2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 979-8-3503-0249-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPRW59228.2023.00060

Appearance Label Balanced Triplet Loss for Multi-modal Aerial View Object Classification

Raghunath Sai Puttagunta [1], Zhu Li [1], Shuvra Bhattacharyya [2], George York [3]
[1] University of Missouri-Kansas City
[2] University of Maryland, College Park
[3] US Air Force Academy
rpyc8@umsystem.edu, lizhu@umkc.edu, ssb@umd.edu, george.york@usafa.edu

Abstract

Automatic target recognition (ATR) using image data is an important computer vision task with widespread applications in remote sensing for surveillance, object tracking, urban planning, agriculture, and more. Although there have been continuous advancements in this task, there is still significant room for further advancements, particularly with aerial images. This work extracts rich information from multimodal synthetic aperture radar (SAR) and electro-optical (EO) aerial images to perform object classification.

Compared to EO images, the advantages of SAR images are that they can be captured at night and in any weather condition. Compared to EO images, the disadvantage of SAR images is that they are noisy. Overcoming the noise inherent to SAR images is a challenging, but worthwhile, task because of the additional information SAR images provide the model.

This work proposes a training strategy that involves the creation of appearance labels to generate triplet pairs for training the network with both triplet loss and cross-entropy loss. During the development phase of the 2023 Perception Beyond Visual Spectrum (PBVS) Multi-modal Aerial Image Object Classification (MAVOC) challenge, our ResNet-34 model achieved a top-1 accuracy of 64.29% for Track 1 and our ensemble learning model achieved a top-1 accuracy 75.84% for Track 2. These values are 542% and 247% higher than the baseline values. Overall, this work ranked 3rd in both Track 1 and Track 2.

1. Introduction

Advancements in deep learning have resulted in advancements in ATR for aerial image classification due to the ability of deep learning models to extract rich visual information. Typically, ATR systems use a single EO sensor, but an ideal ATR would utilize both EO and SAR sensors to complement each other for better aerial image classification. This work explores using both EO and SAR modalities to improve aerial image classification. Fig. 1 shows the EO and SAR images used in our work.

SAR and EO sensors are two of the most widely used sensors in modern remote sensing systems. EO imagery, or traditional overhead imagery, is easy to gather because it is illuminated by sunlight. It is also easy to interpret because it captures light in the familiar visible spectrum, similar to RGB or grayscale images. EO imagery does not perform well in uneven lighting, darkness, and poor weather conditions; however, SAR imagery does perform well in those circumstances because it uses active illumination. SAR imagery also performs especially well when there are multiple objects of interest, or when the object of interest is small, because the SAR sensor captures and stitches together multiple images with multiple polarization combinations [3].

In recent years, there has been extensive work on ATR for aerial image classification using EO images [5, 10, 16, 19, 22, 24, 33]. In comparison, there has been limited work using SAR images [4, 7, 29, 30], and even less work incorporates both EO and SAR images. There was an increase in the amount of work that incorporates both EO and SAR images [12, 15, 23, 28, 31, 32] after the introduction of the MAVOC challenges in NTIRE 21 [18] and PBVS 22 [21].

ATR for the classification of aerial images presents unique challenges when compared with ATR for the classification of other types of images. Aerial images contain highly textured subjects of varying scales. They also have limited pixels for the point of interest because the point of interest appears relatively small from the perspective of an aircraft. The resulting high intraclass variability makes it difficult to differentiate between classes.

The dataset provided in the MAVOC challenge poses several unique challenges, including long-tailed distribution of classes, a lack of pixel registration between EO and SAR

