2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 978-1-6654-8739-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPRW56347.2022.00050

Multi-modal Aerial View Object Classification Challenge Results - PBVS 2022

Spencer Low
Brigham Young University, Provo, Utah
spencerlow@byu.edu

Oliver Nina
Air Force Research Laboratory, Dayton, OH
oliver.nina.1@afresearchlab.com

Angel D. Sappa
ESPOL Polytechnic University, Ecuador
Computer Vision Center, Spain
sappa@ieee.org

Erik Blasch
Air Force Office of Scientific Research, Arlington, VA
erik.blasch.1@us.af.mil

Nathan Inkawhich
Air Force Research Laboratory, Rome, NY
nathan.inkawhich@us.af.mil

Abstract

This paper details the results and main findings of the second iteration of the Multi-modal Aerial View Object Classification (MAVOC) challenge. The primary goal of both MAVOC challenges is to inspire research into methods for building recognition models that utilize both synthetic aperture radar (SAR) and electro-optical (EO) imagery. Teams are encouraged to develop multi-modal approaches that incorporate complementary information from both domains. While the 2021 challenge showed a proof of concept that both modalities could be used together, the 2022 challenge focuses on the detailed multi-modal methods. The 2022 challenge uses the same UNIfied COincident Optical and Radar for recognitioN (UNICORN) dataset and competition format that was used in 2021. Specifically, the challenge focuses on two tasks, (1) SAR classification and (2) SAR + EO classification. The bulk of this document is dedicated to discussing the top performing methods and describing their performance on our blind test set. Notably, all of the top ten teams outperform a Resnet-18 baseline. For SAR classification, the top team showed a 129% improvement over baseline and an 8% average improvement from the 2021 winner. The top team for SAR + EO classification shows a 165% improvement with a 32% average improvement over 2021.

1. Introduction

The goal of Automatic Target Recognition (ATR) models is to accurately recognize, identify, classify and target signatures within remotely sensed imagery [1, 7, 8, 19, 20, 22]. ATR shares many similarities with object detection and labelling in natural imagery. However, most ATR systems are built on complex remote sensing (RS) systems that are often mounted on aircraft or spacecraft. The unique perspective of these aerial view images can challenge identification and labelling in many ways (e.g., there may only be a handful of pixels on target due to limited sensor resolution) [6]. When the image modalities are different such as EO and SAR, less work has been published due to the challenges of non-collocated sensor collections, association of pixel intensities, as well as different image sizes, ground sampling distance (GSD), and image noise [28]. Specifically, SAR, while researched as a single-mode ATR has the benefit of all-weather, all-time, and stand-off results, it also has challenges with signal multi-bounce, shadows, and discerning boundaries of closely-space objects. Hence, there are many unique research challenges when trying to combine EO and SAR for ATR. The 2022 PBVS Multi-modal Aerial View Object Classification (MAVOC) challenge provides an opportunity to study these complicated issues and provides key insights into how to best leverage multi-modal information in ATR models.

There are many benefits to considering multiple sensor types in RS systems. Some modalities are self-illuminated and do not require sunlight to operate. Some microwave band systems can image through clouds and vegetation. Passive, sub-optical sensors can be used to remotely measure temperature, and active radars can identify man-made objects in jungles, or even infer wind speed. Each RS system is often tuned to specific tasks and are rich in information. Despite these RS systems benefits, they are often overlooked in computer vision applications, as it is diffi-

978-1-6654-8739-9/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPRW56347.2022.00050

cult to combine different sensor's data in complementary ways. For this reason, the majority of RS systems leverage only one modality [9]. Visual data is typically the most common as it is widely collected, human interpretable, and composes multiple spectrum bands. Extensions for multi-spectral (MSI) and hyperspectral (HSI) data includes more spectrum bands, but includes 3D large data cubes, requires determining the salient bands for the targets of interest, and necessitates more computation power than the EO domain. Critically, by intelligently fusing various sensors' data, it is expected to observe ATR performance improvements.

The MAVOC challenges use the UNIfied COincident Optical and Radar for recognitioN (UNICORN) dataset. The UNICORN dataset consists of *aligned* SAR and EO aerial view images. These large images are segmented and hand labelled into a variety of classes. Concretely, the MAVOC challenge is divided into two tracks:

- **Track 1: SAR** - Classifiers in Track 1 have EO and SAR data available at training time, but are only tested on SAR data. Track 1 encourages the development of a maximally accurate SAR classifier that can learn using a combination of SAR and EO images, which would enable classification and use in ATR applications if only SAR images are available during deployment.
- **Track 2: SAR + EO** - Classifiers in Track 2 have EO and SAR data available at training and testing time. Contest 2 encourages the development of a maximally accurate SAR and EO classifier, when both SAR and EO images are present together at collection.

