Cross Modality Knowledge Distillation for Multi-modal Aerial View Object Classification

Lehan Yang
The University of Sydney
Camperdown NSW 2006 Australia
lyan3310@uni.sydney.edu.au

Kele Xu*
Key Laboratory for Parallel and
Distributed Processing
Changsha, China
kelele.xu@gmail.com

Abstract

In the case of bad weather or low lighting conditions, a single sensor may not be able to capture enough information for object identification. Compared with the traditional optical image, synthetic aperture radar (SAR) imaging has greater advantages, such as the ability to penetrate through fog and smoke. However, SAR images are of low resolution and contaminated by high-level speckle noise. As a result, it is of great difficulty to extract powerful and robust features from the SAR images. In this paper, we explored whether multiple imaging modalities can improve the object detection performance. Here, we propose a Cross Modality Knowledge Distillation (CMKD) paradigm, and explore two different network structures named CMKD-s and CMKD-m for the object classification task. Specifically, CMKD-s transfers the information captured by the two sensors using the online knowledge distillation, which can achieve cross-modal knowledge sharing and enhance the robustness of the aerial view object classification model. Moreover, leveraging the semi-supervised enhanced training, we proposed a novel method named CMKD-m, which strengthens the model for mutual knowledge transfer. Through quantitative comparison, we found that CMKD-s and CMKD-m outperform the method without knowledge transfer, on the NTIRE2021 SAR-EO challenge dataset.

1. Introduction

Computer vision technology has developed rapidly in recent years, and many related tasks have been well explored [7, 16, 6]. Images acquired using Electro-optical(EO) sensors have been used in many object detection tasks. However, the images obtained by other sensors, such as synthetic aperture radar (SAR), can also assist the EO image analysis when enough information cannot be obtained from the EO image. Currently, the most impressive ways to leverage the knowledge between two types of data is to merge the knowledge of the two domain by transfer learning [12], or transfer learning from simulated data [10]. However, due to the fact that the aerial view sensor has the characteristics of small target resolution and relatively high-level noise, the learning ability of a single model may be severely restricted and not as comprehensive.

Fortunately, we can capture objects at the same time by using EO and SAR sensors to obtain corresponding multi-modal data. Thereby there will be opportunities to obtain more discriminative features from multi-modal data. Here, we propose a method of learning and classifying multi-modal aerial view image data by transferring knowledge. The output distribution is used to transfer knowledge between multi-modal data, so that the student model can obtain more robust cross-modal prior knowledge.

Inspired by the knowledge distillation [5], which uses the teacher model for knowledge transferring purpose, thus enhancing the student model, we designed a dual model for the multi-modal object classification task. The model consists of an SAR image-based teacher model and an EO image-based student model. Specifically, we utilize the output of the teacher model which inputs the corresponding different modal images as a student model, to minimize the distance of teacher model output distribution and student model output distribution. Hence, the EO student model can transfer knowledge from the SAR teacher model, which enhances the robustness of the student model. Based on ResNet [4] backbone, the knowledge distillation works well between these two different modal images. At the same time, inspired by unsupervised neural machine translation[8], we propose a semi-supervised iteration training method. By continuously exchanging the roles of teachers and students in each iteration and between iterations,

*Corresponding author.

teachers and students can learn from each other and improve the performance of the model in the process of mutual enhancement. This method also enables two-way knowledge exchange between SAR image and EO image, instead of transferring SAR knowledge to EO knowledge in the previous method.

In summary, our contribution has the following two major points: 1) We propose a knowledge transfer method for multi-modal aerial view image data. One modal data model learns another modal data, so that the student model can learn more prior knowledge and has better robustness and performance. 2) We propose a semi-supervised training method for multi-modal aerial view image classification, which allows cross-modal knowledge communication in the role rotation of teachers and students, and enhances the performance of teacher and student models.

2. Related work

SAR and EO Classification. There are many works addressed on classification of SAR images. Tzeng *et al*. [13] proposed a method to classify SAR images using fuzzy logic and dynamic neural networks, this is an early attempt to use neural networks to classify SAR images in the early years. Zhao *et al*. [18] addressed on design a end-to-end convolutional neural network to classify SAR images at patch level, and obtained a high accuracy. A single-polarization/supervised SAR image classification system is proposed by Geng *et al*. [3] to improve the problem of noise and speckles in the data, which are not easy to be effectively characterized. For the classification of EO images, traditional visual method[1] is used to extract texture features to classify ships on the sea. Katherine [11] used a VGG-16 based convolutional neural network to detect and classify EO image of ships on the sea. There are also works that use both SAR and EO sensors. Mohammad *et al*. [12] proposed an algorithm to transfer the knowledge of EO domain to SAR domain through transfer learning, which reduces the dependence of SAR images on accurate and large numbers of labeled points.

