Efficient CNN Architecture for Multi-modal Aerial View Object Classification

Casian Miron
casian_miron@yahoo.com

Alexandru Pasarica
alexpasarica@gmail.com

Radu Timofte

MCC Resources S.R.L.
“Gheorghe Asachi” Technical University,
Iasi , Romania

Abstract

The NTIRE 2021 workshop features a Multi-modal Aerial View Object Classification Challenge. Its focus is on multi-sensor imagery classification in order to improve the performance of automatic target recognition (ATR) systems. In this paper we describe our entry in this challenge, a method focused on efficiency and low computational time, while maintaining a high level of accuracy. The method is a convolutional neural network with 11 convolutions, 1 max pooling layers and 3 residual blocks which has a total of 373.130 parameters. The method ranks 3rd in the Track 2 (SAR+EO) of the challenge.

1. Introduction

In air-to-ground applications, the visual distance is always long, in which the target seems to be rather small, and less feature information can be involved. In air-to-ground views, targets tend to have only a few tens of pixels of information, and convolutional neural networks (CNN) have much less information available in feature extraction than in conventional life scenarios [22]. At the same time, the air-to-ground target scene has a large field of view, and the detection target has many environmental changes (occlusion, background interference), so it is hard to acquire satisfied detection results [7].

The use of multi-sensor imagery, with both types of images, EO and SAR, has an advantage in the increase in performance of automatic target recognition (ATR) systems [4]. Electro optical (EO) sensors are used to obtain images that are directly interpretable and intuitive for human operators, alleviating the need for specialized training. Also, these types of images can be used in image processing and computer vision applications such that advanced intelligence can be generated computationally without significant human intervention. Whereas, synthetic aperture radar SAR images are considered more abstract and can be difficult to interpret by both human users and computer vision algorithms. SAR imagery is considered a non-literal imagery type because it does not look like an optical image which is generally intuitive to humans.

This paper is part of the NTIRE 2021 Workshop Challenge on Multi-Modal aerial view object classification [1]. The challenge has two tracks: Track 1 - classification based on the analysis of synthethic aperture radar (SAR) images and Track 2 - classification based on multisensor SAR and electro ocular (EO) images. The paper presents the implementation of an efficient CNN system architecture which was ranked 3rd in Track 2 of the challenge for SAR and EO images object classification.

2. State of the art

The classification of object based on aerial images has multiple applications such as traffic control [2, 15], automatic target recognition [8], agriculture [21, 5], weather monitoring or topological classification [13]. These applications are based on multiple types of images that vary from images acquired using drones, synthethic aperture radar, electro optical sensors, satellite images etc.

The analysis of SAR images for object classification has been previously used in other research papers such as Biondi et al. (2019) [3] which proposes a method for separation of buildings from vegetation based on deep CNN and polarimetric classification, Mdakane et al. (2020) [18] whcih proposes the use of Gradient Boosting Decision Tree Classifier (GBT) for identification of oil spils, Malmgren et al. (2017) [17] which proposes a method for automatic target recognition based on deep learning CNN, Wen et al. (2020) [20] which proposed a dual fast R-CNN for moving target detection.

One of the more challenging applications is the classification of SAR images, due to the high level of complexity and abstractization. Thus, the paper for the Multi-modal

Aerial View Object Classification Challenge written by Liu et al. [16] represents a collection of state of the art methods within this field. The paper presents the top 10 methods proposed for object classification from multisensor images. These methods vary in terms of the proposed CNN architecture from efficient low number of parameters to complex, accuracy oriented architectures with a very high number of parameters that also require longer model training time. There can also be observe a difference in data validation image processing time that is also influenced by the complexity of the trained model and the use of CPU vs GPU.

3. Proposed method

3.1. System architecture

The proposed system architecture for Track 1, presented in Table 1, consists in a convolutional neural network with 11 convolutions, 1 max pooling layers and 3 residual blocks [10]. The convolution layers use rectified linear unit (ReLU) [9, 19] activation function and are followed by batch normalization, its kernel, number of channels and parameters are 3×3, 64 and 36,928, respectively. The total number of parameters is 373,130. In order to compensate the uneven distribution of the dataset classes we used class weights that improve the class difference and do not allow the CNN to skew the results towards the class with the highest representation within the dataset. The diagram of the system architecture is presented in Figure 1. For the first competition track which is focused on classification of SAR data we train a single network with the input $55 \times 55 \times 1$.