The primary metric for measuring performance in both contests is accuracy on a sequestered test set.

The remainder of this document is organized as follows. Section 2 describes more details on the MAVOC challenge problem, including the track definitions, details of the UNICORN dataset, and the evaluation procedures. Section 3 discusses the main results. Section 4 provides methodological details for the top performing approaches in both tracks. Section 5 gives some analysis of the top methods, and Section 6 offers our conclusions.

2. Challenge

The 2022 MAVOC challenge is held jointly with the Perception Beyond the Visible Spectrum (PBVS) workshop following the 2021 competition which was held in conjunction with the 2021 NTIRE workshop [17]. The MAVOC challenge is designed to facilitate innovative approaches in multi-modal classifiers using pairs of SAR and EO images. The SAR images provide a unique challenge to participants as the SAR images are self-illuminated and coherent, which could result in images with unique *SAR shadows* and a tilted perspective. The MAVOC challenge is divided into two tracks focusing on multi-modal models with different utilities.

2.1. Track 1

Track 1 focuses on building a classifier that can be trained on both SAR and EO data but is tested on only SAR data. A resulting classifier should not be dependent on EO data when deployed, but have potentially learned from the combined features found in both SAR and EO images during training. By removing the dependency on having both modalities available at test time, decisions can be made quicker as the computationally expensive rectification preprocessing that is required to align SAR and EO is unnecessary. The multi-modal nature of the different data for training results in a non-trivial task.

2.2. Track 2

Track 2 focuses on building a classifier that is trained on both SAR and EO data and is tested on (SAR, EO) image pairs. Track 2 creates a scenario where the trained ATR models are able to leverage features in both the SAR and EO images during training *and* deployment. Track 2 is expected to have more accurate classifiers as there is more input information at test time and EO images are generally less noisy than their SAR counterparts.

2.3. Dataset

The dataset on which the challenge is formulated is based on the UNIfied COincident Optical and Radar for recognitioN (UNICORN) dataset [14]. The UNICORN data set is chosen because it is both a public dataset and it is an aligned SAR-EO dataset with hand labelled classes. The 2008 UNICORN dataset consists of Wide Area Motion Imagery (WAMI) large format electro-optical (EO) sensor, and Wide Area Synthetic Aperture Radar (SAR). The data was collected from aircraft flown over Dayton, Ohio. The SAR and EO data cover the same approximate field of view, but the reconstructed SAR image has a finer resolution than the EO image. The large SAR and EO images are rectified and aligned using homography algorithms, as shown in Fig. 2. The competition dataset consists of small windowed sections (chips) that are sub-images of the aligned large image. The EO chips are 31×31 px images. Due to the differing resolutions of the large images, the SAR chips are 55×55 px. Each chip contains one of 10 objects to be classified. Figure 1 displays example (SAR, EO) pairs from each class of the dataset. As shown in Table 1, the dataset is partitioned into train, validation, and test sets. The classes in the train set are non-uniformly distributed and follow a long-tail distribution. Importantly, the validation and test set are uniformly distributed across all 10 classes which allow for a true and unbiased measurement of accuracy.