Knowledge Distillation. The concept of knowledge distillation was first proposed by Hinton *et al*. [5], which is intended to allow models with small parameters and weak learning capabilities to have the accuracy of large models. There have been many improvements on the basic methods of knowledge distillation. Sergey *et al*. [15] passes the attention of the teacher network to the student network, which uses spatial-attention to decode the contribution of the input image space to the output. Junho *et al*. [14] innovated at the positional level of knowledge transfer. Knowledge transfer is carried out through the feature flow relationship between layers in the network, allowing students to learn how to learn rather than how to get output.

3. Method

In this section, we individually outline the architecture and training details of the two models: CMKD-s and CMKD-m.

3.1. Architecture

3.1.1 CMKD-s

Here, we propose a method named CMKD-s, to solve SAR and EO object classification using cross-modal knowledge distillation. In this section, we elaborate the key component: cross modality knowledge transfer.

Cross Modality Knowledge Transfer. We propose a dual network, the model that recognizes SAR images is used as the teacher model, and the model that recognizes EO images is used as the student model, as shown in Figure 1. The teacher model is pre-trained on the EO image, and then trained on the SAR image, and the teacher model is obtained that is not too strong but has the ability to provide favorable prior knowledge. In the student model for training EO images, we input matched SAR and EO images to the teacher and student models at the same time, the teacher model can be used for the inference, and the student model is training. The output distribution of the teacher model is transferred to the student model through KL divergence loss as knowledge distillation loss. It can help the student model learn some knowledge not included in the EO image, so as to achieve the purpose of enhancing the student model. The network architecture is the same, based on the same recognition model, but processing data of different modalities. In order to obtain better recognition performance, we use ResNet34 as the teacher and student model, which is not pre-trained on ImageNet. We use different data augmentation methods, including MixUp[17] and AutoAugment[2] for data input augmentation, they can make the distribution of input data more continuous and broader.

3.1.2 CMKD-m

CMKD-m is an improved training method based on CMKD-s. Inspired by the unsupervised neural machine translation, CMKD-m rotates the positions of students and teachers model during the cycle training process, and achieves mutual enhancement in the enhancement iterations of both parties. In this section, we elaborate the two key components: cross modality knowledge transfer and semi-supervised enhanced training.

Cross Modality Knowledge Transfer. The knowledge distillation method of CMKD-m is exactly the same as that of CMKD-s, which achieves the purpose of knowledge transfer by narrowing the distance between the output distribution of the teacher model and the student model.

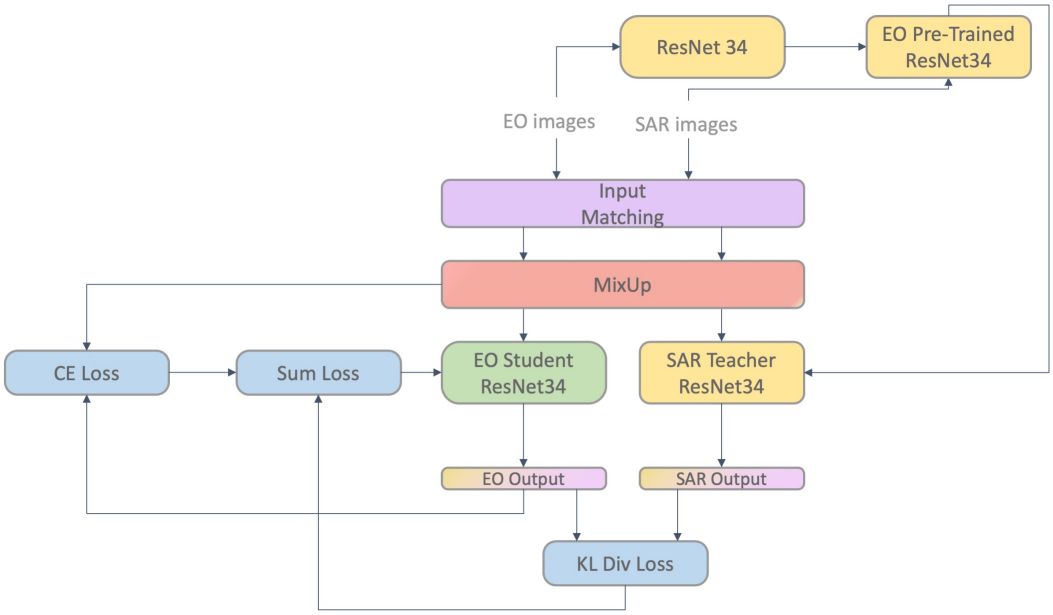

Figure 1. The architecture of the proposed CMKD-s.