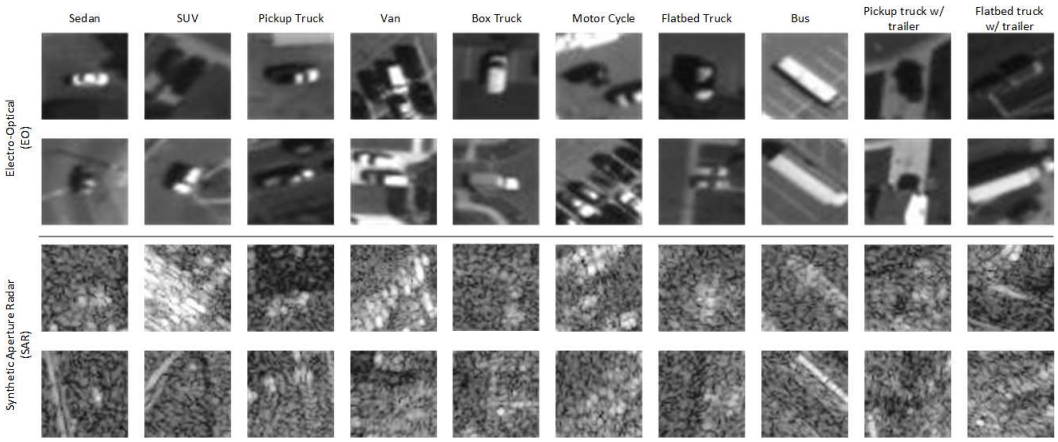


Figure 1. PBVS 23 MAVOC Challenge EO and SAR Images Grouped by Class

images, and low image resolution. The long-tailed distribution of classes can cause bias towards the majority class. The lack of pixel registration (or different fields of view) and the low image resolution make it difficult for models to properly classify the data.

Our work balanced the training dataset through data augmentation and random sampling. Data augmentation techniques were performed on classes with less than 6,000 samples. The data augmentation techniques that were performed included flips, rotations, and affine transformations. For classes with greater than 6,000 images, 6,000 images were randomly selected.

After balancing the training dataset, pre-trained VGG-16 [26] features were extracted. The features were then clustered using a k-dimensional tree (KD-tree) to create appearance labels. Inspired by Liao et al. [16], the appearance labels were then used to construct optimized triplet pairs for triplet loss. Triplet loss ensures that samples of the same class are closer to each other and samples of different classes are further apart from each other in the embedding space [25]. Positive triplet pairs are images from the same class with different appearance labels. Negative triples pairs are images from different classes with the same appearance labels. After the pre-processing steps were completed, the network was trained with both triplet loss and cross-entropy loss.

The MAVOC challenge consists of two tracks. Track 1 requires both the EO and SAR sensor information to be used during the training process and only the SAR sensor information to be used during the evaluation process. In contrast, Track 2 requires both the EO and SAR sensor information to be used during both the training process and the evaluation process. The training datasets are the same for both Track 1 and Track 2. Therefore, the appearance labels are the same for both Track 1 and Track 2.

The contributions from this paper are summarized as follows:

1. Our work balanced the training dataset through data augmentation and random sampling. Data augmentation techniques such as flips, rotations, and affine transformations were performed on classes with less than 6,000 images. For classes with greater than 6,000 images, 6,000 images were randomly selected.
   item Pre-trained VGG-16 [26] features were extracted from the training dataset for both EO and SAR images. The features were then clustered using KD-tree to create appearance labels. Those appearance labels were used to mine optimized triplet pairs for triplet loss.

2. The network was trained with both triplet loss and cross-entropy loss and achieved a top-1 accuracy of 64.29% for Track 1 and a top-1 accuracy of 75.24% for Track 2 during the development phase of the competition.

3. Overall, this work ranked 3rd in both Track 1 and Track 2.

2. Related Work

EO sensors are the most popular sensors in modern remote sensing systems because they are inexpensive, readily available, and easy to interpret under sunny and clear weather conditions. Conversely, SAR sensors excel when there is no lighting, inconsistent lighting, or adverse weather conditions. SAR sensors perform exceptionally well when there are several objects of interest, or when the object is small, as they capture and fuse multiple images with various polarization combinations. SAR sensors are less popular than EO sensors because they are expensive, not readily available, and not easy to interpret due to noise and the fact that they do not capture light in the familiar visible spectrum.

Given that EO sensors are the most popular sensors in modern remote sensing systems, there is a substantial amount of work for ATR aerial image classification using images captured from EO sensors [5,6,10,16,19,22,24,33]. Cheng et al. [5] proposed a large-scale aerial image dataset and experimented with multiple deep learning models to establish a benchmark for aerial images. Liu et al. [19] proposed using a hierarchical Wasserstein distance because the cross-entropy loss that is usually calculated for image classification simply compares the predicted probability with ground truth, ignoring the interclass relationship between images. The inclusion of hierarchical Wasserstein distance achieved better results for aerial image classification. Minetto et al. [22] proposed using ensemble CNNs for aerial image classification. As an initial baseline, ResNet and DenseNet models were trained without any augmentation techniques. These baseline models were fine-tuned with different augmentation techniques and cropping styles to generate a variety of weights. An ensemble majority voting technique was used to get the final predicted class. Muhammet Ali et al. [6] adapted a snapshot ensemble model for aerial image classification. Stochastic gradient descent (SGD) with warm restarts was used to generate weights from several local minima points. The weights were then used for a snapshot ensemble model. Liao et al. [16] proposed a label-splitting strategy and assigned appearance labels using a KD-tree algorithm. These appearance labels were used for training the network with triplet loss. Zhang et al. [33] proposed using Multi-Head Self-Attention (MHSA) for aerial images. Feature maps were extracted using the ResNet architecture. The feature maps were then sent to a transformer encoder block that has layer normalization, MHSA, and a feed forward network.