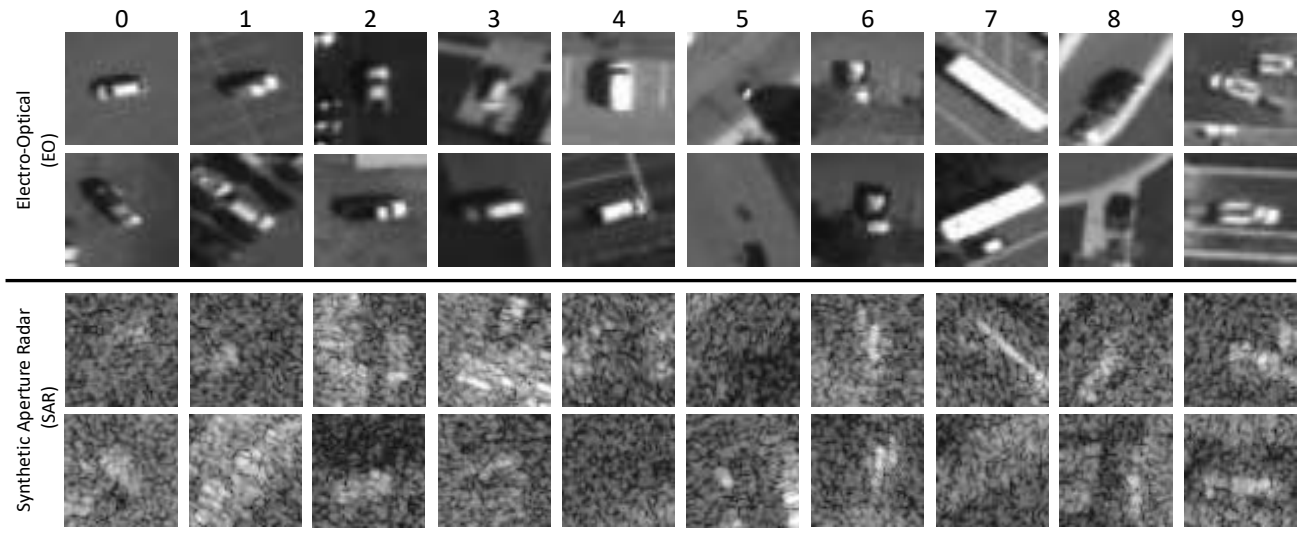

Figure 1. Two sample pairs of EO and SAR chips from each of the 10 classes in the UNICORN Dataset [17].

Figure 2. The aligned scene of the full UNICORN dataset before chipping is performed [14].

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

Class #	Vehicle Type	# Train	# Val	# Test
0	sedan	234,209	77	200
1	SUV	20,089	77	200
2	pickup truck	15,301	77	200
3	van	10,655	77	200
4	box truck	1,741	77	200
5	motorcycle	852	77	200
6	flatbed truck	828	77	200
7	bus	624	77	200
8	pickup truck w/ trailer	840	77	200
9	flatbed truck w/ trailer	633	77	200
	Total	285,772	770	2000

2.4. Evaluation

The scoring uses a top-1% accuracy to determine a classifier's performance. The test set contains 2,000 unlabelled (SAR, EO) chip pairs, with 200 examples for each of the 10 classes. During the testing phase of the competition, teams are allowed up to ten submissions per day. During the evaluation phase, teams submit their label predictions to be evaluated on the competition server. Teams are allowed up to six submissions, which prevents teams from effectively fine-tuning on the test dataset. Accuracy results are made visible during both phases.

2.5. Challenge Phases

The challenge began January 17, 2022, and the test data was released March 10, 2022. The testing phase ended on March 15, 2022 with team submissions finalized.

3. Challenge Results

Eighty-two teams participated in Track 1. Of those 82 participants, 32 teams submitted their algorithms during the development phase, and 33 teams submitted during the testing phase. Track 2 has 77 participants. Of those 77 par-

ticipants, 24 submitted their algorithm during the development phase, and 37 submitted their algorithm during the testing phase. There is both an average performance improvement and a top accuracy improvement when compared to the 2021 NTIRE MAVOC challenge results. Table 2 summarizes the observed improvements of the 2022 PBVS MAVOC challenge over 2021.

Table 2. Various percent improvements of the 2022 PBVS MAVOC challenge over the 2021 NTIRE MAVOC challenge and baselines. The *Baseline* column shows the percent improvement the top scoring team had over the baseline. The *Average of Top 10* column shows the average percent improvement of the top 10 teams from the 2022 challenge over the top 10 teams from the 2021 challenge. The *Top Teams* column shows the percent improvement the top scoring team from the 2022 challenge had over the top scoring team from the 2021 challenge.

Track	Baseline	Average of Top 10	Top Teams
Track 1	+129.57%	+8.59%	+5.26%
Track 2	+165.68%	+32.79%	+9.05%

3.1. Baselines

The baseline classifier utilizes ResNet-18 [3] which has been pre-trained on ImageNet, and fine-tuned on the training and validation sets for 15 epochs. In Track 1, the baseline accuracy was 15.87% and the baseline for Track 2 was 19.23%. Throughout the remainder of this paper, all results may be compared to these values.

3.2. Track 1 SAR Classification Results

For SAR classification, the overall performance increases between the top ten teams from the 2022 PBVS MAVOC challenge when compared with the 2021 NTIRE MAVOC challenge. The percent improvement of the top accuracy score is 5.26%; which is a non-trivial improvement, and can be attributed to newer techniques, and the leading team's novel approach. The test accuracy results of the top 10 teams for SAR classification are show in Table 3 and a breakdown of performance between the 2021 and 2022 MAVOC challenge is shown in Fig. 3.

