Multi-modal Aerial View Object Classification Challenge Results - PBVS 2023

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 Research Laboratory
Arlington, VA
erik.blasch.1@us.af.mil

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

Abstract

This paper presents the findings and results of the third edition of the Multi-modal Aerial View Object Classification (MAVOC) challenge in a detailed and comprehensive manner. The challenge consists of two tracks. The primary aim of both tracks is to encourage research into building recognition models that utilize both synthetic aperture radar (SAR) and electro-optical (EO) imagery. Participating teams are encouraged to develop multi-modal approaches that incorporate complementary information from both domains. While the 2021 challenge demonstrated the feasibility of combining both modalities, the 2022 challenge expanded on the capability of multi-modal models. The 2023 challenge introduces a refined version of the UNICORN dataset and demonstrates significant improvements made. The 2023 challenge adopts an updated UNIfied COincident Optical and Radar for recognitioN (UNICORN V2) dataset and competition format. Two tasks are featured: SAR classification and SAR + EO classification. In addition to measuring accuracy of models, we also introduce out-of-distribution measures to encourage model robustness.

The majority of this paper is dedicated to discussing the top performing methods and evaluating their performance on our blind test set. It is worth noting that all of the top ten teams outperformed the Resnet-50 baseline. The top team for SAR classification achieved a 173% performance improvement over the baseline, while the top team for SAR + EO classification achieved a 175% improvement.

1. Introduction

The objective of Automatic Target Recognition (ATR) models is to accurately detect, recognize, classify, and identify target signatures present in remotely sensed imagery, as evidenced by various sources such as [3, 6, 7, 16, 17, 20]. ATR is similar to object detection and labeling in natural imagery; however, ATR systems are often built on complex remote sensing (RS) systems mounted on aircraft or spacecraft. Hence, there are unique challenges to automatically detect and labeling aerial images, such as limited sensor resolution resulting in only a handful of pixels on the target, as described in [5]. SAR, while researched as a single-mode ATR, offers benefits such as all-weather, all-time, and stand-off results; however, it also presents challenges with signal multi-bounce, shadows, and distinguishing boundaries of closely-spaced objects [9]. Combining EO and SAR for ATR presents many unique research challenges and opportunities.

One significant distinction between ATR models and natural image object detection models is that ATR systems must be capable of detecting out-of-distribution (OOD) samples. A model may encounter data outside of the training set, and high-confidence misclassification can have severe consequences given the applications of ATR systems. Therefore, ATR models must provide a confidence score (or credibility [2]) for each label, with low scores indicating low-confidence [5, 8].

The 2023 PBVS Multi-modal Aerial View Object Classification (MAVOC) challenge (building on the 2022 chal-

lenge [15]) provides an excellent opportunity to study the complex issues associated with EO-SAR image and gain insights on how to optimize the use of multi-modal information in ATR models.

RS systems can benefit from considering multiple sensor types. Each modality may offer a different strength such as self-illuminated systems can operate without sunlight or microwave band system can image through clouds and vegetation. Passive sub-optical sensors can remotely measure temperature, while active radars can identify man-made objects in jungles or infer wind speed. Combinations of passive and active data from RF and EO sensors can enhance object detection, user appreciation, and classification robustness [22]).

Despite their advantages, RS systems are often overlooked in computer vision applications due to the difficulty in combining different sensor data in complementary ways such as signal registration, data scaling, concept alignment, and feature association. As a result, most RS systems rely on a single modality, typically visual data composed of multiple spectrum bands. While extensions like multi-spectral (MSI) and hyperspectral (HSI) data offer more bands, they also require determining the salient bands for the targets of interest and necessitate more computation power than the EO domain. However, by intelligently fusing various sensor data, ATR performance improvements are expected.

2. Challenge

The 2023 MAVOC challenge is held jointly with the Perception Beyond the Visible Spectrum (PBVS) workshop following the 2022 competition and the 2021 workshop held in conjunction with NTIRE [12]. The MAVOC challenge is designed to facilitate innovative approaches in multi-modal classifiers using pairs of SAR and EO images. Participants are evaluated on the top-1% accuracy and area under the receiver operating characteristic (AUROC). This is to encourage the design of models that excel both at labelling and detecting out-of-distribution samples.

The SAR images provided pose a unique challenge to participants due to their self-illuminated and coherent nature, leading to images with distinct *SAR shadows* and a tilted perspective. The challenge is bifurcated into two tracks, each emphasizing multi-modal models with differing utilities.