Semi-Supervised Enhanced Training. We propose a training method for SAR and EO image classification models. As shown in Figure 2, the entire training process is divided into n iterations. In each round and between each round, the two models are calling for the roles of teachers and students to enhance each other. In the first iteration, as a teacher model, the EO model first uses the pre-trained model on the SAR image to routinely train on the EO image. Then in the training process of the student SAR model, the knowledge of the teacher model is transferred to the student model. In the subsequent iterations, first the EO model and the SAR model exchange the roles of teacher and student, and EO serves as a student model to obtain knowledge from the teacher SAR model that from the previous iteration. The roles are exchanged again, and the SAR model is used as a student model to gain knowledge from the EO teacher model of this iteration. Repeat this training processing until the iteration ends. Since EO images are easier to obtain features by the model than SAR images, we assume that the EO model has relatively more knowledge. Therefore, when the EO model is used as a teacher model, the knowledge distillation coefficient α will be relatively high, and when the SAR model is used as a teacher model, α will be relatively low, where α is the weight of the knowledge distillation loss when calculate the sum of loss.

3.2. Loss function

We utilize two losses in the main part of training, they are weighted cross entropy loss L_{wce} and weighted KL divergence loss. L_{wKLdiv}, as Eq. 1.

$$L_{total} = \alpha L_{wKLdiv} + (1 - \alpha) Lwce \tag{1}$$

Weighted Cross Entropy Loss Likely to many single-label classification tasks, we chose cross-entropy loss. However, considering the extremely unbalanced distribution of the training dataset, we introduced weight to weight the loss under each category, which alleviated the model offset caused by the imbalance of the data.

$$L_{wce} = weight[class] \cdot L_{ce}(M(I), L) \tag{2}$$

where weight stands for the weight of every class, calculate by the number of samples. Class is the index of category. L_{ce} donates the original cross entropy loss, where $M(I)$ is the output of model M when input the input I.

Weighted KL Divergence Loss. Similarly, we also use weights on the KL divergence loss to balance the impact of the data set. The loss is calculated between the output of the teacher model and the student model, which is used to narrow the output distribution of students and teachers to

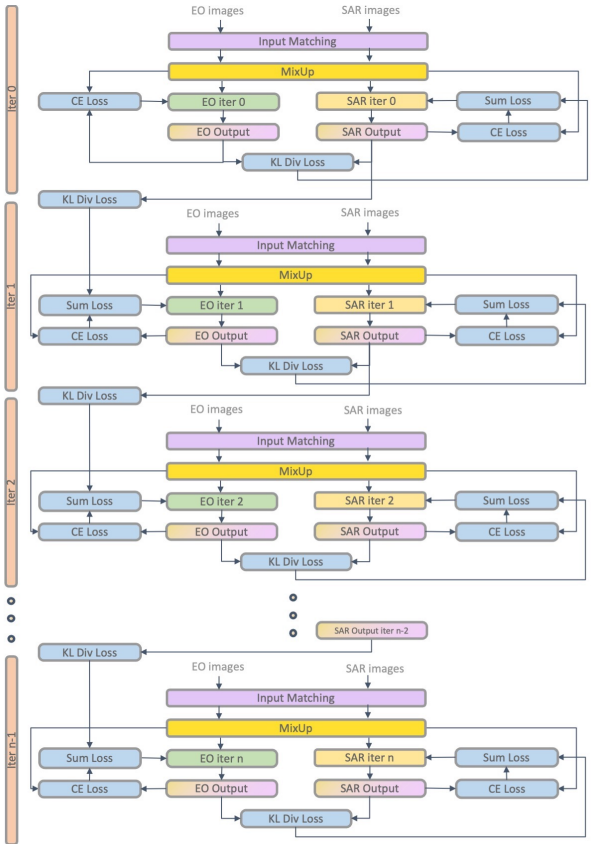

Figure 2. The architecture of the proposed CMKD-m.

achieve the purpose of knowledge transfer.

$$L_{wKLdiv} = weight[class] \cdot D_{KL}(T(I_t), S(I_s)) \quad (3)$$

where weight[class] donates the weight of each class, $T(I_t$ is the output of teacher model T when input teacher input, $S(I_s)$ is the output of student model S when input student input, D_{KL} stands for the original Kullback–Leibler divergence loss. The input of student and teacher model can vary. Specifically, when transferring EO knowledge to SAR model, EO is the input of teacher model while SAR is the input of student model. When we transfer SAR knowledge to EO model, SAR will be the input of teacher model while EO will be the input of student model.