3.3. Track 2 Results

In SAR + EO classification, we see significant performance improvements when compared with the 2021 NTIRE MAVOC challenge. As shown in Table 2, there is a 9.05% improvement between the two competition's top accuracy scores. We also observed more competitive results from all participating teams. The test accuracy results of the top 10 teams from EO + SAR classification are show in Table 4 and a comparison with 2021 results is shown in Fig. 4.

Table 3. Top-10 Teams for Track 1 (SAR)

Rank	Team	Accuracy
1	**USTC-IAT-United**	**36.44**
2	NLPR-RVG	31.23
3	NYCX	28.09
4	Moyu	27.97
5	Tassel_tzw	27.48
6	mytry1	26.76
7	MEYE	26.63
8	priyakansal	25.67
9	anirudhsikdar3	25.30
10	robot-1	25.06
	baseline	15.87

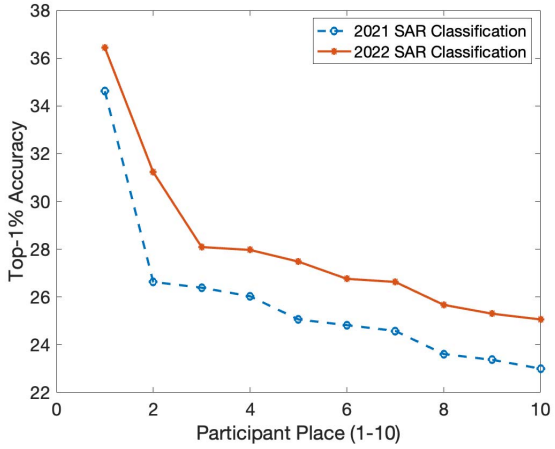

Figure 3. The performance differences between the 2021 NTIRE and 2022 PBVS MAVOC challenges. This figure plots the performance of the ten top performing teams for SAR classification.

4. Challenge Methods

This section briefly summarizes the approaches used by the teams that submitted their models and documentation for prize consideration. Not all teams submitted their methods and are subsequently absent from this paper. We examine the submitted methods from the top teams in each track. This section consists of edited summaries submitted by each team.

4.1. Track 1 - SAR

4.1.1 Rank 1: USTC-IAT-United

Team USTC-IAT-United proposes a novel two stage approach. In stage 1, a supervised training step is conducted based on a combination of SeNet [5] and MobileNet-V3 models [4]. In stage 2, pseudo-labels are generated followed

Table 4. Top-10 Teams for Track 2 (EO + SAR)

Rank	Team	Accuracy
1	**USTC-IAT-United**	**51.09**
2	SunshineMu	46.85
3	SH	41.77
4	NLPR-RVG	37.65
5	sumanth_udupa	34.26
6	jzsherlock	33.17
7	adityakane	30.39
8	hsansui	28.81
9	CiuchitiPoc	28.57
10	mcc	27.85
	baseline	19.23

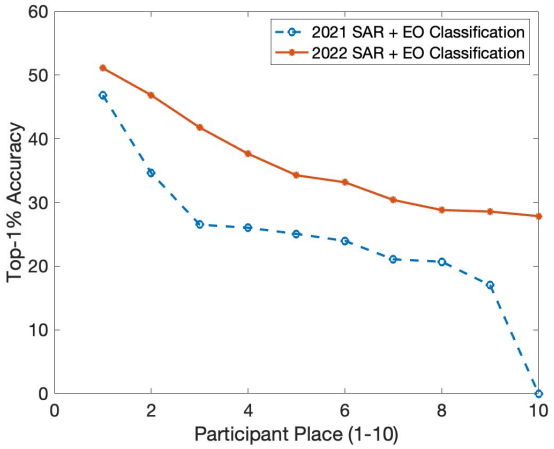

Figure 4. The performance differences between the 2021 NTIRE and 2022 PBVS MAVOC challenges. This figure plots the performance of the ten top performing teams for track 2.

by k-means clustering. SeNet is utilized to adapt to the different image resolutions of the UNICORN dataset. The SeNet is also used to capture the fine-grained detail contained in the images. The team found that there were multiple images that varied in their angle, lighting, and sharpness of the same scene. They exploited the collection of variations for scene clustering, which provided significant improvements to their classifier's performance. The classifier performs feature preprocessing and feature enhancement on the validation set and test set features are extracted by global average pooling, with k-means to cluster pictures of similar scenes into the same cluster, and assigns all pictures in the cluster to the same label. The pseudo-labeling strategy based on scene clustering can be employed as a post-processing method and also suitable for semi-supervised learning.