2.1. Track 1

The primary objective of Track 1 is to develop a classifier that can be trained on both SAR and EO data and tested solely on SAR data. The resulting classifier should not rely on EO data when deployed, but instead, learn from the merged features found in both SAR and EO images during training. By eliminating the need for both modalities at test time, decisions can be made faster, as the computationally expensive rectification preprocessing required to align SAR and EO is not required. The diverse nature of the training data from different modalities makes it a challenging task.

2.2. Track 2

The primary objective of Track 2 is to develop a classifier trained on both SAR and EO data, and tested on (SAR, EO) image pairs. Training with EO/SAR data allows the ATR models to utilize features from both modalities during both training and deployment, potentially resulting in more accurate classifiers. Compared to Track 1, Track 2 benefits from the additional input information at test time, as EO images are typically less noisy than SAR images.

2.3. AUROC

The 2023 MAVOC challenge introduces the use of AUROC to determine a model's ability to detect out-of-distribution samples. The AUROC ranges from 0 to 1. A score of 0.5 corresponds with random guessing, and a score of 1 corresponds with perfect confidence. Out-of-distribution (OOD) samples or negative samples are shuffled into the validation and test sets, but not provided during training. Moving OOD samples to testing discourages the use of additional training classes as a catch-all for the negative samples.

2.4. Dataset

The foundation of the challenge is built upon the UNIfied COincident Optical and Radar for recognitioN (UNICORN) dataset [11], which offers a publicly available and aligned SAR-EO dataset with hand-labeled classes. The 2023 MAVOC challenge utilizes a refined version of the 2008 UNICORN dataset (UNICORN-V2). The UNICORN-V2 dataset consists of Wide Area Motion Imagery (WAMI) large format electro-optical (EO) sensor [18] and Wide Area Synthetic Aperture Radar (SAR) data, collected from an aircraft flown over Dayton, Ohio. This dataset exhibits improved alignment accuracy, and contains an increase in labelled images. A baseline pre-trained Resnet-50 model performed similiary between the two version of the UNICORN dataset scoring between 15-18% accuracy on a held out test set. We also observed that models and approaches from the 2022 challenge performed nearly identically on the new dataset.

While the SAR and EO data cover the same approximate field of view, the reconstructed SAR image has a finer resolution than the EO image. These large SAR and EO images are rectified and aligned using homography algorithms, as depicted in Figure 2. The competition dataset comprises small windowed sections (chips) that are sub-images of the aligned large image. Each chip contains one of 10 objects to

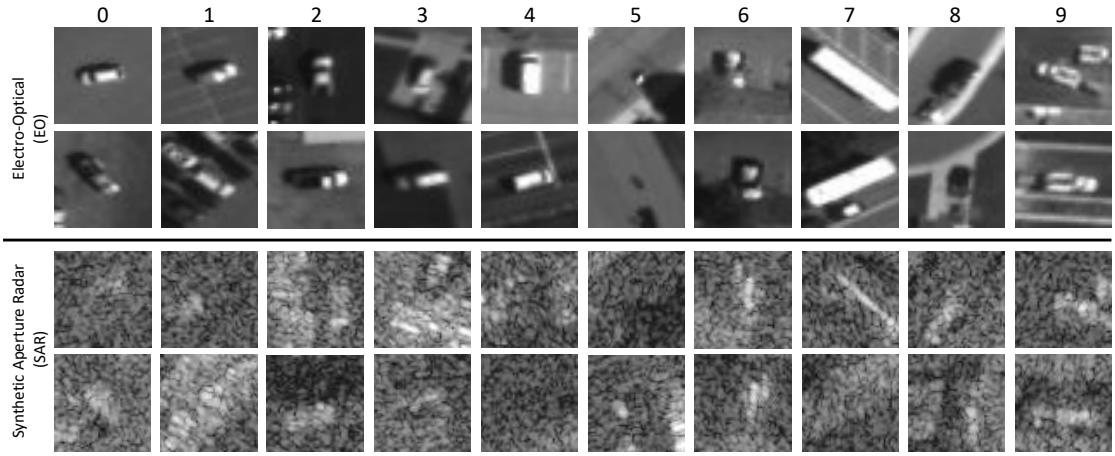

Figure 1. Two sample pairs of EO and SAR chips from each of the 10 classes in the UNICORN Dataset [12].

