NTIRE 2021 Multi-modal Aerial View Object Classification Challenge

Jerrick Liu Nathan Inkawhich Oliver Nina Radu Timofte Yuru Duan
Gongzhe Li Xueli Geng Huanqia Cai

Abstract

In this paper, we introduce the first Challenge on Multi-modal Aerial View Object Classification (MAVOC) in conjunction with the NTIRE 2021 workshop at CVPR. This challenge is composed of two different tracks using EO and SAR imagery. Both EO and SAR sensors possess different advantages and drawbacks. The purpose of this competition is to analyze how to use both sets of sensory information in complementary ways. We discuss the top methods submitted for this competition and evaluate their results on our blind test set. Our challenge results show significant improvement of more than 15% accuracy from our current baselines for each track of the competition.

1. Introduction

Automatic target recognition (ATR) is a well-known problem in the area of computer vision. Although there has been significant progress in this area in recent years, a large amount of work remains, particularly in the area of ATR for aerial view images. For this challenge, we define ATR for aerial view imagery as the classification of targets in image chips cropped from larger frames. ATR from aerial view images presents a unique set of challenges due to the nature of the target and the ratio of the target size to the background of the image. There are many other issues that are specific to ATR in aerial view images that include: lower resolution of the target, target texture, light reflectance, etc. We consider these issues in our first Multi-modal Aerial View Object Classification Challenge (MAVOC) whose primary goal is to spark new innovations in the area of ATR research through development of new algorithms that utilize different types of sensors that could complement each other for target classification tasks.

Modern remote sensing (RS) systems are equipped with a variety of sensor types which ultimately define the data modality that the ATR algorithms operate in. One popular modality is electro-optical (EO), which effectively captures images in the visible spectrum. These images are human interpretable and the collection of such data, although challenging, is still rather straightforward (e.g. one may use the Google Earth [2] platform to collect aerial imagery in the EO domain). In the past, there have been many research efforts for ATR in the EO domain [13, 24, 20, 15]. Another popular sensor modality for ATR algorithms to work with is synthetic aperture radar (SAR) data. While being less human interpretable, SAR has the benefit of operating at night and in varying weather conditions, which gives it a distinct advantage over EO sensors in certain applications [18]. However, the collection process of SAR data is much more complex and expensive, meaning, the extent of SAR-ATR research lags behind EO-ATR research in several respects [11, 12, 10, 21, 18, 5, 17]. It is also worth mentioning a few other modalities common in ATR research: hyper-spectral, multi-spectral, and infrared [1]; but for the purposes of this challenge, we focus exclusively on EO and SAR data.

Despite the advantages and disadvantages of each individual data modality, a traditional RS system often leverages only a single modality. Thus, an area of significant promise is designing ATR algorithms that utilize multiple data modalities. This is promising because a multi-modal ATR system may be able to mitigate the drawbacks associated with each sensor-type. For instance, EO sensors have a fundamental dependency on visible light which allows them to capture details of the target such as color and texture. On the other hand, SAR sensors do not rely on light as they are self illuminating. Using both EO and SAR would allow for accurate imaging regardless of background lighting. SAR sensors are unable to resolve details as finely as EO sensors due to their lower operating frequencies. In the case of moving vehicles, the signatures will be acceptable in the EO domain but will be different in the SAR domain due to the properties of the radar systems and the algorithms used to form the back-projected SAR images. The goal of this challenge is to develop multi-modal ATR algorithms that leverage both EO and SAR data. Intuitively, having information from both sensor types will provide more feature-

*Oliver Nina (oliver.nina.1@afresearchlab.com), Bob Lee, Jerrick Liu, Chris Menart, Nathan Inkawhich, and Radu Timofte are the NTIRE 2021 challenge organizers. The other authors participated in the challenge.
Appendix A contains the authors' team names and affiliations.
https://data.vision.ee.ethz.ch/cvl/ntire21/

rich information to the ATR algorithm so that it can achieve higher overall performance. We also hope to encourage the innovation of fundamentally new ATR techniques that look beyond a "straightforward" integration of the data types.