Given that SAR sensors are not as popular as EO sensors in modern remote sensing systems, there are fewer works for ATR aerial image classification using images captured from SAR sensors. Chen et al. [4] proposed using a sparsely connected CNN, instead of a typical CNN, for SAR data. A typical CNN would have a large number of parameters and overfit limited SAR data. Ding et al. [7] improved SAR image classification by augmenting the limited SAR datasets with image translation, random speckle noise, and pose synthesis. Lin et al. [17] proposed a novel convolutional highway unit inspired by long short-term memory (LSTM) networks. There is an adaptive gating mechanism in each convolutional highway layer that reduces the number of parameters in the layer when compared to a traditional convolutional layer. The reduced number of parameters makes it easier to train the model on a limited dataset. Wang et al. [30] proposed a modified Squeeze-and-Excitation block which leads to better feature extraction from SAR images.

A multi-modal approach incorporating both EO and SAR sensor data can be leveraged to improve ATR aerial image classification accuracy. Furthermore, in low light or adverse weather conditions where EO data is unsuitable, SAR data can be used as a supplement. In recent years, the MAVOC challenges in NTIRE 21 [18] and PBVS 22 [21] have sparked interest in combining EO and SAR data for ATR aerial image classification. There are now many publications that are based on the datasets shared in the MAVOC challenges [13, 15, 23, 28, 31, 32]. Yang et al. [31] proposed an ensemble learning framework with a cascaded expert branch and a parallel expert branch. The cascaded expert branch has multiple ResNet-50 models that learn different features from the same input. The parallel expert branch is trained on a dataset that has been re-balanced using the strategy proposed in [9]. In both the cascaded expert branch and the parallel expert branch, a decision-based voting fusion is done to predict the class of each image. Miron et al. [23] proposed an efficient CNN architecture for the NTIRE 21 MAVOC challenge [18]. The architecture achieved 26.51% top-1 accuracy with a 0.02 second CPU runtime and only 0.3 million parameters. Li et al. [15] proposed a two-stage shake-shake network to address the class imbalance in the PBVS 22 MAVOC challenge [21] training dataset. Inspired by shake-shake regularization [8], the authors proposed a regularization term γ. The γ regularization term is learned by a residual block during training to prevent overfitting. In the first training stage, the model is trained on all of the datasets. In the second training stage, the dataset is balanced, the classifier layer (fully connected layer) is trained, and the weights of the other layers are frozen. Yu et al. [32] proposed a pseudo-labeling strategy based on the k-means++ [2] scene clustering algorithm. The clustering is done in post-processing to mitigate bias caused by using multiple models. The strategy achieved the highest top-1 accuracies in both Track 1 and Track 2 of the PBVS 22 MAVOC challenge [21]. Udupa et al. proposed a multi-modal domain fusion strategy for the PBVS 22 MAVOC challenge [21]. The primary objective of domain fusion is to build a domain invariant model. To learn the domain gap between EO and SAR sensor data, sliced Wasserstein discrepancy [14] was used as a loss function.

3. Proposed Method

This section outlines the details of our proposed framework and is organized into two subsections. The first subsection details our label splitting strategy and the second subsection details our network architecture and loss function.

3.1. Label Splitting Strategy

The training datasets provided in the NTIRE 21 [18], PBVS 22 [21], and PBVS 23 MAVOC challenges [1] all have a long-tail class distribution. Data augmentation techniques were performed on classes containing less than 6,000