This USTC-IAT-United team used the same general architecture for both Track 1 and Track 2. The main difference between the two Tracks is in Track 1, the team trains three SAR image classifiers, while in Track 2 they train three EO classifiers and one SAR classifier.

4.1.2 Rank 2: NLPR-RVG

Team NLPR-RVG proposes a method which employs the feature similarity loss and the transductive learning for boosting the performance of SAR aerial view object classification. The team firstly uses a network comprised of a Swin-transformer (Base) [18] and a three-layers multi-layer perceptron (MLP) to predict the class labels from the input SAR images. Considering that the distribution of the training data is unbalanced, they randomly select 500 samples in each class from all the training data for training the network in each epoch and employ on-the-fly data augmentations at the training stages.

For learning more powerful features from the noisy SAR images, the team proposes a feature similarity loss, which is inspired by the supervised contrastive learning [12] and the triplet loss [24] methods. In each training batch, the feature similarity loss pulls together the features from the same class in the feature space and pushes apart the features from different classes. By using the feature similarity loss, the predicted ratio of class 'sedan' on the testing data is reduced from 45.0% to 14.5% after training 100 epochs. Since the testing data is approximately uniformly distributed among the ten classes, it indicates that the feature similarity loss could also alleviate the influence of the unbalanced training data distribution to some extent.

It is noted that the performances of the network are obviously different between the training data and the validation/testing data, which indicates there is a potential domain shift between the training and validation/testing data. Addressing the domain-shift problem, the team trains the network in the transductive setting [21] and a two-stage training strategy is proposed. At the first training stage, the network is trained on the labeled training data. At the second training stage, the network firstly predicts the pseudo labels of the testing data before each epoch. Then, the testing data and the corresponding pseudo labels which have higher confidence in each class are added to the training data in the epoch. The network is trained on both the labeled training data and the selected testing data with the pseudo labels.

The loss function for training the network is a weight combining of the cross-entropy loss L_{CE} and the feature similarity loss L_{sim}, which is formulated as:

$$L = L_{CE} + \lambda L_{sim} \quad , \tag{1}$$

where λ is a preseted weight parameter. The team sets $\lambda = 1$ on the first training stage and set $\lambda = 0.5$ at the second training stage. The similarity loss L_{sim} is formulated as:

$$L_{sim} = \frac{1}{B^2}\sum_{i=1}^{B}\Big[-\sum_{j\in P(i)} F_i^{\top}\cdot F_j + \sum_{k\in N(i)} \max\left(F_i^{\top}\cdot F_k, \epsilon\right)\Big] \tag{2}$$

where B denotes the batch size, F denotes the feature vector extracted by the backbone. $P(i)$ is the positive feature vectors, which are extracted from the images in the same class with that produces F_i, while $N(i)$ is the negative feature vectors. ϵ is a predefined threshold which is set to $\epsilon = 0.1$.

4.1.3 Rank 3: NYCX

Team NYCX uses a two-stage training method to deal with the long-tailed distribution of the data. In the first stage, the team uses the four classes with the most samples (10,000 samples per class) to train a ResNet-50 network with basic feature extraction and discrimination capabilities. The second stage selects 624 samples from each class in the data set (as many as possible) to form a balanced data set. They use the ResNet-50 network trained in the first stage and replaced its classification head with a 10-class MLP to fine-tune on the balanced data set, to get the final classification network.

4.1.4 Rank 4: Moyu

The inspiration for team Moyu's approach comes from [11] and [10]. A two-stage training strategy is used to decouple the learning procedure into the representation learning stage and the classification learning stage. For the training phase, the team trains the Shake-Shake model [2] with the complete dataset to learn the feature representation, then they freeze the parameters of the feature extractor and only train the classifier with the class-balanced dataset. The team chose Shake-Shake32 as the backbone of the feature extractor, which proves to have great performance when facing over-fitting datasets. A fully connected (FC) layer serves as the classifier. The distribution of the given SAR image dataset is long-tailed and the percent of class "0" is close to 80%. To alleviate the over fitting problem caused by too many head-classes, the team first extends the dataset for all classes except the "sedan" class by random rotation, random horizontal flip, and vertical flip. On the first stage, model is trained using the full expanded data (still long-tailed) to get a useful pre-trained weights. However, since there is less information for the tail-classes of the dataset, the model will still be biased towards the head-classes. For the next stage, they establish a class-balanced dataset containing of 50000 images (5000 images per class) by random selecting from the extended dataset. The team freezes the parameters of the feature extractor and only train the FC layer on this class-balanced dataset. The model can achieve higher classification performance on all categories after two stage training strategy. In the test phase, the team first employs the test time augmentation (TTA [25]) while they set the number of images generated by one image to 12 to obtain the classes probability distribution of all the samples. They then use a post-processing approach, called the Classification with Alternating Normalization (CAN [10]), to get the final results. The CAN is a non-parametric post-processing trick and can improve classification accuracy for some challenging examples by re-adjusting category probability distribution using the prior category distributions of dataset.