Ultimately, the MAVOC challenge is divided into two tracks, each with a slightly different emphasis. The focus of the first track is to train a maximally accurate classifier of SAR data. While labeled EO and SAR data are both available during training time, the classifier is tested exclusively on SAR data. The idea behind this track is to encourage the development of a classifier that can learn from both EO and SAR data, but whose primary focus is to improve on SAR classification. This could be useful when only SAR data is available at test time. The second track focuses on the development of a maximally accurate classifier of both EO and SAR data that takes EO/SAR pairs of images as input during both training and test time. The idea behind this track is to propose new classifiers that can leverage EO and SAR data to improve on EO and SAR classification when both are available.

For both tracks, participants are scored mainly on the accuracy of their method's design while also considering the creativity and novelty of their approach. Additionally, for both tracks, participants may use both data sets for training. As a part of this challenge problem, we also release a new dataset that has paired EO+SAR data representations. Details on our new dataset will be released in a separate future publication. However, we give a brief overview of our dataset in this paper. We believe our dataset is unique and will help practitioners to perform better experiments when studying EO and SAR classification tasks.

As a result of this challenge, many potential solutions to our multi-modal ATR problem have been proposed. For instance, the winner of track SAR, proposed a feature fusion technique used to generate new features to create diverse data. The winner of track EO+SAR, used a data augmentation, batch balance, and noisy student technique to improve accuracy. Overall, through execution of this challenge we have improved our baselines more than 15% in each track.

The remainder of this document is organized as follows: Section 2 describes our challenge and different tracks in the challenge, as well as our dataset, Section 3 explains the challenge results, and Section 4 explains the winning submissions for each track. We conclude with Section 5.

2. Challenge

Our first Multi-modal Aerial View Object Classification (MAVOC) challenge was held jointly with the New Trends in Image Restoration and Enhancement (NTIRE) workshop. This challenge is one of the NTIRE 2021 associated challenges: nonhomogeneous dehazing [4], defocus deblurring using dual-pixel [3], depth guided image relighting [6], image deblurring [19], multi-modal aerial view imagery classification [14], learning the super-resolution space [16], etc.

Our MAVOC challenge presents different tasks for predicting the class label of aerial low resolution images based on a set of prior examples of images and their class labels. We separate these tasks into two main tracks:

2.1. Track 1: SAR

The first competition track focuses on classification of SAR data. The goal is to train a classifier that is maximally accurate on a held-out test set of SAR chips from 10 classes (see Table 1). Participants are welcome to use both the EO and SAR training data sets to accomplish this task. Due to the nature of SAR images, achieving high accuracy on these chipped SAR images is non-trivial. We have provided a baseline model for track 1 as a comparison. See table 2 for our baseline accuracy.

2.2. Track 2: EO + SAR

The second competition track focuses on classification of EO and SAR data. As opposed to just SAR, the goal of this track is to train a classifier on either the SAR or the EO domain that is maximally accurate on its own domain as well as the other domain. This track requires contestants to jointly utilize the EO and SAR datasets. We provided a baseline model for track 2 for contestants to use as a comparison. See Table 3 more details regarding our baseline accuracy.

2.3. Dataset

The data for this challenge consists of two types of small windowed regions (chips) generated from large images captured by several aircraft mounted EO and SAR sensors. The EO chips are 31×31 px images. The SAR chips cover the same approximate field of view as the corresponding EO images and have a finer resolution than EO images. Due to the SAR processing the chips vary in pixel size but are generally around 55×55 px. Figure 1 provides samples of EO and SAR chips. The targets belong to a list of 10 classes that correspond to a training set of non-uniformly distributed number of samples per class (see below) whereas the validation set is based on a small uniformly distributed number of samples per class.

The dataset is divided into:

- Training set: This set resembles the data which is non-uniform and imbalanced (i.e. some classes have more samples than others)
- Validation set: This set is a uniformly distributed among all classes with < 100 samples per class
- Test set: This split resembles the validation test with a uniform distribution of testing images among the classes.