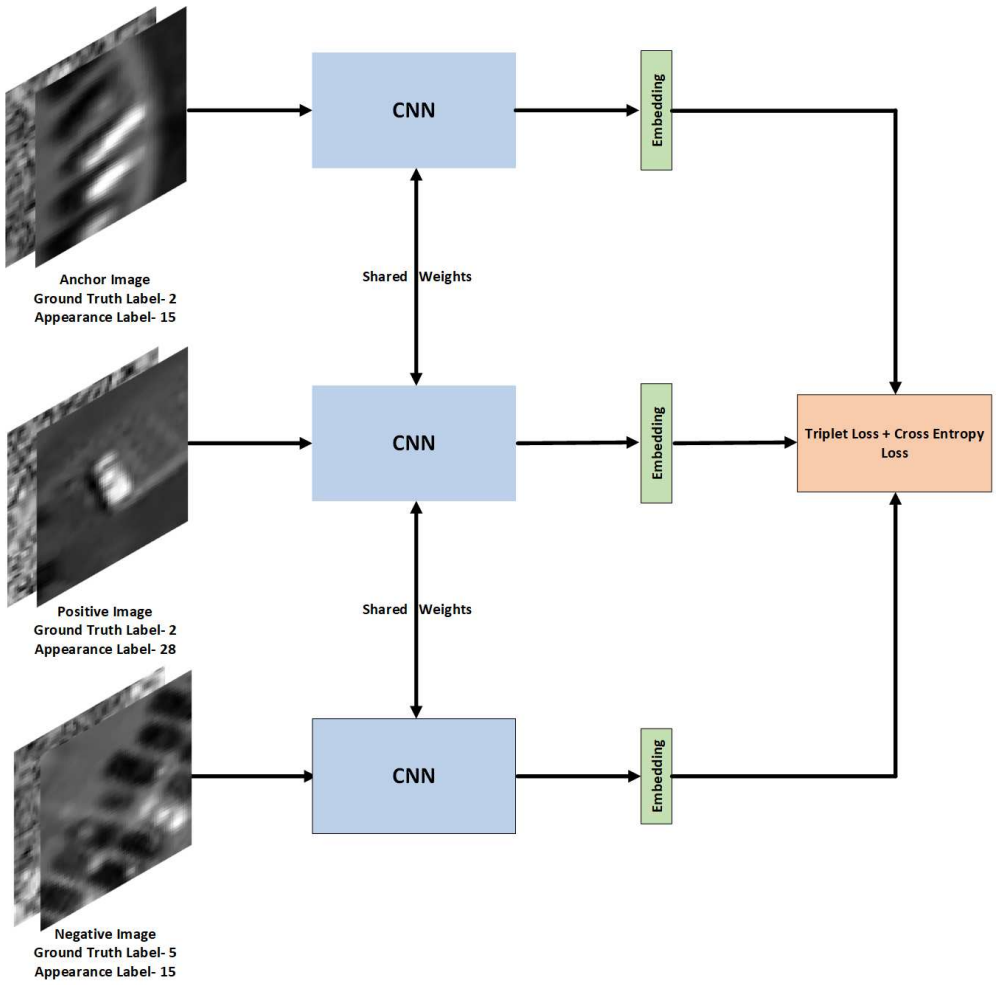


Figure 2. Proposed Training Framework for Track 2

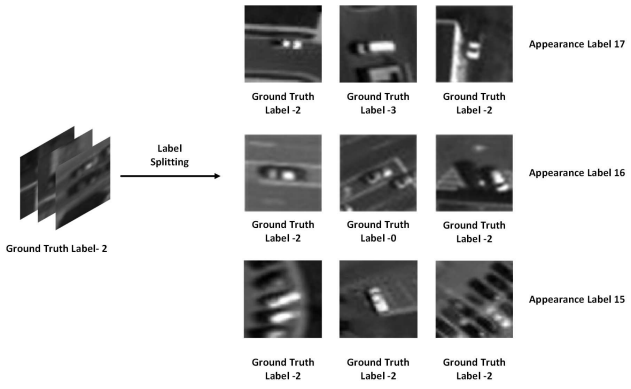


Figure 3. Label Splitting for EO Images

images to overcome the long-tail class distribution. The data augmentation techniques that were performed included flips, rotations, and affine transformations. For classes with greater than 6,000 images, 6,000 images were randomly selected.

Fig. 1 shows the high intraclass variability and low interclass variability that is common in aerial image datasets. To overcome these variabilities, the network was trained using triplet loss. Triplet loss is achieved through mining triplet pairs. In our work, inspired by Liao et al. [16], we implemented a triplet mining technique called label splitting. Label splitting groups visually similar images together by labeling them with the same appearance label. Since the goal of triplet loss is to group instances of the same class closer together and instances of different classes further apart, this improves the quality of the triplet pairs. Positive triplet pairs are images from the same class with different appearance labels. Negative triples pairs are images from different classes with the same appearance labels.