The method proposed for Track 2 consists of a pair of SAR and EO images and for that we train 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 is given by averaging the Softmax vectors of the two networks. For SAR images we took the last 100 images from each class and made a validation set, for EO we took the last 100 images from each class for validation. Adam optimizer [23] is used for training the networks. We use the default hyper-parameter. The networks are trained for 100 epochs. No extra-data or augmentation was used. The analysis using the CNN architecture was done using a CPU (inference time per sample is approximately 0.02s). The system does not rely on network pretraining.

3.2. Datasets

The NTIRE 2021 workshop on Multi-modal Aerial View Object Classification Challenge proposed for classification two datasets: Track 1 dataset which requires the classification of SAR images and Track 2 dataset which is based on both SAR and EO images. Both datasets contain images from 10 classes presented in detail in Table 2. The distribution of images in each class is highly skewed due to the high number of images in the class ”Sedan” which has approximately 10 times more images than the next class ”SUV”. The proposed methods for each dataset use a training set created from selection of SAR and EO images.

Table 1. Proposed CNN system architecture

No.	Layer	K	#param	Dimensions
0	Input Image	-	-	HxWx1
1	C+R+D(0,3)	3x3	640	H-2xW-2x64
2	MaxPooling	2x2	-	(H-2)/2x(W-2)/2x64
3	C+R+D(0,3)	3x3	36928	(H-2)/2x(W-2)/2x64
4	BN	-	256	(H-2)/2x(W-2)/2x64
5	ADD(2+4)	-	-	(H-2)/2x(W-2)/2x64
6	C+R+D(0,3)	3x3	36928	(H-2)/2x(W-2)/2x64
7	BN	-	256	(H-2)/2 x (W-2)/2x64
8	C+R+D(0,3)	3x3	36928	(H-2)/2x(W-2)/2x64
9	BN	-	256	(H-2)/2x(W-2)/2x64
10	ADD(5+9)	-	-	(H-2)/2x(W-2)/2x64
11	C+R+D(0,3)	3x3	36928	(H-2)/2-2x(W-2)/2-2x64
12	BN	-	256	(H-2)/2-2 x (W-2)/2-2x64
13	C+R+D(0,3)	3x3	36928	(H-2)/2-4x(W-2)/2-4x64
14	BN	-	256	(H-2)/2-4 x (W-2)/2-4x64
15	C+R+D(0,3)	3x3	36928	(H-2)/2-4x(W-2)/2-4x64
16	BN	-	256	(H-2)/2-4 x (W-2)/2-4x64
17	C+R+D(0,3)	3x3	36928	(H-2)/2-4x(W-2)/2-4x64
18	BN	-	256	(H-2)/2-4 x (W-2)/2-4 64
19	ADD(14+18)	-	-	(H-2)/2-4x(W-2)/2-4x64
20	C+R+D(0,3)	3x3	36928	(H-2)/2-6x(W-2)/2-6x64
21	BN	-	256	(H-2)/2-6 x (W-2)/2-6x64
22	C+R+D(0,3)	3x3	36928	(H-2)/2-8x(W-2)/2-8x64
23	BN	-	256	(H-2)/2-8x(W-2)/2-8x64
24	C+R+D(0,3)	3x3	36928	(H-2)/2-10x(W-2)/2-10x64
25	BN	-	256	(H-2)/2-10x(W-2)/2-10x64
26	GlobalAveragePooling			64
27	D(0.3)	-	-	64
28	Dense(10)+SM		650	10
-	Output		373.130	

BN= BatchNormalization,C= Convolution,
D= Dropout,R= ReLU, SM = softmax

3.2.1 Track 1: SAR imagery

The first dataset is based on synthetic aperture radar (SAR) images for model training and image classification. The dataset contains 293772 SAR greyscale images with the spatial resolution of approximately 56x56 pixels. The presented approach for track 1 is based on all the SAR images from the dataset, divided into the training set and validation set. In order to obtain the validation set, we selected the last