4.2. Track 2 - SAR + EO

4.2.1 Rank 1: USTC-IAT-United

Team USTC-IAT-United uses the same architecture as described in Section 4.1.1, with a small change. Instead of training three SAR classifiers, they train three EO classifiers and one SAR classifier.

4.2.2 Rank 4: NLPR-RVG

Team NLPR-RVG considers that both the EO images and SAR images are available in the Track2 in both training and testing phases. They propose to ensemble the two trained models from the two kinds of images (i.e., EO images and EO-SAR images). The EO-SAR pairs are made by concatenating the images. Considering the successful applications of transductive learning in many vision tasks, they propose a transductive learning framework for the training of EO images and EO-SAR images, which firstly selects a subset of the unlabeled test samples with relatively higher confidence scores predicted by an inductive baseline model, and then reassigns these selected samples with pseudo labels either based on the K-Nearest Neighbors (K-NN) classification algorithm in the feature space spanned by the feature vectors resulting from a pretrained feature extractor, or based on the predictions of the updated baseline model. Finally, the baseline model is updated by jointly using the labeled training samples and the selected pseudo-labeled test samples, and the final predictions on the test samples are co-predicted by the models trained on both EO images and EO-SAR images.

4.2.3 Rank 5: sumanth_udupa

Team sumanth_udupa trained their classifier based on the philosophy that the domain gap between the EO and SAR can be reduced. The EO and SAR models have a feature extractor and a classifier.

The team begins with two ResNet-models that have been pre-trained on ImageNet. One model is then fine-tuned on EO images, and the other is fine-tuned on both EO and

SAR images. The team uses Focal Loss to fine tune their models [15]. Because of the long-tailed distribution of the dataset, they perform data-augmentation on the less represented classes. They use a weighted random sampler to increase the representation of the tail classes.

They used sliced Wasserstein discrepancy as the domain-gap loss [13]. They take the outputs of the feature extractors of the EO model and the SAR model, and minimize the domain gap in the embedding space. To make the learning more efficient, they use class-conditional domain adaptation which they achieve by passing the EO and SAR image pair at a time to the two models, effectively making it shared feature learner. They also use the unlabelled data points and frames provided to them (validation and test set) to minimize the cross-domain Wasserstein loss. The classifiers of both the models are trained using the Focal loss. Focal loss worked better than cross-entropy loss suggesting that some samples are easier to learn than the others.

For Track 2, they just use the ResNet-50 EO model for EO image and SAR image and take weighted average of the two results and achieve competitive results.

4.2.4 Rank 6: jzsherlock

Due to the large difference between SAR and EO images, team jzsherlock constructed a dual-stream network structure to encode the SAR and EO images separately to feature vectors, and then concatenated the two vectors to an a fully connected layer to generate the class prediction. They apply MobileNetV2 [23] structure with pretrained weights from the timm [26] library as encoders for reasons from two aspects: 1) efficiency and lightweight, which can reduce the model size (compared with ResNet-34 [3] based model (170.6M), MobileNetV2 based model is only 18.5M), and boost training speed while maintaining or even slightly boosting the performance, as well as 2) reduce overfitting, because larger models tend to overfit from the available data.

Using the dataset imbalance of each class, the team first uses under-sampling to select same number of k samples for each class, where k is set to the minimum of number of samples per class. Then 20% of each class is randomly separated out as a validation set while the remaining data are used for training. Focal loss [16] is used as the loss function instead of commonly used cross entropy loss to cast more attention on hard samples to tackle with overfitting. Label smoothing and various data augmentation are also used in calculating focal loss for the same reason.