Figure 2. The aligned scene of the full UNICORN dataset before chipping is performed [11].

be classified. A typical EO chip is 31 $\times$ 31 px and a typical SAR chip is 55 $\times$ 55 px due to the differing resolutions of the large images. Figure 1 shows examples of (SAR, EO) pairs from each class of the dataset. The UNICORN-V2 dataset is created using a refined homography and alignment process. The alignment enables both more chips to be identified and more precise. As demonstrated in Table 1, the dataset is partitioned into train, validation, and test sets, with the train set classes non-uniformly distributed, following a long-tail distribution. However, the validation and test sets are uniformly distributed across all 10 classes, enabling a true and unbiased measurement of accuracy.

The out-of-distribution images are pulled from other classes in the UNICORN-V2 dataset. These are classes with fewer examples than the flatbed truck w/ trailer class. These classes are shown in Table 2.

Table 1. Details of the UNICORN-V2 Dataset used as the in-distribution classes in this challenge (counts represent the number of (EO, SAR) pairs).

Class #	Vehicle Type	# Train	# Val	# Test
0	sedan	364,228	77	200
1	SUV	43,642	77	200
2	pickup truck	24,420	77	200
3	van	17,159	77	200
4	box truck	3,414	77	200
5	motorcycle	2,351	77	200
6	flatbed truck	1,233	77	200
7	bus	1,130	77	200
8	pickup truck w/ trailer	971	77	200
9	flatbed truck w/ trailer	714	77	200
	Total	459,262	770	2000

Table 2. Details of the UNICORN-V2 Dataset used as the out-of-distribution classes in this challenge (counts represent the number of (EO, SAR) pairs).

Class #	Vehicle Type	# Train	# Val	# Test
0	other	-	77	1,151
1	sedan w/ trailer	-	77	872
2	dismount	-	77	609
3	SUV w/ trailer	-	77	681
4	plane	-	77	432
	Total	-	770	3,745

2.5. Evaluation

Submissions are evaluated using a weighted average of the top-1% accuracy and AUROC of the model. The test set contains 2,000 unlabelled (SAR, EO) chip pairs, with 200 examples for each of the 10 classes, and 3745 out-of-distribution samples. The weighting is shown in Eq. 1.

$$\text{Score} = 0.75\ \text{Accuracy} + 0.25\ \text{AUROC} \quad (1)$$

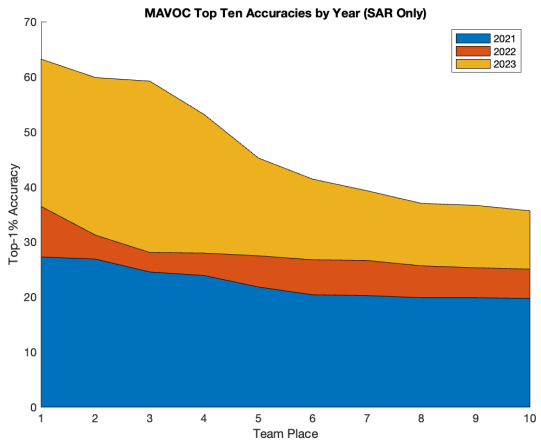

Figure 3. The performance differences between the 2021 NTIRE, 2022 PBVS, and 2023 PBVS MAVOC challenges. This figure plots the Top-1% accuracy of the ten top performing teams for SAR classification.

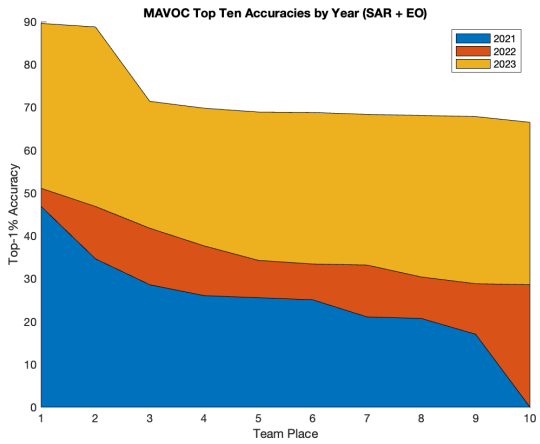

Figure 4. The performance differences between the 2021 NTIRE, 2022 PBVS, and 2023 PBVS MAVOC challenges. This figure plots the Top-1% accuracy of the ten top performing teams for SAR + EO classification.