4. Experiments

In this section, we first introduce the details of the dataset and training method, and then quantitatively evaluate the performance of our method on the dataset. Finally, an ablation study is conducted to demonstrate the effectiveness of each component of the network. In the track 2 of NRTIRE-21 Multi-modal Aerial View Object Classification Challenge [9], our Top-1 Accuracy ranked 6th place on the final leaderboard.

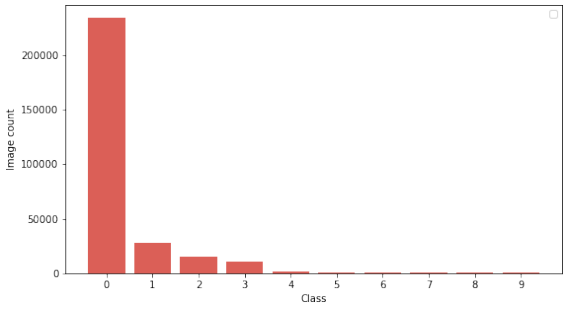

Figure 3. The distribution of NTIRE21 Aerial View Object Dataset.

4.1. Dataset

NTIRE21 Aerial View Object Dataset. NTIRE21 Aerial View Object Dataset is a large data set corresponding to EO and SAR images. All images are single object small resolution images cropped from the pictures taken by EO and SAR sensors. The train set includes 293772 images, it can be seen from the Figure 3 that the distribution between the categories of the dataset is very unbalanced, and the images of the first category "sedan" are much more than other categories. As shown in Figure 4, images are randomly selected from each category. Except the image is set to be completely black because the analog sensor cannot capture it, it can be seen that the noise of the SAR image is relatively high. Compared to SAR images, the robust feature learning can be easier for the EO image.

4.2. Experimental Setting

Training Details. For the two methods we proposed, we used the same method to divide the dataset, randomly selecting 70% of each category as the training set and 30% as the validation set. In the stage of data pre-processing, we apply 30×30 as the resolution of EO images, 48×48 to SAR images. And use AutoAugment to automatically enhance and augment the data, and use MixUp to make the distribution of training data more continuous. Besides, we adopt SGD optimizer with 0.9 momentum and 0.01 learning rate, which reduces to half every 20 epochs. Our model was implemented with PyTorch with 1 Tesla M40 GPU.

Measure Metric. Similar to the evaluation indicators required by the NTIRE 2021 Multi-modal Aerial View Object Classification Challenge, we also selected Top-1 accuracy as the metric for the model evaluation, which is usually used for image classification tasks.

Post Processing. Although we have made efforts to solve the problem of uneven distribution of the dataset, the actual output of the final model is still offset. On the test set of the NTIRE 2021 Multi-modal Aerial View Object Classification Challenge, the output distribution without post-

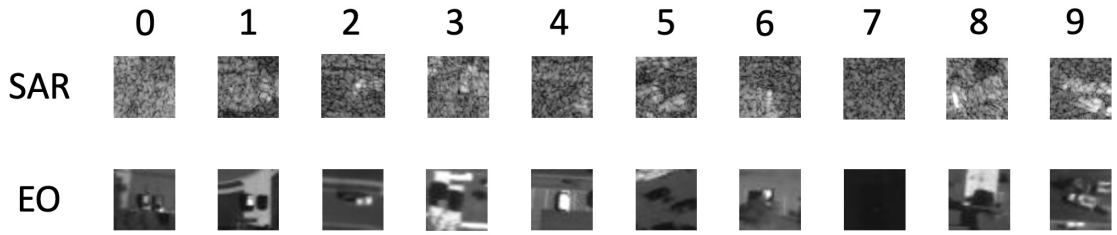

Figure 4. Samples from the NTIRE21 Aerial View Object Dataset.

Methods	Top-1 Accuracy
Baseline	0.22337662
CMKD-s	**0.31168831**
CMKD-m	0.23376623
CMKD-s(post processed)	**0.36363636**

Table 1. Results CMKD-s and CMKD-m

processing will be biased towards the first category. Since the ground truth distribution of the test set is known to be uniform, we performed prediction processing on the first four classes with the most data distribution in post processing. The method is to sort the activation values in the four categories, and one-tenth of the number of test sets is used as the output number n of each category. The top n images with the activation value of these categories are output as this category, and the remaining images are output to the category with the second highest activation value of the image.