The team utilized semi-supervised strategy in training as well. After the accuracy in validation set is stable, the test set was inferenced with the trained model, and the results with max probability higher than a threshold is added to the training dataset with the inferenced class as their pseudo label. Then the model continued to train some epoches using the combined train dataset. The model weights with best validation accuracy are used for the final inference to generate results of test dataset.

Even if all the strategies are taken to reduce the class bias of models with more samples, the final outputs still tend to bias for the common classes (the first 4 classes which have over 10k samples) over long-tailed classes (the last 6 classes, which all have less than 1k samples, except class "box truck", which is less than 2k). Inspired by [27], the team re-calibrated the output using post-processing: the first four classes are sorted by the predicted probabilities and labeled one-tenth of test samples for each class. Then the remaining samples which have relatively lower probabilities of the common classes are labeled by their maximum probability class. Experiments show that removing the influence of common classes can obtain a more balanced result for the long-tailed classes.

4.2.5 Rank 9: CiuchitiPoc

Team CiuchitiPoc uses a model that consists of a convolutional neural network (CNN) with 11 convolutions, and 1 max pooling layer. The convolution layers use the rectified linear unit (ReLU) activation function and are followed by batch normalization, its kernel, number of channels and parameters are 3x3, 32 and 9248, respectively. The total number of parameters is 94,410. In order to compensate for the uneven distribution of the dataset classes, the team used class weights that improve the class difference and don't allow the CNN to skew the results towards the class with the highest representation within the dataset. The results are given by averaging the output of the two networks. For SAR images, the team took the last 100 images from each class and made a validation set. For EO the team took the last 100 images from each class for validation.

5. Analysis

The 2022 MAVOC challenge had significant performance increases when compared with the 2021 NTIRE MAVOC challenge. The average performance between the top ten teams increased as well as the performance of the best classifiers. Approaches from both years used novel and unique methods to address the unbalanced dataset. However, the 2022 teams used newer techniques such as Swin-Transformers, self-supervised learning, and semi-supervised learning. These techniques are state-of-the-art and were published during 2021. Hence, the top performing 2022 teams leverage recent techniques to outperform the top performers from the 2021 competition. From the competition results, many of these techniques are well suited for ATR applications.

In both the 2021 and 2022 competitions, teams addressed

the data imbalance distribution of the training set. Because the training set has a long-tailed distribution, teams developed methods to increase the frequency of the tail classes. They used various data augmentation techniques, or weighted sampling methods. However, the validation and test set are uniformly distributed. Had the validation and test set matched the distribution of the train set, there might not have been such a consistent effort towards data augmentation.

6. Conclusions

The 2022 MAVOC challenge encourages the development of multi-modal models that excel at image classification which aligns with ATR applications. The MAVOC challenge used the UNICORN dataset that consists of aerial view SAR and EO image pairs; which has unique challenges due to the view point of the SAR and EO images. The challenge comprises two contests: (1) learning from both SAR and EO, but can be used with only SAR inputs, and (2) training and testing with both SAR and EO images at test time. Congratulations to the winners for superior results, participants who enter, and those all those facilitating the challenge which will motivate future work in multi-modal ATR.

Acknowledgements

We would like to thank Dr. Vince Velten and Eric Todd for reviewing this paper, and Bob Lee for providing logistic support for the 2022 PBVS MAVOC Challenge. We would also like to thank Angel Wheelwright for helping run the competition. This documentation has been approved for public release by the US Air Force Research Laboratory with PA approval number AFRL-2022-1668.

Appendix A. Teams Information

We acknowledge the participants. We used edited versions of team submissions for method explanations.

MAVOC 2022 organization team:
Members: Spencer Low, Dr. Oliver Nina, Dr. Angel Sappa, Bob Lee
Affiliation: BYU, AFRL, ESPOL, CVC, WBI

A.1. USTC-IAT-United:
Members: Jun Yu, Hao Chang, Keda Lu, Liwen Zhang, Shenshen Du
Affiliation: University of Science and Technology of China

A.2. NLPR-RVG:
Members: Zhengming Zhou, Jiayin Sun, Qiulei Dong
Affiliation: The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

A.3. NYCX:
Members: Rui Zhang, Yuhui Wu, Zhiwen Wang
Affiliation: University of Electronic Science and Technology of China

A.4. Moyu:
Members: Linpeng Pan, Gongzhe Li,Linwei Qiu, Zhiwen Tan, Fengying Xie, Haopeng Zhang
Affiliation: Beihang University

A.5. NLPR-RVG:
Members: Jiayin Sun, Zhengming Zhou and Qiulei Dong
Affiliation: The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