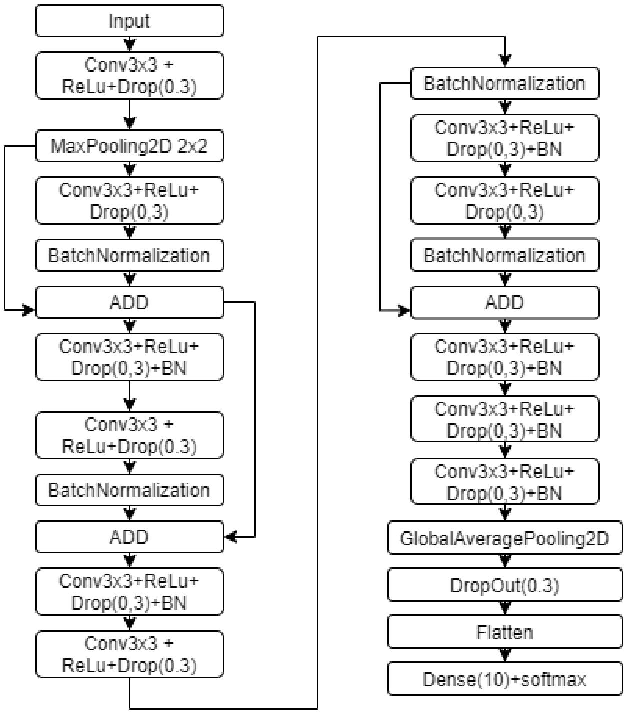

Figure 1. Proposed CNN system architecture

100 images from each class and the remaining images were used as the training set. Sample SAR images from the first dataset are presented in Figure 2.

3.2.2 Track 2: SAR+EO imagery

The second dataset proposed for the NTIRE 2021 competition consists of mulstisensor images acquired using electro optical (EO) and synthethic aperture radar (SAR) methods. The purposed of this approach is to compensate the abstract nature of SAR images that can be difficult to interpret to human users with more intuitive images obtained using EO sensors. The dataset contains the previously mentioned SAR images also used in Track1 and an additional 293772 EO greyscale images with the spatial resolution 32x32 pixels, with the same class distribution as the SAR images. The approach used when analysing images from this dataset is to train two models, one on SAR images and one on EO images and the output Softmax vectors from each model are average and applied on the validation set. The same approach as track 1 was applied in the separation into training and validation sets: the last 100 images from each class for both SAR and EO images were used for validation and the remaining images of each class were used in model training. Sample SAR and EO images are presented in Figures 2 and 3.

Table 2. Total number of images for each subclass

No.	Class	#SAR images	#EO images
1	Sedan	234,209	234,209
2	SUV	28,089	28,089
3	Pickup	15,301	15,301
4	Van	10,655	10,655
5	Box truck	1,741	1,741
6	Motorcycle	852	852
7	Flatbed truck	828	828
8	Bus	624	624
9	Pickup with trailer	840	840
10	Flatbed with trailer	633	633
-	Total	293722	293722