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

Rank	Team	Total ↑	Accuracy ↑	AUROC ↑
1	**fcai21**	**0.65**	**63.20**	**0.71**
2	Running	0.64	59.20	0.80
3	qingqing	0.61	53.15	0.85
4	Raghunath19	0.61	59.85	0.64
5	papu	0.45	41.40	0.56
6	yangyang6	0.45	45.24	0.44
7	USTC-IAT-United	0.44	34.55	0.71
8	VL	0.42	39.30	0.52
9	hsansui	0.42	34.55	0.66
10	jsyoon	0.41	34.75	0.60
	2022 Best (USTC)	-	36.44	-

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

Rank	Team	Total ↑	Accuracy ↑	AUROC ↑
1	**overfittinghhh**	**0.84**	**89.60**	**0.68**
2	txccc	0.84	88.77	0.68
3	doggie	0.74	69.80	0.85
4	Raghunath19	0.71	71.40	0.70
5	fcai21	0.71	68.35	0.79
6	sunny088	0.71	68.80	0.76
7	mcc	0.70	68.90	0.74
8	cathy122	0.70	68.10	0.76
9	CCCathy	0.69	67.85	0.74
10	wwwwww	0.67	66.50	0.70
	2022 Best (USTC)	-	51.09	-

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 and confidence score 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. Results are made visible during both phases.

2.6. Challenge Phases

The challenge began January 11, 2023, and the test data was released March 1, 2023. The testing phase ended on March 7, 2023 with team submissions finalized.

3. Challenge Results

One hundred and nineteen teams participated in Track 1. Of those 119 participants, 44 teams submitted their algorithms during the development phase, and 43 teams submitted during the testing phase. Track 2 had 105 participants. Of those 105 participants, 32 submitted their algorithm during the development phase, and 47 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 and 2022 PBVS MAVOC challenge results.

3.1. Track 1 SAR Classification Results

In Track 1 of the 2023 competition, teams displayed exceptional performance, surpassing their counterparts from the previous year by a significant margin. Notably, the top model of 2023 achieved a remarkable increase of 73% in its top-1% accuracy compared to the leading model of 2022. Figure 3 provides a clear visualization of the performance gap between the top ten teams of each year's MAVOC challenge. Additionally, Table 3 presents a summary of the performance of the highest-performing teams.

3.2. Track 2 SAR + EO Classification Results

Additionally, teams in Track 2 of the 2023 competition outperformed the 2022 competition teams by a considerable margin. The top-1% accuracy of the top model of 2023 increased by 75% over the top performing model of 2022. Figure 4 illustrates the performance difference between the top 10 teams from each year of the MAVOC challenge. Table 4 is a summary of the performance of the top performing teams.

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

4.1.1 Rank 1: fcai

The team fcai presents a novel three-stage approach for improving the performance of computer vision models on imbalanced datasets. Their method utilizes a ResNet-101 backbone pre-trained on ImageNet for feature extraction. In the first stage, the model is trained on the entire dataset to learn rich feature representations. However, they observe severe imbalances in both the distribution of classes, with 80% of the samples belonging to class 0 and only 0.15% to class 9, and the distribution of samples within each class, particularly in the head classes. This leads to redundancy and overfitting issues. To address these challenges, they employ the Class-Balanced Loss to re-weight the loss of each class based on its effective sample size. In the second and third stages, they generate reliable pseudo-labels and utilize semi-supervised learning to leverage the underlying structure and distribution of the data to improve model performance and reduce reliance on labeled data. In the second stage, they predict the remaining samples in the test set using a high confidence model and cluster them using DBSCAN and k-means. They filter out outlier clusters and generate pseudo-labels to create a balanced dataset for semi-supervised training. To further enhance data diversity, team fcai apply SAR dataset-specific data augmentation strategies. In the third stage, they introduce a novel Reliable Sample Pool (RSP) to enhance the model's confidence in predicting in-distribution data and its out-of-distribution detection ability. The RSP stores the top-N samples with the highest confidence scores for each class prediction on the test set after each epoch. These reliable samples are included in the training set of the next epoch to strengthen the model's trust in in-distribution data. The capacity of the RSP is a learnable hyperparameter, and the samples in the pool will also be updated as the model changes. Additionally, they fine-tune the model using the reliable samples in the RSP.

4.1.2 Rank 2: Running

Team Running proposes a novel domain alignment embedded contrastive learning semi-supervised network (DCsemNet) for aerial view object classification. The schematic of the proposed DCsemNet for aerial view object classification is depicted in Figure 5, which is described in detail as follows. Team Running first pre-processes the training data. Through data analysis, they found that the training data is long-tailed and there are many repeated scene samples. Therefore, they design a key frame extraction strategy to clean the data, thereby reducing the samples of similar scenes and reducing serious data imbalance.