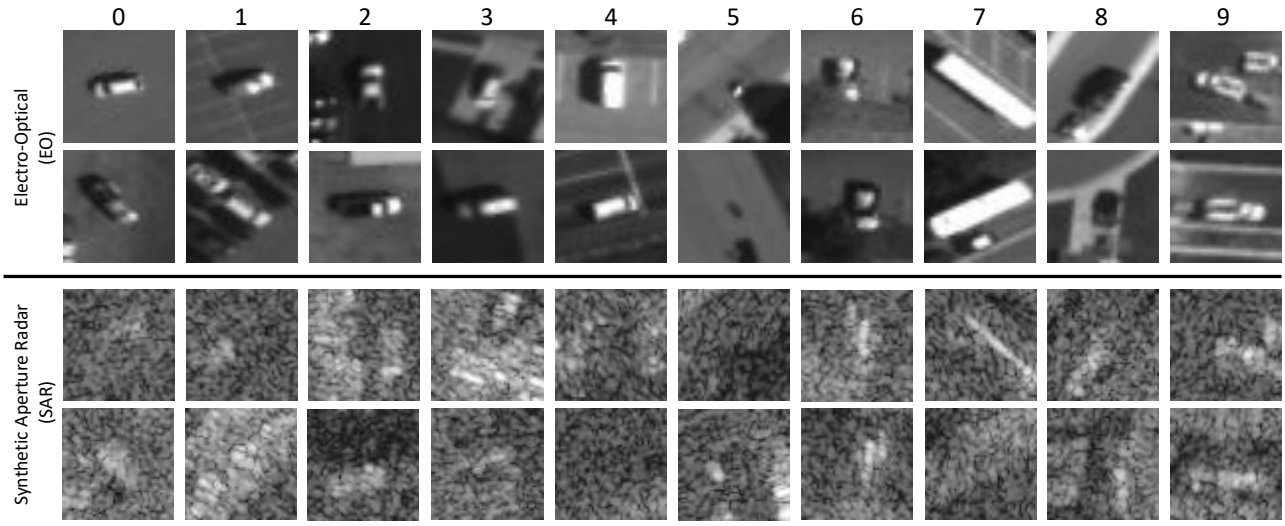

Figure 1: Two sample pairs of EO and SAR chips from each of the 10 classes in the Unicorn Dataset.

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

Class #	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

The images belong to one of 10 categories: 0 to 9. The train data contains both SAR and EO images and the class labels. The valid/test data contains SAR images for the SAR track and both SAR and EO images for (SAR+EO) track of the challenge. The purpose is to use the provided (SAR+EO) train image to maximize the classification accuracy when the inputs are only SAR images or both SAR and EO images. See Table 1 for a breakdown on the images for each class. Note that instead of the class name now it is simply labeled 0-9.

2.4. Evaluation

The evaluation is comprised of the comparison between the predictions with the reference ground truth labels. As often employed in the literature, the standard classification accuracy top-1 percent will be used. For each dataset, the results over all the processed images belonging to it will be reported.

2.5. Challenge Phases

The challenge was separated into two distinct phases, development and testing. The development phase began on Jan 1, 2021 and continued until March 15, 2021 when the testing phase began. The testing phase concluded on March 20, 2021, concluding the challenge.

3. Challenge Results

Throughout our competition, there were 163 participants in Track 1, and 160 participants in Track 2. During the development phase, a total of 827 algorithms were submitted for Track 1, and 593 algorithms were submitted for Track 2. During the Testing phase, there were a total of 50 submissions. This section reports the five best performing algorithms submitted during the testing phase. The results are organized by track.

3.1. Baselines

For the competition, we provided a simple AlexNet baselines model pretrained on ImageNet with no data augmentation done on the training and testing dataset. The baseline results are shown in Tables 2 and 3.