Our proposed method is shown in Fig. 2. Feature extraction is performed on VGG-16 [26] fully-connected layers for both EO and SAR images. The dimensionality of the features is then reduced using principal component analy-

sis (PCA). Reducing the dimensionality of the features decreases the computation required for the subsequent KD-tree and makes the features more robust to noise. PCA is performed separately for EO and SAR features. After PCA is performed, the EO and SAR features are combined as input to the KD-tree.

KD-tree is an unsupervised clustering algorithm. Each image is clustered into a KD-tree node and the node number is the label that is assigned to the image. The number of images in a given KD-tree node depends upon how many images have a similar visual appearance. The number of KD-tree nodes depends upon the depth of the KD-tree. In our case, the labels generated from the KD-tree are appearance labels. To select the triplet mining anchors, an image was randomly chosen from each tree node. To determine the positive triplet pairs for a given anchor, images with the same ground truth label are selected from different tree nodes than the anchor tree node. To determine the negative triplet pairs for a given anchor, images with different ground truth labels are selected from the same tree node as the anchor tree node. This strategy for determining the positive and negative triplet pairs ensures that we choose triplet pairs that are visually similar to each other. Fig. 3 shows how label splitting groups visually similar images together by labelling them with the same appearance label

3.2. Network Architecture and Loss Function

This section describes our proposed network architecture and loss function. The highest performing single model is a pre-trained ResNet-34 [11]. The loss function is a combination of triplet loss and cross-entropy loss. Triplet loss minimizes the intraclass distance and maximizes interclass distance. Cross-entropy loss also maximizes interclass distance. The equations for loss functions are defined below:

$$L_{CE} = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \hat{x_{a\,i}} + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \hat{x_{a\,i}} + b_j}} \quad (1)$$

$$L_{triplet} = max(d(\hat{x}_a, \hat{x}_p) - d(\hat{x}_a, \hat{x}_n) + margin, 0) \quad (2)$$

$$L_{multi-loss} = \alpha * L_{triplet} + (1 - \alpha) * L_{CE} \quad (3)$$

where $\hat{x_{ai}} \in \mathbb{R}^d$ is the ith feature that belongs to the y_ith class. d, $W \in \mathbb{R}^{d \times n}$, and $b \in \mathbb{R}^d$ denote the feature dimension, last connected layer, and bias term, respectively. $\hat{x_a}$, $\hat{x_p}$, and $\hat{x_n}$ are the anchor, positive image, and negative image, respectively. The regularization term, or α, used for training the multi-loss loss function was 0.8.

Fig.2 shows our proposed Track 2 training framework. The anchor and positive triplet pairs share the same ground truth label; however, the anchor has a different appearance label. The anchor and negative triplet pairs have different ground truth labels but share the same appearance label. The anchors, positive triplet pairs, and negative triplet pairs are given as input to the CNN and trained together using the same network weights. The initial convolutional layer has 1 channel for Track 1 and 2 channels for Track 2. After extracting the embeddings, or features, from the CNN, the model is optimized with a multi-loss function that combines triplet loss with cross-entropy loss. For both Track 1 and Track 2, the embeddings generated by the fully-connected layer have a dimension of 512 for calculating the triplet loss. A second fully-connected layer with a dimension of 10 was added for calculating cross-entropy loss.

4. Experimental Results

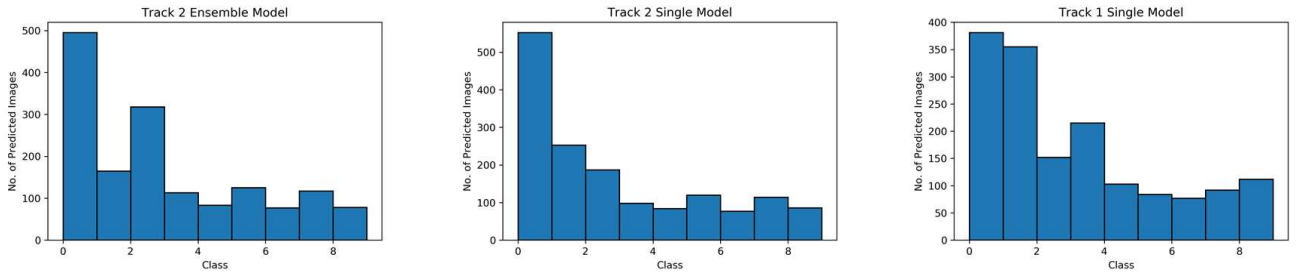
4.1. Dataset

Our proposed method was trained, evaluated, and tested on the PBVS 23 [1] MAVOC challenge dataset. The public dataset includes ground truth labels for training, but does not include ground truth labels for validation or testing. Participants evaluate model performance by submitting their pre-trained model on the challenge website during the validation phase to receive a validation result and during the testing phase to receive a testing result. The validation phase is limited to 60 submissions per participant and the testing phase is limited to 6 submissions per participant. Although the challenge contains two tracks, the dataset is the same for both tracks. Both Track 1 and Track 2 allow submissions to be trained on both the given EO data and the given SAR data. During evaluation and testing, Track 1 evaluates model performance using only SAR labels and Track 2 evaluates model performance using both EO and SAR labels.

