UNSUPERVISED DOMAIN ADAPTATION USING A TEACHER-STUDENT NETWORK FOR CROSS-CITY CLASSIFICATION OF SENTINEL-2 IMAGES

Jingliang Hu[1], Lichao Mou[2], Xiao Xiang Zhu[1,2*]

[1] Signal processing in Earth observation (SiPEO), Technische Universität München (TUM),
80333 Munich, Germany - Jingliang.Hu@tum.de
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR),
82234 Weßling, Germany - (lichao.mou, xiaoxiang.zhu)@dlr.de

KEY WORDS: Cross-city classification, Deep learning, Mean teacher model, Teacher-student network, Transfer learning, Unsupervised domain adaptation

ABSTRACT:

A machine learning algorithm in remote sensing often fails in the inference of a data set which has a different geographic location than the training data. This is because data of different locations have different underlying distributions caused by complicated reasons, such as the climate and the culture. For a large scale or a global scale task, this issue becomes relevant since it is extremely expensive to collect training data over all regions of interest. Unsupervised domain adaptation is a potential solution for this issue. Its goal is to train an algorithm in a source domain and generalize it to a target domain without using any label from the target domain. Those domains can be associated to geographic locations in remote sensing. In this paper, we attempt to adapt the unsupervised domain adaptation strategy by using a teacher-student network, mean teacher model, to investigate a cross-city classification problem in remote sensing. The mean teacher model consists of two identical networks, a teacher network and a student network. The objective function is a combination of a classification loss and a consistent loss. The classification loss works within the source domain (a city) and aims at accomplishing the goal of classification. The consistent loss works within the target domain (another city) and aims at transferring the knowledge learned from the source to the target. In this paper, two cross-city scenarios are set up. First, we train the model with the data of the city Munich, Germany, and test it on the data of the city Moscow, Russia. The second one is carried out by switching the training and testing data. For comparison, the baseline algorithm is a ResNet-18 which is also chosen as the backbone for the teacher and student networks in the mean teacher model. With 10 independent runs, in the first scenario, the mean teacher model has a mean overall accuracy of 53.38% which is slightly higher than the mean overall accuracy of the baseline, 52.21%. However, in the second scenario, the mean teacher model has a mean overall accuracy of 62.71% which is 5% higher than the mean overall accuracy of the baseline, 57.76%. This work demonstrates that it is worthy to explore the potential of the mean teacher model to solve the domain adaptation issues in remote sensing.

1. INTRODUCTION

According to the United Nations (UN)[1], more than 55.3% of the world's population lived in urban areas in 2018, and the number is still growing. Mapping the urban regions globally provides strategic geographic information for the development of the human kind. Current state-of-the-art global urban mapping delivers a global layer of binary mask, urban vs. non-urban, such as the World Settlement Footprint (WSF) (Marconcini et al., 2019). However, binary maps are not able to provide information within cities, such as functionality and morphological structure of blocks. Those information are very relevant. For example, the evaluation of the Sustainable Development Goals (SDGs) relies on those geographic information within cities (Paganini et al., 2018; Melchiorri et al., 2019). Currently, some efforts have been done toward providing detailed urban maps on the global scale (Demuzere et al., 2019; Yoo et al., 2019). All those studies have pointed out a technical issue for achieving their goals, the cross-city classification challenge. For a global task, a classification algorithm is trained over data sets of a limited number of cities, and is applied over all cities globally. During the inference, the accuracy of the trained algorithm is often not acceptable. This is because the data of different cities change due to different climates, environments, cultures, and so on. This issue is so relevant in remote sensing because no one can avoid it when a large scale or a global scale task is under consideration. To tackle this issue, domain adaptation is an option from the methodological aspect.

Domain adaptation in the context of this paper refers training in a source domain and testing in a target domain for the same task, according to the description in (Pan, Yang). For a global scale remote sensing tasks, the target domain normally has no labeled data samples or occasionally a few labeled ones. This work focuses on the former case, a.k.a. unsupervised domain adaptation. Among literature, there are some studies (Demuzere et al., 2019; Yoo et al., 2019) that test the transferability of various algorithms in remote sensing tasks. However, to our best knowledge, only few studies have developed strategies to improve the transferring capability of their algorithms. Tong et al. has developed a strategy to improve the transferring capability of their algorithm. This work trains a deep network in the source domain, predicts labels of instances from the target domain with the trained network, selects reliable predictions in the target domain based on defined criterion, and tunes the trained network with the selected reliable samples. Their experiments have shown considerable improvements. However, the selection of reliable predictions in this framework requires human interaction and empirical experiences. It might be an issue in practice when dealing with big data. Therefore, it would be more practical to have an end-to-end learnable solution. Fang

*Corresponding author
[1] United Nations, The World's Cities in 2018

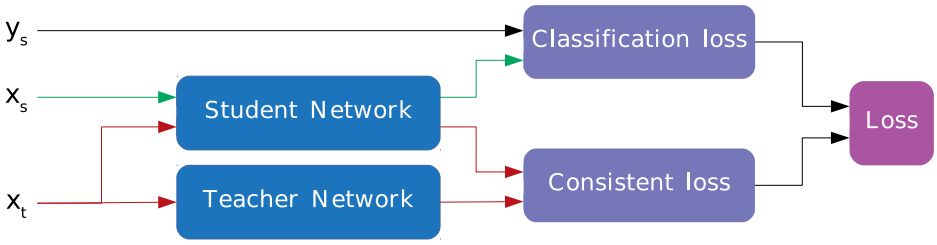

Figure 1. The structure of the mean teacher model implemented in this work, modified from (French et al., 2017).

et al. and Liu et al. have both applied a generative adversarial network strategy to deal with domain adaptation for land cover mapping using very high resolution (VHR) optical aerial images. However, it is very expensive to access VHR optical aerial images with a consistent quality or a global coverage.