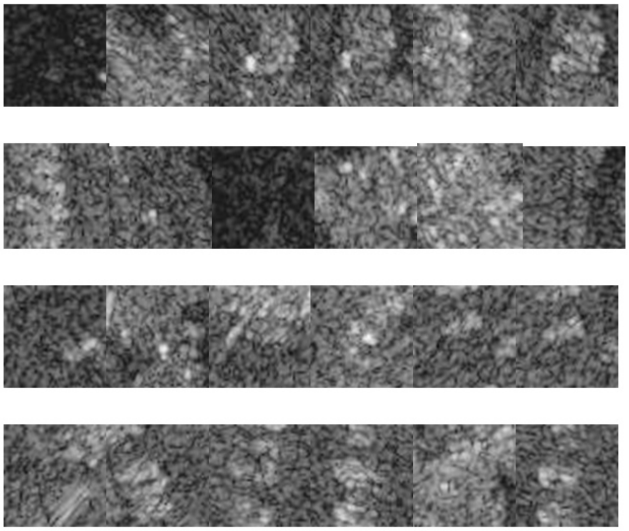

Figure 2. SAR images from the datasets

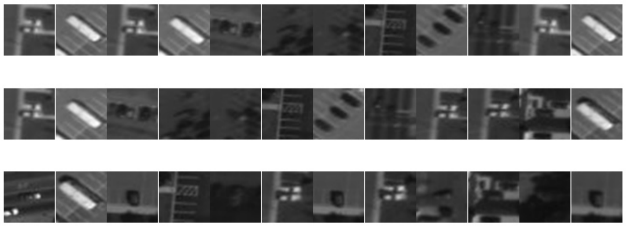

Figure 3. Electro optical images from the datasets

3.3. Class weigthed Softmax Cross-Entropy Loss

The Class-Balanced Loss can be used in training highly imbalanced datasets by introducing a weighting factor that is inversely proportional to the effective number of samples. The datasets from both challenge tracks, detailed in Table 2, present a main class that has approximately 10 times higher number of images than the next class, also there are another 6 classes which have under 1800 images.

The Softmax is represented by the feature vector that is produced in the final layer of the proposed CNN architecture. [14]. The Softmax activation function determines the probabilities for each class where the sum of all probabilities is equal to 1. This activation of the CNN model has the advantage of normalizing the outputs and each value in the output of the Softmax function is interpreted as the probability of membership for each class [12]. The Softmax is determined by the following equation:

$$Softmax = (\frac{(exp(z_i)}{\sum_{j=1}^{C}(exp(z_j))}) \tag{1}$$

The Softmax function takes into consideration each class z = [z_1, z_2, z_3, ..., z_C], where C=10 represents the total number of clases in the analysed datasets, as mutually exclusive and computes the probability distribution over all classes [6]. Given a sample with class label z_i, the Softmax cross-entropy ($CE_{Softmax}$) loss for this sample is written as:

$$CE_{Softmax} = -log(\frac{(exp(z_i)}{\sum_{j=1}^{10}(exp(z_j))}) \tag{2}$$

Suppose class i has n_i training samples, the class-balanced ($CB_{Softmax}$) Softmax cross-entropy loss is [11]:

$$CB_{Softmax} = -\frac{NoI}{C * n_i} log(\frac{exp(z_i)}{\sum_{j=1}^{10}(exp(z_j))}) \tag{3}$$

where $\frac{NoI}{C*n_i}$ is the class weight factor and the NoI represents the total number of images which is equal to 293722.

4. Experimental results

The first track results were obtained using a single model with 27 layers (11 convolutional layers) on the training set of SAR images. Given the complex type of images that are analyzed, the performance of the training stage is reduced. This is also reflected in the overall standing where out proposed method is ranked 13th based on accuracy, as presented in Table 3.

Table 3. Results obtained for track 1 test data. (*) Our runtime is reported on CPU, whereas the other runtimes are self-reported on GPU.

Method	Accuracy [%]	Runtime [s]
1 duanyuru	34.615	0.43
2 meye66	26.634	0.02
3 ulosc	26.392	n/a
4 yangchris11	26.029	0.001
5 ga_z_a	25.061	0.04
6 XuYifei	24.818	n/a
7 BONG	24.576	0.04
8 zhangxs	23.608	0.006
9 oooo0	23.366	0.0048
13 Casian	20.944	0.02*

The second Track proposed solution was implemented using Python scripting language and deployed on a 13 Gb GPU for the training stage of the challenge, whereas the test stages required an approximate computational time of 0.02s per sample on a single CPU core and on the GPU is 0.5ms per image this results in a inference time of 1ms for both images EO+SAR. The results for Track 2 were obtained using two networks for each type of images, either SAR or EO. One network is trained on the SAR images dataset with a input resolution of 56x56x1 and the other one is trained on the EO images dataset with input 32x32x1. The final result is determined by computing the average of the output Softmax vectors of the two networks. The overall results for Track 2 are presented in Table 4 where we can observe that out proposed method is ranked 3rd based on accuracy. For the CNN architecture used in Track 2 we also performed an ablation study in order to determine if the performance of the trained model is influenced by the number of layers and parameters used. This can be an indicator of the ratio between CNN object classification efficiency and accuracy.