Sample images from the dataset can be seen in Fig.1. The images in the dataset are small regions of larger images taken by EO and SAR sensors mounted on multiple aircraft. The EO sensor typically has a spatial resolution of 31 $\times$ 31 and the SAR sensor typically has a spatial resolution of 56 $\times$ 56. As shown in Table 1, the training dataset is severely imbalanced. The first 4 classes account for almost 98% of the dataset size. However, the validation and testing datasets are balanced with approximately the same number of images per class.

4.2. Implementation Details

Our work balanced the training dataset through data augmentation and random sampling. After balancing the training dataset, all of the images in the training dataset were resized to 224 $\times$ 224 and inputted into the pre-trained VGG-16 [26] model for feature extraction. The features were then clustered using a KD-tree of depth 7 to create 2^7, or 128, appearance labels. Inspired by Liao et al. [16], the ap-



A) Histogram of No. of Predicted Images per Class for Track 2 Ensemble Model

B) Histogram of No. of Predicted Images per Class for Track 2 Single Model

C) Histogram of No. of Predicted Images per Class for Track 1 Single Model

Figure 4. Histograms of No. of Predicted Images per Class for Different Tracks and Models in the Development Phase

Class	Class Name	# Training Samples
0	Sedan	364,228
1	SUV	43,642
2	Pickup Truck	24,420
3	Van	17,159
4	Box Truck	3,414
5	Motorcycle	2,351
6	Flatbed Truck	1,233
7	Bus	1,130
8	Pickup Truck w/ Trailer	971
9	Flatbed Truck w/ Trailer	714

Table 1. PBVS 23 MAVOC Challenge Training Dataset Class Distribution

pearance labels were then used to construct approximately 500,000 optimized triplet pairs for triplet loss. After the pre-processing steps were completed, the model was trained with both triplet loss and cross-entropy loss. Although the highest performing single model was ResNet-34 with pre-trained ImageNet weights, experiments were also performed on other models such as the EfficientNet-B0 [27] and Swin-T [20]. The experiments were implemented using PyTorch on an NVIDIA RTX A6000 GPU. Each model was trained for 5 epochs with a batch size of 64, learning rate of 1e-4, and Adam optimizer with default parameters. It took around 5 hours to train each model.

4.3. Results

Table 2 shows the results of Track 2 during the development, or validation, phase. The 1,571 images in the validation dataset are equally distributed across the 10 classes. Table 2 shows the final weighted score for each model in the development phase. The weights used for the final score were not made public for the participants. The single model with the highest total score and highest top-1 accuracy was ResNet-34. This model outperformed the EfficientNet-B0 model, the Swin-T model, and the provided baseline model that was given by the competition. The baseline model was a ResNet-18 model that was trained with cross-entropy loss. Our ensemble model outperformed the baseline model in top-1 accuracy by 247%. The ensemble model consists of ResNet-34, EfficientNet-B0, and Swin-T. The final output from the ensemble model is the weighted sum of scores from the constituent models. ResNet-34 is weighted by 0.5, EfficientNet-B0 is weighted by 0.23, and Swin-T is weighted by 0.25. Compared to ResNet-34, the ensemble model final score is 0.01 higher and the ensemble model final accuracy is 1% higher.

Table 3 shows the Track 2 test phase results. The test dataset contains 5,745 images that are equally distributed across the 10 classes. Our ensemble model ranked 3rd out of all of the Track 2 submissions.

Table 4 shows the Track 1 development phase results. The Track 1 development phase results only use SAR images as input to the network. The ResNet-34 model achieved a top-1 accuracy of 64.29. Similar to Track 2, the ResNet-34 model achieved better performance than the EfficientNet-B0 and Swin-T models. Due to the limited timeframe of the competition, we did not have time to submit the ensemble model like we did for Track 2.

Table 5 shows the Track 1 test phase results. Our ResNet-34 model placed 3rd in the competition.