Team Running adopts a two-stage training strategy to optimize the performance of the proposed model. In the first stage, based on the modal complementary between the EO and SAR data, they propose a domain alignment embedded contrastive learning feature representation network (DCFRNet). The proposed DCFRNet extracts the domain invariant features of the bi-modal data, by minimizing the maximum mean discrepancy (MMD), conditional maximum mean discrepancy (CMMD), and cosine-similarity of the two data features, which reduces the distribution difference between the SAR data features and the bi-modal fusion data features. The proposed DCFRNet then embeds domain alignment into a balance contrastive learning network to increase the inter-class discrimination and use the logit compensation strategy to eliminate the bias caused by data imbalance, so as to optimize the long-tail problem.

In the second stage, they propose a high-confidence pseudo-label generation based semi-supervised optimization strategy (HPGSem). They obtain the trained classification network in the first stage to generate pseudo-label of the testing data. They train a Resnet-based classification network to output the probability, and obtain the confidence of each sample by enhancing the energy function. Then, they train a binary classification network for out-of-distribution detection tasks by using high-confidence in-distribution samples and low-confidence out-of-distribution samples. The confidence of the binary classification results is used as the discriminant inside and outside the distribution, and the out-of-distribution samples are eliminated. The proposed HPGSem introduces confidence learning to obtain the high-confidence pseudo label. After that, semi-supervised optimization strategy is adopted to fine-tune the model obtained in the first stage, so as to obtain the final high-precision classification results. Due to the characteristics of repeated scenes in the test data, they design a post-

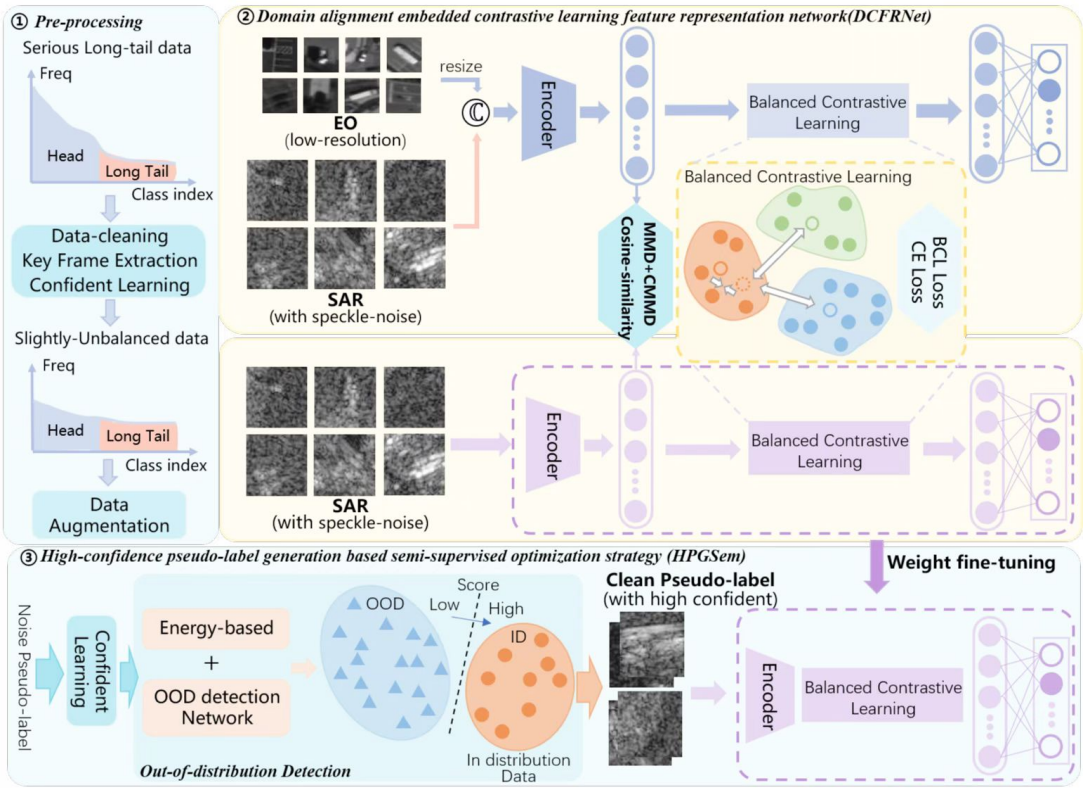

Figure 5. Team Running's model architecture. This architecture is similar to team doggie's architecture. This is due to both teams comprising of the same members.

processing method based on the similarity clustering of image scenes. The labels of the same scene in each cluster are uniformly corrected using the majority vote procedure.