3.2. Track SAR Results

The top three ranked teams for track 1 SAR achieved improved results when compared to our baseline results. The methods used by these teams are described in great detail in the following sections. The top ten ranked teams for Track 1 SAR are as follows:

Table 2: Top-10 Teams for Track 1 (SAR)

Rank	Team	Accuracy
1	**MISL-SAR**	**34.62**
2	Moyu	26.63
3	Sol Cummings	26.39
4	UW-IPL	26.03
5	Ga_z_a	25.06
6	Dian	24.82
7	BONG	24.58
8	Zhangxs	23.61
9	Oooo0	23.37
10	LeonShangguan	23.00
	SAR Baseline	15.87

3.3. Track EO + SAR Results

Similarly, the top ten ranked teams for track 2 EO+SAR achieved improved results when compared to our baseline. The methods used by these teams are described in great detail in the following sections. This track saw fewer successful submissions as teams ranked 10 and above scored an accuracy of 0. Thus, this table only contains teams ranked 1-9. The top nine ranked teams for track 2 are as follows:

Table 3: Top-9 Teams for Track 2 (EO + SAR)

Rank	Team	Accuracy
1	**XD-IPIU**	**46.85**
2	CVPRer	34.63
3	Casian	26.51
4	MichaelXin	26.03
5	LeonShangguan	25.06
6	Xsourse	23.97
7	UW-IPL	21.07
8	Benjamin666	20.70
	EO + SAR Baseline	19.23
9	Vamshi	17.01

4. Challenge Methods

The methods to train a classifier to classify SAR and EO/SAR pair images varied from team to team. In this section we describe the methods used by the top three teams of each track.

4.1. Track SAR Methods Descriptions

Classifying SAR images has been historically a non-trivial task. The methods used by the top three teams provide valuable insight on the future of SAR classification. These methods span from using techniques such as feature fusion, data augmentation and multi-stage model training. We describe each method of the top three competitors in detail in the following sections.

4.1.1 Rank 1: Team MISL-SAR

Team MISL-SAR used a ResNet [7] based architecture. A total of 587,544 (SAR+EO) images were used in the experiment. Because the data distribution is severely unbalanced, with "sedan" accounting for 79.7% of the data and "bus" only accounting for 0.2% of the data, the team randomly selected 6,000 SAR images and corresponding 6,000 EO images from each category as the data set, of which 5,000 were used for training the model and 1,000 were used for testing results. For categories with less than 6,000 images, a simple data enlargement was performed.

The team then, randomly selected 1,000 images with known labels to predict. The correct number of predictions is calculated and the prediction accuracy is obtained. During training, the team's data often contains redundant information. This redundant information has a negative effect on the feature learning effect of the neural network. Therefore, from this perspective, when designing their deep neural networks, the participants deliberately enhanced the ability of the deep neural network to eliminate redundant information. The team noted that ResNet is the better choice at the moment.

Furthermore, the team introduced the concept of feature fusion. Feature fusion is used to generate new fusion features from existing feature sets. The most diverse information can be obtained from the multiple original feature sets involved in the fusion. The team found that feature fusion can eliminate redundant information resulting from the correlation between different feature sets and make subsequent decisions possible.

The team further experimented with EfficientNetB1[22] and found that it performed well in the local test set, but the score was not ideal. Instead, ResNet152, which performed worse in the local test set than EfficientNetB1, scored slightly better in the final test. Moving forward with ResNet152 as the team's architecture of choice, the network was modified a little, and the fusion of EO image and SAR image was added. Although the local test set did not perform well, the final score was the best.

4.1.2 Rank 2: Team Moyu

Team Moyu proposed a novel three-stage training procedure by decoupling the information rich head-class data and the rest-classes data and transferring model's expression ability from parts classes dataset to all classes dataset: the team used the complete dataset for rough training, then

the team used classes 1-9 to train a model, and after that the team used class-balanced datasets to fine tune the whole model and classifier individually.

The team's training procedure contained three stages. The first stage was to use the whole dataset to train a rough model. The reason of using all images was that the network model can always be more general when observing more data. Besides, as the team used SGD as their optimizer, the momentum of SGD will drive the model to an area that is more smooth to class 0 because of the long-tailed distribution.