Fig. 4 shows histograms of the number of predicted images per class during the development phase. Although the ground truth labels are unknown, the fact that there is an equal distribution of classes in the validation dataset is known. It is clear from Fig. 4 that the distribution of classes is not equal in the model predictions. All of our models are biased toward class 0, the sedan class.

5. Conclusion

Inspired by Liao et al. [16], this work proposes a training strategy that involves the creation of appearance labels to generate triplet pairs for training the network with both triplet loss and cross-entropy loss. The appearance labels

Model	Final Score	Accuracy (top-1)%	AUROC	TNR at TPR95	Run Time on CPU
Baseline	0.28	21.82	0.48	0.03	-
ResNet-34	0.77	74.81	0.84	0.20	0.028
EfficientNet-B0	0.75	74.03	0.79	0.26	0.024
Swin-T	0.69	69.35	0.70	0.01	0.074
Ensemble Model	0.78	75.84	0.84	0.18	-

Table 2. PBVS 23 MAVOC Development Phase Results for Track 2

Team	Final Score	Accuracy (top-1)%	AUROC	TNR at TPR95
Team A	0.84	89.60	0.68	0.02
Team B	0.84	88.77	0.68	0.03
Team C	0.74	69.80	0.85	0.56
Our Ensemble Model	**0.71**	**71.40**	**0.70**	**0.05**
Team D	0.71	68.35	0.79	0.10
Team E	0.71	68.35	0.79	0.10
Team F	0.71	68.80	0.76	0.07
Our ResNet-34 Model	**0.70**	**68.90**	**0.74**	**0.10**

Table 3. PBVS 23 MAVOC Test Phase Results for Track 2

are created using the KD-tree algorithm. Positive triplet pairs are defined as having the same ground truth label, but a different appearance label, when compared to an anchor image. In contrast, a negative triplet pair is defined as having the same appearance label, but a different ground truth label, when compared to an anchor image. Triplet loss minimizes the high intraclass variability and low interclass variability of aerial images. We constructed approximately 500,000 optimized triplet pairs for triplet loss. After the pre-processing steps were completed, we conducted experiments using the PBVS 23 [1] MAVOC challenge dataset and the ResNet-34, Swin-T, EfficientNet-B0, and ensemble learning models.

The PBVS 23 [1] MAVOC challenge contains two tracks. Both Track 1 and Track 2 allow submissions to be trained on both the given EO data and the given SAR data. During evaluation and testing, Track 1 evaluates model performance using only SAR labels and Track 2 evaluates model performance using both EO and SAR labels. During the development phase, our ResNet-34 model achieved a top-1 accuracy of 64.29% for Track 1 and our ensemble learning model achieved a top-1 accuracy 75.84% for Track 2. These values are 542% and 247% higher than the baseline values. Overall, this work[1] ranked 3rd in both Track 1 and Track 2.

[1]This work is supported in part by NSF Award 2148382 and AFRL SFFP at USAF Academy.

References

[1] PBVS 23. Pbvs 23 workshop website. https://pbvs-workshop.github.io/challenge.html, 2023. 3, 5, 7

[2] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 3

[3] General Atomics. General atomics. https://www.ga-ccri.com/how-can-sar-imagery-improve-on-optical-overhead-imagery, 2023. 1

[4] 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. 1, 3

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 1, 3

[6] Muhammet Ali Dede, Erchan Aptoula, and Yakup Genc. Deep network ensembles for aerial scene classification. *IEEE Geoscience and Remote Sensing Letters*, 16(5):732–735, 2019. 3

[7] Jun Ding, Bo Chen, Hongwei Liu, and Mengyuan Huang. Convolutional neural network with data augmentation for sar target recognition. *IEEE Geoscience and Remote Sensing Letters*, 13(3):364–368, 2016. 1, 3

[8] Xavier Gastaldi. Shake-shake regularization. *CoRR*, abs/1705.07485, 2017. 3

[9] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. *CoRR*, abs/1908.03195, 2019. 3

Model	Final Score	Accuracy (top-1)%	AUROC	TNR at TPR95	Run Time on CPU
Baseline	0.22	10.00	0.57	0.07	-
ResNet-34	0.68	64.29	0.78	0.13	0.028
EfficientNet-B0	0.62	59.35	0.71	0.09	0.024
Swin-T	0.62	60.13	0.67	0.05	0.074