4.1.3 Rank 4: Raghunath19

To address the issue of imbalanced training data, team Raghunath19 implemented a strategy to ensure that each class had an equal number of 6,000 images. For classes that originally had over 6,000 images, they randomly selected 6,000 images to include in the balanced dataset. Meanwhile, for classes with fewer than 6,000 images, they applied various data augmentation techniques such as flipping, rotation, and affine transform to expand the dataset to 6,000 images. As a result, their new training data comprises a total of 60,000 images, which is a substantial decrease from the original dataset size of approximately 459,000 images. They experimented with several different architectures and found that the most effective approach for their validation/test strategy involved using a ResNet-34 [4] network with pre-trained weights. In training the ResNet-34 network, they employed a combination of triplet loss [21] and cross-entropy loss, inspired by the work of [1]. To reduce the dimensionality of individual EO and SAR features, they performed principal component analysis (PCA) and built a k-d tree of depth 128 on the resulting features to obtain the appearance label of each training sample in the tree node. To obtain positive and negative triplet loss samples, they searched the k-d tree for each anchor in each tree node. Specifically, they identified positive samples that belonged to the same class as the anchor, but were not located in the same node, and negative samples that were located in the same node as the anchor, but were not of the same class. This process resulted in about 500,000 triplet pairs for network training, with approximately equal numbers of triplet pairs for each positive and negative class pair.

Team Raghunath19 used a similar architecture for the Track 1 and Track 2 challenges. The main difference in between them is the initial convolution layer has 1 channel and 2 channels for Track 1 and Track 2 challenges respectively. For Track 2 they also implemented an ensemble method where they used ensemble fusion technique to achieve better results. The models that were used for ensemble fusion were ResNet-34 [4], EfficientNet-B0 and Swin Transformer [13]

4.1.4 Rank 6: yangyang6

Team yangyang6 used the same approach detailed in the next section: Track 2, "overfittinghhh".

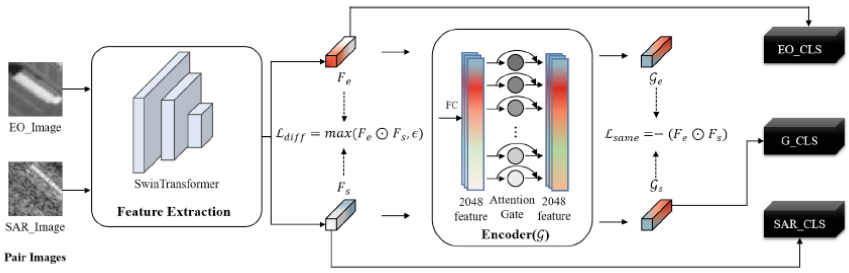

Figure 6. Team VL's model overview.

4.1.5 Rank 8: VL

Team VL aims to obtain a SAR classification model similar to a EO classification model. The structure of their proposed method is shown in Figure 6. Their pipeline consists of data processing, model design, training, and inference. To deal the long-tail problem, they use a non-uniform enhancement strategy to increase the sample number of tail classes. Then they randomly select 5000 images from each class to get a uniform train dataset. The enhancements includes crop, rotation, and resize.

In the train stage, the pair SAR and EO images are fed to the model to get their features Fs and Fe. And then they are input to a encoder module G to generate consistent features Gs and Ge. Team VL assumes the output of the encoder to be consistent regardless of whether the input is a SAR object or EO object. And they use a same loss Lsame to supervise the learning of the consistent features. At the same time, the Fs and Fe should be different and a difference loss Ldiff is used to supervise this learning process. Then the Fs, Fe, and Gs are fed to classifiers to predict the classification probability. The cross-entropy function is used to supervise them. The training pipeline contains two stage. In the Stage 1, the uniform training dataset (SAR and EO images) is used to train the model with the pre-trained Swin Transformer on ImageNet (100 epochs). In the Stage 2, they first generate pseudo-labels of test images according to the high score prediction of the Stage 1 model. And then the Stage 1 model is finetuned for 19 epochs with the pseudo-dataset.

4.1.6 Rank 10: jsyoon

Team jsyoon proposed a system that is based on transfer learning methods using the ConvNext model. The proposed system automatically cleaned the input dataset through the perceptual hash-based deduplication algorithm [10] until each label has fewer than 4000 images. They removed the error data, which has zero value of the sum of pixel values except 'Sedan(0)'. They employed various data augmentation processes in the training phase (flipping, rotate, Gaussian noise, random resize, crop) [19]. The x-large size of the pre-trained ConvNext model was utilized for the classification model. They used a cosine annealing-based learning rate scheduler to overcome the overfitting in the training phase [14].