Table 4. Results obtained for track 2 test data. (*) Our runtime is reported on CPU, whereas the other runtimes are self-reported on GPU.

Method	Accuracy [%]	Runtime [s]
1 shirly	46.852	0.035
2 caihuanqia	34.625	0.018
3 Casian	26.513	0.02*
4 MichaelXin	26.029	0.15
5 LeonShangguan	25.061	0.0581
6 xsourse	23.971	0.018
7 yangchris11	21.065	0.001
8 benjamin666	20.702	0.033
9 vamshi	17.01	n/a

The datasets are heavily skewed in favor of images of sedan type cars, which can make it difficult to obtain an accurate training stage. This is why a weighted class method was used in order to prevent over training of the CNN towards this type of class. The total number of images for each class is presented in Table 2.

4.1. Ablation study

The ablation study experiment was performed in order to see that if we modify the architecture by stripping in different stages layers we obtain a faster CNN and maintain the accuracy of the assembler used on the Track2 EO+SAR challenge. The five experimental setups are obtained as follows:

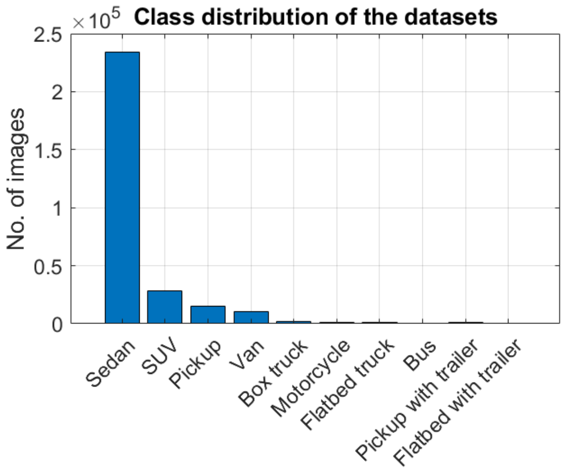

Figure 4. Distribution of images from the datatsets into classes

- Setup 0: We remove from the architecture presented in Table 1 the following layers: 20,21,22,23,24,25 and use 32 as the number of channels for the convolutions and the dropout 0.1 resulting in a lightweight architecture with 47,530 parameters;
- Setup 1: We use the same architecture as presented in Table 1, the only difference is that the number of channels is 32 instead of 64, resulting in the total number of parameters of 94.410 ;
- Setup 2: modify Setup 1 by removing the skip connection layers: 5,10,19;
- Setup 3: We modify the architecture presented in Table 1 by removing the skip connection layers: 5,10,19
- Setup 4: submitted CNN architecture to the challenge

The experimental results obtained for the setups of the ablation study are presented in Table 5:

Table 5. Ablation study of the impact of number of parameters over model accuracy

Setup	**#param**	**Acc[%]**	**speed**
Setup 0	2x47.530	20	2x12ms
Setup 1	2x94.410	21.42	2x18ms
Setup 2	2x94.410	24.935	2x17ms
Setup 3	2x373,130	20.389	2x19ms
Setup 4	2x373,130	22.078	2x20ms

5. Conclusions

The proposed method for Track 2 of the NTIRE Multi-modal Aerial View Object Classification Challenge relies on an efficient CNN architecture. This aspect is supported by the reduced number of parameters used in the system architecture. Overall, for the 27 layers of the CNN architecture a total of 373,130 parameters were used. The results obtained using the proposed method are below two other methods submitted for the NTIRE 2021 Challenge, but its novelty comes from the reduced processing time and low number of parameters used during the training stages. Also, it is worth mentioning that the method proposed in this paper is the only one that uses CPU processing during the test stages, whereas all others rely on GPU processing, which is generally necessary when working with complex CNNs.