Table 4. PBVS 23 MAVOC Development Phase Results for Track 1

Team	Final Score	Accuracy (top-1)%	AUROC	TNR at TPR95
Team A	0.65	63.20	0.71	0.03
Team B	0.64	59.20	0.80	0.35
Team C	0.61	53.15	0.85	0.50
Our ResNet-34 Model	**0.61**	**59.85**	**0.64**	**0.05**

Table 5. PBVS 23 MAVOC Test Phase Results for Track 1

[10] Xiaobing Han, Yanfei Zhong, Liqin Cao, and Liangpei Zhang. Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing*, 9(8), 2017. 1, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[12] Chowdhury Sadman Jahan, Andreas Savakis, and Erik Blasch. Cross-modal knowledge distillation in deep networks for sar image classification. In *Geospatial Informatics XII*, volume 12099, pages 20–27. SPIE, 2022. 1

[13] Chowdhury Sadman Jahan, Andreas Savakis, and Erik Blasch. Sar image classification with knowledge distillation and class balancing for long-tailed distributions. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5, 2022. 3

[14] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. *CoRR*, abs/1903.04064, 2019. 3

[15] Gongzhe Li, Linpeng Pan, Linwei Qiu, Zhiwen Tan, Fengying Xie, and Haopeng Zhang. A two-stage shake-shake network for long-tailed recognition of sar aerial view objects. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 248–255, 2022. 1, 3

[16] Rijun Liao, Zhu Li, Shuvra S. Bhattacharyya, and George York. Aerial image classification with label splitting and optimized triplet loss learning. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2021. 1, 2, 3, 4, 5, 6

[17] Zhao Lin, Kefeng Ji, Miao Kang, Xiangguang Leng, and Huanxin Zou. Deep convolutional highway unit network for sar target classification with limited labeled training data. *IEEE Geoscience and Remote Sensing Letters*, 14(7):1091–1095, 2017. 3

[18] Jerrick Liu, Nathan Inkawhich, Oliver Nina, Radu Timofte, Yuru Duan, Gongzhe Li, Xueli Geng, and Huanqia Cai. Ntire 2021 multi-modal aerial view object classification challenge. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 588–595, 2021. 1, 3

[19] Yishu Liu, Ching Y Suen, Yingbin Liu, and Liwang Ding. Scene classification using hierarchical wasserstein cnn. *IEEE Transactions on Geoscience and Remote Sensing*, 57(5):2494–2509, 2018. 1, 3

[20] 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. *CoRR*, abs/2103.14030, 2021. 6

[21] Spencer Low, Oliver Nina, Angel D. Sappa, Erik Blasch, and Nathan Inkawhich. Multi-modal aerial view object classification challenge results - pbvs 2022. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 349–357, 2022. 1, 3

[22] Rodrigo Minetto, Maurício Pamplona Segundo, and Sudeep Sarkar. Hydra: an ensemble of convolutional neural networks for geospatial land classification. *CoRR*, abs/1802.03518, 2018. 1, 3

[23] Casian Miron, Alexandru Pasarica, and Radu Timofte. Efficient cnn architecture for multi-modal aerial view object classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 560–565, 2021. 1, 3

[24] Raghunath Sai Puttagunta, Renlong Hang, Zhu Li, and Shuvra Bhattacharyya. Low resolution recognition of aerial images. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019. 1, 3

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. 2

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4, 5

[27] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 6

[28] Sumanth Udupa, Aniruddh Sikdar, and Sundaram Suresh. Multi-modal domain fusion for multi-modal aerial view object classification. 12 2022. 1, 3

[29] Simon A. Wagner. Sar atr by a combination of convolutional neural network and support vector machines. *IEEE Transactions on Aerospace and Electronic Systems*, 52(6):2861–2872, 2016. 1

[30] Li Wang, Xueru Bai, and Feng Zhou. Sar atr of ground vehicles based on esenet. *Remote Sensing*, 11(11), 2019. 1, 3

[31] Cheng-Yen Yang, Hung-Min Hsu, Jiarui Cai, and Jenq-Neng Hwang. Long-tailed recognition of sar aerial view objects by cascading and paralleling experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 142–148, 2021. 1, 3

[32] Jun Yu, Hao Chang, Keda Lu, Liwen Zhang, and Shenshen Du. Scene clustering based pseudo-labeling strategy for multi-modal aerial view object classification. 05 2022. 1, 3

[33] Jianrong Zhang, Hongwei Zhao, and Jiao Li. Trs: Transformers for remote sensing scene classification. *Remote Sensing*, 13(20):4143, 2021. 1, 3