The second stage was to use a "class-0-removed" dataset to train a model having prominent feature expression ability about class 1-9 targets. By discarding class-0 images, the distribution of rest data was more balanced, thus resolving the most troublesome difficulty caused by long-tailed images.

However, due to stage 1, the model still memorized some information of class 0. Besides, because class-0 images account for more than 80% of the dataset, the class 1-9 dataset was quite small that made training much faster.

The third stage was to use a class-balanced dataset which contained the same number of 10 categories of images to fine tune the model. In this stage, the team established a sub-dataset containing 50,000 images (5,000 per class) by random sampling and image augmentation. First, the team trained the model 20 epochs to transfer the 9-class model to a 10-class model. Next, the team trained another 10 epochs by freezing backbone parameters and only adjusting the classifier. After these two sub-stages, the team's final model obtained good classification accuracy on all 10 categories in the validation set.

4.1.3 Rank 3: Sol Cummings

The third ranked team, Sol Cummings, proposed using MobileNetV3 [8] and CutMix [26] as well as many data augmentations to get better results.

The training strategies proposed by Sol aimed to address the following problems: an imbalanced dataset, the limited number of overall samples, and low image resolution.

Under sampling is used to combat the imbalance in classes within the training dataset. In the proposed solution, the number of samples used for each class is capped at approximately 1,400 samples. Tentative results indicated under sampling outperformed oversampling.

Data augmentation is then employed to artificially introduce more training samples. A modification to CutMix is proposed that mixes images but preserves the center region of patches, while not blending labels. The intuition of preserving the center region is to not interfere with the regions where the properties of each class are most prominent. Along with the modified CutMix augmentation, 90 degree rotations, horizontal/vertical flips, and random cropping were combined. The stem block in the MobileNetV3 large architecture was altered from stride 2 to 1 in order to preserve spatial resolution. The modification in the model architecture prevented the loss of information from the patches early on, and improved overall scores.

4.2. Track EO + SAR Method Descriptions

Further, will discuss the methods used by the top three ranked teams in Track 2 to perform image classification on the combination of EO + SAR images. The techniques used for the methods in this track range from: data augmentation, batch balance, low-level image processing and semi-supervised learning. We describe each method for the top three competitors in this track in detail.

4.2.1 Rank 1: Team XD-IPIU

Team XD-IPIU used a fairly straightforward training method that proved to be quite effective. The strategy consisted of

- Data Augmentation: Random crops, random brightness and contrast changes, random flips (horizontal and vertical) were applied to the dataset.
- Batch Balance: Due to the unbalanced nature of the dataset, the team balanced out the images in each class for each batch during each iteration of the training process.
- Noisy Student [25]: The team also applied a semi-self learning model to improve the robustness and generalization ability of the network through training.

4.2.2 Rank 2: Team CVPRer

Team CVPRer had a specific data processing step as a main component to their method and it is described as follows:

Data Processing: Given that provided training set is class unbalanced, while the test set is uniformly scattered in all classes, using all the training data could lead to a wrongly biased prediction according to the team's experiments. Further attempts to over-sample classes containing less data was proven ineffective, showing difficulty in network convergence as well as a decrease in accuracy. As a result, the team selected a subset of the training set, which has around 800 samples per class, to achieve a similar data distribution as the test set, and the team randomly sampled from it during training.

The team composed the input by resizing and concatenating 2 duplicated EO images and their matching SAR image together as a 3-channel image. Several data augmentation methods were applied to increase the data diversity. Due to the poor quality of the input SAR image, the team

also applied a median blur on the original SAR image before any further augmentation to reduce any noise level. Also, because EO+SAR image pairs are acquired through airborne devices, these images can be augmented through random rotation to simulate different relative positions and angles between the device and the target object when the images are produced. Besides, the team found that many training images contained a large amount of redundant information with its target in the middle of the image taking up only around 1/4 in space. To focus the network on the main target and prevent it from picking up useless, and even confusing, input regions, the team cropped the input image to 0.65 of it original size after rotation. Finally, the team normalized the input value to -1 to 1 for better numeric property.