4.3. Results

The results of both our methods and comparison with our baseline on the testset of NTIRE21 Aerial View Object Dataset shown in Table 1. It can be seen that the two methods we proposed, CMKD-s and CMKD-m, have improved compared with the baseline that does not use cross-modal knowledge distillation, especially CMKD-s that uses post processing has the greatest improvement. For the small increase of CMKD-m, we think it is caused by the poor processing of SAR image noise, which makes the EO image model receive noise disturbance in the process of learning the SAR model. In the process of iteration training, both the SAR model and the EO model were subjected to greater bias.

4.4. Ablation Study

As mentioned earlier, we used MixUp and AutoAugment, two powerful data augmentation tools, in the training strategy of the model. These two tools have improved the performance of our model. However, in order to prove that the cross modality knowledge distillation method we proposed does positively improve the performance of the model, we conducted ablation experiments on the three components of our method: MixUp, AutoAugment, and Cross Modality Knowledge Distillation. The ablation experiments as following: 1) baseline: use only basic image changes for preprocessing: random rotation, random resized crop, random horizontal flip, random vertical flip, and only use ResNet34 as the network backbone; 2) baseline + MixUp: Alpha = 0.4 is used as the mix parameter of MixUp; 3) baseline + AutoAugment;: the implementation of AutoAugment is added to the base to achieve the purpose of enhancing data; 4) baseline + MixUp + AutoAugment: MixUp and AutoAugment are simultaneously implemented on the base. On this basis, we compared the methods using CMKD-s components: 5) baseline + CMKD-s; 6) baseline + MixUp + CMKD-s; 7) baseline + AutoAugment + CMKD-s; 8) baseline + MixUp + AutoAugment + CMKD-s. In detail, we trained these networks for 120 epochs on the NTIRE EO trainset, and use Top-1 Accuracy as a performance indicator on the NTIRE EO testset.

As shown in Table 2, every element in our network has an important impact on the performance of the network. MixUp and AutoAugment alone or together with the network will have a positive improvement in the accuracy of the network. When CMKD-s acts on the network, the top-1 accuracy is significantly improved. Especially when CMKD-s works with MixUp and AutoAugment, the performance of the network is improved more significantly. Hence, it can be proved that the CMKD-s we proposed is efficient. We also compared the accuracy after post-processing. It can be seen that post-processing is easier to improve performance when the accuracy of the model is already high. And it may cause performance degradation when the accuracy of the original model is insufficient.

5. Conclusion

In this paper, we propose an aerial view object classification method based on knowledge distillation to obtain cross-modal knowledge. We use the output of parallel data

Methods	CMKD-s	Top-1 Accuracy	Top-1 Accuracy (Post-processing)
Baseline	-	0.13506494	0.18571429
Baseline + MixUp	-	0.14558440	0.14415584
Baseline + AutoAugment	-	0.19090909	0.25714286
Baseline + MixUp + AutoAugment	-	0.22337662	0.20649351
Baseline	√	0.14025974	0.12467532
Baseline + MixUp	√	0.17142857	0.21298701
Baseline + AutoAugment	√	0.20259740	0.24935065
Baseline + MixUp + AutoAugment (**Ours**)	√	**0.31168831**	**0.36363636**

Table 2. Results of CMKD-s and CMKD-m.

in different modal models to allow student models to learn robust prior knowledge. In addition, we introduced a semi-supervised iteration training method, which enables cross-modal models to enhance each other in iterating training. Compared with the baseline that does not use knowledge distillation, we have achieved a huge improvement, which proves the effectiveness of the method.

References

[1] V. F. Arguedas. Texture-based vessel classifier for electro-optical satellite imagery. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3866–3870, 2015. 2

[2] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[3] J. Geng, H. Wang, J. Fan, and X. Ma. Deep supervised and contractive neural network for sar image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4):2442–2459, 2017. 2

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2

[6] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 1

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[8] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 1

[9] 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. 4

[10] 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. *I E E E Geoscience and Remote Sensing Letters*, 14(9):1484–8, 2017. 1

[11] Katherine Rice. Convolutional neural networks for detection and classification of maritime vessels in electro-optical satellite imagery, 2018. 2

[12] Mohammad Rostami, Soheil Kolouri, Eric Eaton, and Kyungnam Kim. Sar image classification using few-shot cross-domain transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1, 2

[13] Yu-Chang Tzeng and Kun-Shan Chen. A fuzzy neural network to sar image classification. *IEEE Transactions on Geoscience and remote Sensing*, 36(1):301–307, 1998. 2

[14] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[15] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2

[16] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 1

[17] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2

[18] J. Zhao, W. Guo, S. Cui, Z. Zhang, and W. Yu. Convolutional neural network for sar image classification at patch level. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 945–948, 2016. 2