On the other hand, our proposed method for Track 1 of the challenge which relies on the use of a single model trained on SAR images, has lower accuracy and ranks 13th in the competition due to the fact that SAR images are more complex and require the use of a CNN architecture focused on accuracy that uses a high number of parameters.

References

[1] NTIRE 2021. Ntire 2021 workshop. https://competitions.codalab.org/competitions/28095#learn_the_details. Accessed: 12.02.2021. 1

[2] Bilel Benjdira, Taha Khursheed, Anis Koubaa, Adel Ammar, and Kais Ouni. Car detection using unmanned aerial vehicles: Comparison between faster r-cnn and yolov3. In *2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS)*, pages 1–6. IEEE, 2019. 1

[3] Filippo Biondi. Multi-chromatic analysis polarimetric interferometric synthetic aperture radar (mca-polinsar) for urban classification. *International Journal of Remote Sensing*, 40(10):3721–3750, 2019. 1

[4] Ilker Bozcan and Erdal Kayacan. Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8504–8510. IEEE, 2020. 1

[5] Pei-Chun Chen, Yen-Cheng Chiang, and Pei-Yi Weng. Imaging using unmanned aerial vehicles for agriculture land use classification. *Agriculture*, 10(9):416, 2020. 1

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. 4

[7] Recai Alper Emek and Nusret Demir. Building detection from sar images using unet deep learning method. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 44:215–218, 2020. 1

[8] Antonio-Javier Gallego, Antonio Pertusa, and Pablo Gil. Automatic ship classification from optical aerial images with convolutional neural networks. *Remote Sensing*, 10(4):511, 2018. 1

[9] Kazuyuki Hara, Daisuke Saito, and Hayaru Shouno. Analysis of function of rectified linear unit used in deep learning. In *2015 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2015. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[11] Quentin Jodelet, Xin Liu, and Tsuyoshi Murata. Balanced softmax cross-entropy for incremental learning. *arXiv preprint arXiv:2103.12532*, 2021. 4

[12] Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou, et al. Efficient softmax approximation for gpus. In *International Conference on Machine Learning*, pages 1302–1310. PMLR, 2017. 4

[13] Ren N Keyport, Thomas Oommen, Tapas R Martha, KS Sajinkumar, and John S Gierke. A comparative analysis of pixel-and object-based detection of landslides from very high-resolution images. *International journal of applied earth observation and geoinformation*, 64:1–11, 2018. 1

[14] Takumi Kobayashi. Large margin in softmax cross-entropy loss. In *BMVC*, page 139, 2019. 4

[15] Chih-Yi Li and Huei-Yung Lin. Vehicle detection and classification in aerial images using convolutional neural networks. In *VISIGRAPP (5: VISAPP)*, pages 775–782, 2020. 1

[16] 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

[17] David Malmgren-Hansen, Anders Kusk, Jørgen Dall, Allan Aasbjerg Nielsen, Rasmus Engholm, and Henning Skriver. Improving sar automatic target recognition models with transfer learning from simulated data. *IEEE Geoscience and remote sensing Letters*, 14(9):1484–1488, 2017. 1

[18] Lizwe Wandile Mdakane and Waldo Kleynhans. Feature selection and classification of oil spill from vessels using sentinel-1 wide-swath synthetic aperture radar data. *IEEE Geoscience and Remote Sensing Letters*, 2020. 1

[19] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 2

[20] Liwu Wen, Jinshan Ding, and Otmar Loffeld. Video sar moving target detection using dual faster r-cnn. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2984–2994, 2021. 1

[21] Beibei Xu, Wensheng Wang, Greg Falzon, Paul Kwan, Leifeng Guo, Zhiguo Sun, and Chunlei Li. Livestock classification and counting in quadcopter aerial images using mask r-cnn. *International Journal of Remote Sensing*, 41(21):8121–8142, 2020. 1

[22] Dongfang Yang, Xing Liu, Hao He, and Yongfei Li. Air-to-ground multimodal object detection algorithm based on feature association learning. *International Journal of Advanced Robotic Systems*, 16(3):1729881419842995, 2019. 1

[23] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–2. IEEE, 2018. 2