As for training, the team used EfficientNetB6 [22], B7 [22], and B8 [22] pretrained on ImageNet data set and modified the network with a SE Block [9] and an ECA Block [23] on the selected subset of training data. The team only loaded the part of parameters with their corresponding structure unchanged. For new or changed layers, the team used Kaiming initialization. During training, 32x32 EO images and 57x57 SAR images were combined, augmented, and then resized to 299x299 to fit the input of the model. The batch size was then set to 24 using 2 NVIDIA Tesla P40 GPUs. The team minimized the cross-entropy loss with SGD optimizer using the "Reduce Learning Rate On Plateau" policy as learning rate schedule. The base learning rate was set to 10^{-3}.

To further improve the prediction on weak classes, the team applied semi-supervised learning. First, the team trained the model with labeled data in the selected training set. The model is then used with the unlabeled test set to predict the pseudo labels and add some of those to the training set according to prediction confidence. Finally, the team trained the model the same way as the first time to obtain the improved model.

4.2.3 Rank 3: Team Casian

Team Casian proposed a system architecture that consisted of a convolutional neural network with 11 convolutions, 1 max pooling layers and 3 residual blocks. The convolution layers used rectified linear unit (ReLU) activation function and were followed by batch normalization. Its kernel, the number of channels and parameters are 3×3, 64 and 36,928, respectively. The total number of parameters was 373,130. Moreover, the Adam optimizer was used for training the networks and the default hyper-parameters were used. The networks were trained for 100 epochs each. In order to compensate the uneven distribution of the dataset classes, the team used class weights that improved the class difference and do not allow the CNN to skew the results towards the class with the highest representation within the dataset. The team trained two convolutional neural networks: one for SAR images with the input $56 \times 56 \times 1$ and one for EO images with the input $32 \times 32 \times 1$. The result was given by averaging the predictions of the two networks and achieved reasonable results.

5. Conclusion

In this paper, we reported on our first Challenge on Multi-modal Aerial View Object Classification (MAVOC) in conjunction with the NTIRE 2021 workshop at CVPR. Object classification on EO and SAR domain is a non-trivial task and could be quite difficult in some instances. This is due to the nature of EO and SAR sensors and the type of signal used to capture and render their image frames.

The purpose of this challenge was to encouraged the development of novel techniques to improve image classification in electro-optical and synthetic aperture radar images in a multi-domain environment. The participation and number of submissions was impressive with hundreds of participants and submissions which sparked interest in this problem and produced innovative and interesting solutions. Overall this challenge proved to be successful in the number of submissions and in accuracy improvement upon our standard baselines with an improvement of more than 15% in accuracy in both tracks.

We would like to congratulate the winners of the NTIRE 2021 Multi-modal Aerial View Object Classification Challenge for achieving impressive results in their respective tracks. The methods submitted to this competition and discussed above allow us to look forward into continuing improving the design of algorithms that utilize and leverage multi-modal imagery for aerial view object classification.

Acknowledgements

We would like to thank Chris Menart (AFRL), Bret Minnehan (AFRL), Todd Rovito (AFRL), Bob Lee (WBI), Sahil Jain (UIUC), Peter Hall (OSU), Ryan Wolf (UIUC), Arvind Saligrama (Stanford) and Spencer Lowe (BYU) for their help with this competition in their different roles in organizing, preparing and hosting this competition. We would also like to give special thanks to each one of the participants of this competition and for their submissions.

A. Teams and Affiliations

NTIRE 2021 Team

Title:
NTIRE 2021 Multi-modal Aerial View Object Classification Challenge
Members:
Oliver Nina[1] *(oliver.nina.1@afresearchlab.com)*, Jerrick

Liu[2], Chris Menart[1], Bob Lee[3], Radu Timofte[5], Nathan Inkawhich.[4]
Affiliations:
[1] Air Force Research Laboratory, USA
[2] University of Illinois at Urbana-Champaign, USA
[3] Wright Brothers Institute, USA
[4] Duke University, USA
[5] Computer Vision Lab, ETH Zürich, Switzerland

Ranked 1 Team (Track SAR)

Method:
Feature Fusion
Team Name:
MISL-SAR
Members:
Yuru Duan[6] *(1216761190@qq.com)*, Wei Wei[6], Lei Zhang[6], Songzheng Xu[6], Yuxuan Sun[6], Jiaqi Tang[6]
Affiliations:
[6] Chang'an Campus, Northwestern Polytechnical University, Xi'an, Shaanxi Province, China.