A.6. sumanth_udupa:
Members: Sumanth Udupa, Aniruddh Sikdar, Suresh Sundaram
Affiliation: IISc, Bengaluru

A.7. jzsherlock:
Members: Sumanth Udupa, Aniruddh Sikdar, Suresh Sundaram
Affiliation: IISc, Bengaluru

A.8. CiuchitiPoc:
Members: Casian Miron, Daria Miron,Oana Moraru
Affiliation: MCC Resources SRL, Romania, Iasi

References

[1] S. Chen, H. Wang, F. Xu, and Y. Jin. Target classification using the deep convolutional networks for sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4806–4817, 2016. 1

[2] Xavier Gastaldi. Shake-shake regularization of 3-branch residual networks. In *ICLR*, 2017. 6

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 7

[4] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *CoRR*, abs/1905.02244, 2019. 4

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. 4

[6] Nathan Inkawhich, Eric Davis, Matthew 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. 1

[7] Nathan Inkawhich, Eric Davis, Uttam Majumder, Chris Capraro, and Yiran Chen. Advanced techniques for robust sar atr: Mitigating noise and phase errors. In *IEEE International Radar Conference (RADAR)*, 2020. 1

[8] Nathan Inkawhich, Matthew Inkawhich, Eric Davis, Uttam 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. 1

[9] Nathan Inkawhich, Jingyang Zhang, Eric K. Davis, Ryan Luley, and Yiran Chen. Improving out-of-distribution detection by learning from the deployment environment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2070–2086, 2022. 2

[10] Menglin Jia, Austin Reiter, Ser Nam Lim, Yoav Artzi, and Claire Cardie. When in doubt: Improving classification performance with alternating normalization. In *EMNLP*, 2021. 6

[11] 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. 6

[12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 5

[13] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. *CoRR*, abs/1903.04064, 2019. 7

[14] Colin Leong, Todd Rovito, Olga Mendoza-Schrock, Christopher Menart, Jason Bowser, Linda Moore, Steve Scarborough, Michael Minardi, and David Hascher. Unified coincident optical and radar for recognition (unicorn) 2008 dataset, 2008. 2, 3

[15] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 7

[16] 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. 7

[17] Jerrick Liu, Nathan Inkawhich, Oliver Nina, Radu Timofte, Sahil Jain, Bob Lee, Yuru Duan, Wei Wei, Lei Zhang, Songzheng Xu, Yuxuan Sun, Jiaqi Tang, Xueli Geng, Mengru Ma, Gongzhe Li, Huanqia Cai, Chengxue Cai, Sol Cummings, Casian Miron, Alexandru Pasarica, Cheng-Yen Yang, Hung-Min Hsu, Jiarui Cai, Jie Mei, Chia-Ying Yeh, Jenq-Neng Hwang, Michael Xin, Zhongkai Shangguan, Zihe Zheng, Xu Yifei, Lehan Yang, Kele Xu, and Min Feng. NTIRE 2021 multi-modal aerial view object classification challenge. *CoRR*, abs/2107.01189, 2021. 2, 3

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5

[19] Uttam Majumder, Erik Christiansen, Qing Wu, Nate Inkawhich, Erik Blasch, and John Nehrbass. High-performance computing for automatic target recognition in synthetic aperture radar imagery. In Igor V. Ternovskiy and Peter Chin, editors, *Cyber Sensing 2017*, volume 10185, pages 76 – 83. International Society for Optics and Photonics, SPIE, 2017. 1

[20] Uttam K. Majumder, Erik P. Blasch, and David A. Garren. *Deep Learning for Radar and Communications Automatic Target Recognition*. Artech House, 2020. 1

[21] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *ICCV*, pages 6082–6091, 2019. 5

[22] Timothy Ross, Stephen Worrell, Vincent Velten, John Mossing, and Michael Bryant. Standard sar atr evaluation experiments using the mstar public release data set. In *SPIE Conference on Algorithms for Synthetic Aperture Radar Imagery V*, 1998. 1

[23] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 5

[25] Divya Shanmugam, Davis W. Blalock, Guha Balakrishnan, and John V. Guttag. When and why test-time augmentation works. *ArXiv*, abs/2011.11156, 2020. 6

[26] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 7

[27] Lehan Yang and Kele Xu. Cross modality knowledge distillation for multi-modal aerial view object classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–387, 2021. 7

[28] Y. Zheng, E. Blasch, and Z. Liu. *Multispectral Image Fusion and Colorization*. Press Monographs. SPIE Press, 2018. 1