4.2. Track 2

4.2.1 Rank 1: overfittinghhh

Team overfittinghhh employs a two stage model. In Stage 1 they trained multiple 10-class classifiers and used them to vote for each sample, taking the most-voted class as the final label and the average confidence of these models as the final confidence score. They found that some classes (class 0: sedan, 1: suv, 2: pickup truck, 3: van) are visually similar, so they trained a two-stage classification module to optimize their performance.

In Stage 2, they added a classification model and a clustering module to re-identify the samples that are easily confused (class 0 - 3). Specifically, they used the Perceptual hash algorithm to cluster the samples that were classified as class 0 - 3, and then classified the samples within each cluster using the 4-class classifier. The most frequently occurring category was taken as the label for all samples in the cluster. To fuse SAR and EO, they first extract features using a backbone model and then test Transformer and traditional concatenation feature fusion strategies, as well as a concatenation module based on channel self-attention. Ultimately, they used a concatenation feature fusion strategy based on channel attention. They used the output results of multiple models to vote for the final result. The confidence of the sample is calculated from the average of multiple voting models. In the experiment, they found that the distribution of class 0 - 3 categories is similar, so they trained a two-stage module to further optimize.

4.2.2 Rank 3: doggie

Team doggie proposed a balanced contrastive learning semi-supervised network (BCLsemNet) for multi-modal aerial view object classification. This architecture is similar to the one used by team Running, as both teams are comprised of the same team members. The architecture of the proposed BCLsemNet is depicted in Figure 7. They

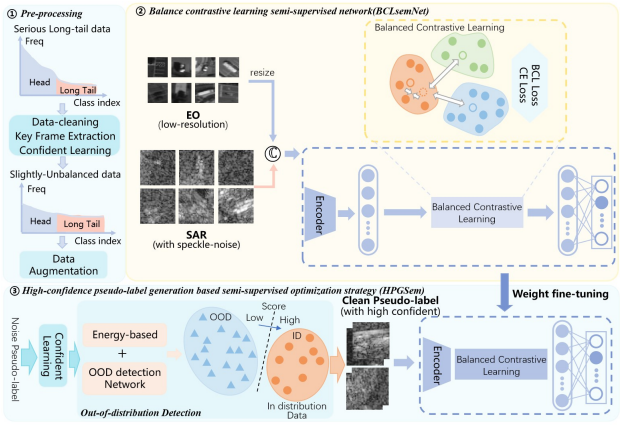

Figure 7. Architecture of the balance contrastive learning semi-supervised network (BCLsemNet) for multi-modal object classification. This architecture is similar to team Running's architecture. This is due to both teams comprising of the same members.

first pre-process the training data. Because the training data is extremely unbalanced and there are many similar scene samples, they propose a key frame extraction strategy to clean the data, thereby reducing the samples of similar scenes and alleviating the long tail problem of the data.

They adopt a two-stage training strategy to optimize the performance of the model. In the first stage, they propose a balanced contrastive learning feature representation network (BCLNet). Due to the modal complementarity between EO data and SAR data, they stitch EO data and SAR data according to the channel dimension. The spliced data is input into the balanced contrastive learning network to increase the inter-class gap and reduce the intra-class difference. For the long-tail problem of data, they embed class-averaging and class-complement strategies in contrastive learning to optimize the long-tail problem.

In the second stage, they propose a high-confidence pseudo-label generation based semi-supervised optimization strategy (HPGSem). They obtain the trained classification network in the first stage to generate pseudo-label of the testing data. By screening and optimizing the pseudo-labels generated in the first stage, highly reliable pseudo-labels are obtained for semi-supervised training. Specifically, they design a binary classification network for out-of-distribution detection tasks. They train a Resnet-based classification network to output the probability of each predicted sample, post-process the probability value, and enhance the energy function to obtain each sample. According to the level of its energy value, some samples are selected according to the category as the inside and outside the distribution samples. These samples are given 1-0 labels and sent to the binary classification network for training. The confidence of the binary classification results is used as the discriminant inside and outside the distribution, and the out-of-distribution samples are eliminated. Then, the proposed HPGSem introduces confidence learning to evaluate the quality of the generated pseudo-label, and chooses the high-confidence pseudo-label. After obtaining the high-confidence pseudo label, a semi-supervised optimization strategy is adopted to fine-tune the classification model obtained in the first stage, so as to obtain the final high-precision classification network. Due to the characteristics of repeated scenes in the test data, after obtaining the test labels, they design a post-processing method based on the similarity clustering of image scenes. The labels of the same scene in each cluster are uniformly corrected using the majority vote procedure.