Ranked 2 Team (Track SAR)

Method:
A Novel Three-stage Training Strategy for Long-Tailed Classification
Team Name:
Team Moyu
Members:
Gongzhe Li[7] *(gzLi20@buaa.edu.cn)*, Zhiwen Tan[7], Linpeng Pan[7]
Affiliations:
[7] Beihang University, China.

Ranked 3 Team (Track SAR)

Method:
Improving Performance of Lightweight Models Through Training Strategies
Team Name:
Sol Cummings
Members:
Sol Cummings[8] *(scummings@berkeley.edu)*
Affiliations:
[8] PASCO Corporation, Japan

Ranked 1 Team (Track EO+SAR)

Method:
EO and SAR Data Fusion
Team Name:
XD-IPIU
Members:
Xueli Geng[10] *(xlgeng@stu.xidian.edu.cn)*, Mengru Ma[10]
Affiliations:
[10]Key Laboratory of Intelligent Perception and Image Understanding (IPIU) of Xidian University, China.

Ranked 2 Team (Track EO+SAR)

Method:
Enhanced Efficientnet for EO and SAR Images Classification
Team Name:
CVPRer
Members:
Huanqia Cai[11] *(369454780@qq.com)*, Chengxue Cai[11]
Affiliations:
[11] Dengzhuang South Road, Haidian District, China.

Ranked 3 Team (Track EO+SAR)

Method:
Efficient CNN Architecture for Multi-modal Aerial View Object Classification
Team Name:
Team Casian
Members:
Casian Miron[12] *(casian_miron@yahoo.com)*, Alexandru Pasarica[12]
Affiliations:
[12]Gheorghe Asachi Technical University, Romania.

References

[1] 2021 IEEE GRSS Data Fusion Contest: Track DSE - Detection of Settlements without Electricity. https://www.grss-ieee.org/community/technical-committees/2021-ieee-grss-data-fusion-contest-track-dse/. Accessed: 2021-04-16. 1

[2] Google Earth. https://earth.google.com/web/. 1

[3] Abdullah Abuolaim, Radu Timofte, Michael S Brown, et al. NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[4] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2021 nonhomogeneous dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[5] 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, Aug 2016. 1

[6] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. NTIRE 2021 depth guided image relighting challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*, 2019. 5

[9] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. 6

[10] N. Inkawhich, E. Davis, M. J. Inkawhich, U. Majumder, and Y. Chen. Training sar-atr models for reliable operation in open-world environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–1, 2021. 1

[11] N. Inkawhich, E. Davis, U. Majumder, C. Capraro, and Y. Chen. Advanced techniques for robust sar atr: Mitigating noise and phase errors. In *IEEE International Radar Conference (RADAR)*, pages 844–849, 2020. 1

[12] N. Inkawhich, M. J. Inkawhich, E. K. Davis, U. K. Majumder, E. Tripp, C. Capraro, and Y. 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

[13] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *CoRR*, abs/1802.07856, 2018. 1

[14] Jerrick Liu, Oliver Nina, Radu Timofte, et al. NTIRE 2021 multi-modal aerial view object classification challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[15] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021. 1

[16] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 learning the super-resolution space challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

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

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

[19] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on image deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[20] Xiaoman Qi, Panpan Zhu, Wang Yuebin, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P. Takis Mathiopoulos. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, 11 2020. 1

[21] 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

[22] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 4, 6

[23] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020. 6

[24] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge J. Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983. IEEE Computer Society, 2018. 1

[25] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020. 5

[26] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019. 5