*, pages 844–849. IEEE, 2020. 2

[25] Nathan Inkawhich, Matthew J Inkawhich, Eric K Davis, Uttam K Majumder, Erin Tripp, Chris Capraro, and Yiran Chen. Bridging a gap in sar-atr: Training on fully synthetic and testing on measured data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2942–2955, 2021. 2

[26] Nathan A Inkawhich, Eric K Davis, Matthew J Inkawhich, Uttam K Majumder, and Yiran Chen. Training sar-atr models for reliable operation in open-world environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3954–3966, 2021. 2

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[28] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018. 2

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[30] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008. 2

[31] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021. 2

[32] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 774–782, 2021. 7

[33] Hemani Parikh, Samir Patel, and Vibha Patel. Classification of sar and polsar images using deep learning: A review. *International Journal of Image and Data Fusion*, 11(1):1–32, 2020. 2

[34] Katherine Rice. Convolutional neural networks for detection and classification of maritime vessels in electro-optical satellite imagery. Technical report, Naval Postgraduate School Monterey United States, 2018. 1, 2

[35] Low S., Nina O., Sappa A., Blasch E., and Inkawhich N. Multi-modal aerial view object classification challenge results - pbvs 2022. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, June 19-24*, 2022. 1, 2

[36] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 4

[37] Yu-Chang Tzeng and Kun-Shan Chen. A fuzzy neural network to sar image classification. *IEEE Transactions on Geoscience and remote Sensing*, 36(1):301–307, 1998. 1, 2

[38] Lei Wang, K Andrea Scott, Linlin Xu, and David A Clausi. Sea ice concentration estimation during melt from dual-pol

sar scenes using deep convolutional neural networks: A case study. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4524–4533, 2016. 2

[39] Puyang Wang, He Zhang, and Vishal M Patel. Sar image despeckling using a convolutional neural network. *IEEE Signal Processing Letters*, 24(12):1763–1767, 2017. 2

[40] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liang-pei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 2

[41] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 6