4.2.3 Rank 4: Raghunath19

The same base approach is used as in Track 1 with a minor additions. For Track 2 they implemented an ensemble method where they used ensemble fusion technique to achieve better results. The models that were used for ensemble fusion were ResNet-34 [4], EfficientNet-B0 and Swin Transformer [13].

4.2.4 Rank 5: fcai21

Despite differences between EO and SAR images, they assume that the SAR image domain provides knowledge representation beyond the EO image domain. Hence, they utilize high-confidence samples from the SAR image classification network that their team developed in Track 1 as pseudo-labels. To address the issue of long-tail distribution, they propose a two-stage network with shared pooling.

5. Conclusions

Submission performances for both tracks of the 2023 PBVS MAVOC challenge increased dramatically when compared with the 2022 challenge results. In Track 1, we observed a 73% increase in accuracy. Similarly, in Track 2, we observed a 75% increase in accuracy. These score improvements can be attributed to advancements in the architecture and training methods. The observed results demonstrate many new approaches to accommodate sparse and non-uniformly distributed data.

The 2023 MAVOC challenge demonstrated remarkable growth in both participation and the performance of the participants. We have observed year over year growth in both of these metrics. Additionally, the introduction of an out-of-distribution score has provided insight into these ATR models and enable more informed application.

Acknowledgements

We would like to thank Angel Wheelwright, Justice Wheelwright, and Eve Myadze-Pike for support in running the competition.

References

[1] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. Group-sensitive triplet embedding for vehicle reidentification. *IEEE Transactions on Multimedia*, 20(9):2385–2399, 2018. 6

[2] Erik Blasch, Audun Josang, Jean Dezer, and et al. Urref self-confidence in information fusion trust. In *Int'l Conf. on Information Fusion*, 2014. 1

[3] 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

[4] 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. 6, 8

[5] 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

[6] 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

[7] 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

[8] 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. 1

[9] Sadman C. Jahan, Andreas Savakis, and Erik Blasch. Cross-modal knowledge distillation in deep networks for sar image classification2. In *Proc. Spie 12099*, 2022. 1

[10] Tanuj Jain, Christopher Lennan, Zubin John, and Dat Tran. Imagededup. https://github.com/idealo/imagededup, 2019. 7

[11] 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

[12] 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

[13] 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. 6, 8

[14] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. 7

[15] Spencer Low, Oliver Nina, Angel D. Sappa, Erik Blasch, and Nathan Inkawhich. Multi-modal aerial view object classification challenge results-pbvs 2022. In *Proceedings of the IEEE conference on computer vision and pattern recognition - workshop*, 2022. 2

[16] 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

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

[18] Kannappan Palaniappan, Mahdieh Poostchi, Hadi Aliakbarpour, and et al. Moving object detection for vehicle tracking in wide area motion imagery using 4d filtering. In *International Conference on Pattern Recognition (ICPR)*, 2016. 2

[19] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary R. Bradski. Kornia: an open source differentiable computer vision library for pytorch. *CoRR*, abs/1910.02190, 2019. 7

[20] 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

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. 6

[22] Asad Vakil, Jenny Liu, Peter Zulch, and et al. A survey of multimodal sensor fusion for passive rf and eo information integration. *IEEE Aerospace and Electronic Systems Magazine*, 36(7):44–61, 2021. 2

Appendix A. Teams Information

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

MAVOC 2023 organization team:

Members: Spencer Low, Dr. Oliver Nina, Dr. Angel Sappa, Dr. Nathan Inkawhich

Affiliation: BYU, AFRL, ESPOL, CVC, WBI

fcai:

Members: Feng Cai, Keyu Wu, Feng Wang, Haipeng Wang

Affiliation: Fudan University

Running:
Members: Wenqian Dong, Shaoxiong Hou, Jiahui Qu, Jizhou Cui, Jie He, and Ling Huang
Affiliation: Xidian University

Raghunath19:
Members: Raghunath Sai Puttagunta, Dr. Zhu Li
Affiliation: University of Missouri-Kansas City

yangyang6 and overfittinghhh:
Members: Yang Yang, Tong Xin, Lu Xiang Zhe

VL:
Members: Weilong Guo
Affiliation: Technology and Engineering Center for Space Utilization, Chinaese Academy of Sciences

jsyoon:
Members: Jiseok Yoon, Ik Hyun Lee, Sunder Ali Khowaja
Affiliation: IKLAB, Tech University of Korea, University of Sindh

doggie:
Members: Shaoxiong Hou, Wenqian Dong, Jiahui Qu, Junying Ren, Yuanbo Yang, Yvshan Xie
Affiliation: Xidian University