Pursuing an end-to-end network solution to the domain adaptation problem with no labeled data available in the target domain, a model draw the authors' attentions, the mean teacher model (Tarvainen, Valpola; French et al., 2017). The mean teacher model was originally (Tarvainen, Valpola) designed as a temporal ensemble solution for semi-supervised learning, and later it was modified in (French et al., 2017) to deal with domain adaptation problems. The modified version produced the state-of-the-art classification accuracy over multiple benchmark data sets of unsupervised domain adaptation. In this paper, the authors attempt to investigate the performance of the mean teacher model in terms of the domain adaptation problem in remote sensing. In the section 2, the cross-city problem, the mean teacher model, and the data used in this paper are introduced. Section 3 illustrates the experiment results upon which a discussion is carried out. Section 4 concludes this paper.

2. METHOD

2.1 Problem statement

In this paper, the cross-city classification challenge is formatted as a domain adaptation problem. The data of one city with annotations are treated as the source domain. The data of the source domain is represented as $(\mathbf{X}_s, \mathbf{Y}_s)$, where $\mathbf{X}_s$ presents the data and $\mathbf{Y}_s$ indicates the corresponding label. The data of another city without any annotation is treated as the target domain and represented as $\mathbf{X}_t$. The task is to estimate the label $\hat{\mathbf{Y}}_t$ of data $\mathbf{X}_t$ in the target domain.

2.2 Mean teacher model

The structure of the mean teacher model is illustrated in Figure 1. It consists of a student network and a teacher network.

The student network takes the data of the source domain to accomplish the supervised classification by minimizing the cross entropy loss. Meanwhile, both the teacher network and the student network take the data of the target domain, and the consistent loss (mean square error) encourages the two networks providing identical outputs for the same sample. The consistent loss is aim to bridge the gap between the source and target domains. It should be pointed out that the consistent loss encourages consistent predictions of the teacher and student networks. As the teacher network is a temporal ensemble version of the student network, the teacher network is more robust than the student network. With the consistent objective, the teacher network guides the student network on predicting the data samples of the target domain.

The teacher network is a temporal ensemble version of the student network because its weights are updated by an exponential moving average (EMA) of the weights of the student network (1).

$$W_t(I_n) = (1-\alpha)W_t(I_{n-1}) + \alpha W_s(I_n); \quad (1)$$

where I_n is the n^{th} iteration, α is a weight value ranging from 0 to 1, and W_t and W_s are weights of the teacher and student networks, respectively.

The advantages of this update strategy are:

1. Computation cost is lower than optimizing the teacher network directly.
2. The teacher network is a temporal ensemble of the student network, which is robust.

On the other hand, its disadvantages are as follows:

1. The performance of the teacher network heavily depends on the student network.
2. The teacher network barely brings diversity for the consistent loss which is important for domain adaptation.
3. The setting of α is complicated.

2.3 Data

This paper investigates the mean teacher model on the cross-city classification of the local climate zone (Stewart, Oke; Bechtel et al., 2015). The data set used in this paper is a part of the So2Sat LCZ42 dataset (Zhu et al., 2019). It has about 400,000 pairs of the Sentinel-1 and Sentinel-2 patches with annotated local climate zone labels. The Sentinel-2 patches of the city Moscow and the city Munich are used in this paper. The data patches have a size of 32 by 32 by 10. The ten channels are the ten out of the thirteen bands of the Sentinel-2 data where the first, the ninth, and the tenth bands are abandoned. More details about the data can be found in (Zhu et al., 2019; Schmitt et al., 2019). The numbers of samples of classes are shown in the table 1. For the sake of simplicity, the classes of the two cities are kept as the same ones.

3. EXPERIMENT AND DISCUSSION

3.1 Experiment setting

Two cross-city scenarios are set up in the experiments:

Class name	Compact mid-rise	Compact low-rise	Open high-rise	Open mid-rise	Open low-rise	Large low-rise	Sparsely built	Heavy industry
Class code	**2**	**3**	**4**	**5**	**6**	**8**	**9**	**10**
Moscow	330	12	861	429	117	334	403	229
Munich	336	3	24	359	765	727	138	63
Class name	Dense trees	Scattered trees	Bush, scrub	Low plants	Bare rock or paved	Water		Total
Class code	**11**	**12**	**13**	**14**	**15**	**17**		
Moscow	808	66	2	861	12	338		4802
Munich	775	48	35	775	9	775		4832

Table 1. Number of reference data of classes for both cities.

	Train	Test	Overall accuracy	Average accuracy	kappa coefficient
Baseline	Munich	Moscow	52.21% $\pm$ 1.74	32.60% $\pm$ 0.83	0.4561 $\pm$ 0.0185
Mean teacher	Munich	Moscow	53.38% $\pm$ 2.97	32.42% $\pm$ 1.50	0.4599 $\pm$ 0.0322
Target domain	Moscow	Moscow	81.07% $\pm$ 0.99	55.01% $\pm$ 1.80	0.7831 $\pm$ 0.0113
Baseline	Moscow	Munich	57.76% $\pm$ 2.16	38.45% $\pm$ 2.06	0.5200 $\pm$ 0.0239
Mean teacher	Moscow	Munich	62.71% $\pm$ 6.29	40.84% $\pm$ 2.99	0.5715 $\pm$ 0.0689
Target domain	Munich	Munich	86.75% $\pm$ 0.86	63.79% $\pm$ 1.56	0.8465 $\pm$ 0.0100

Table 2. Cross-city classification results of the baseline network and the mean teacher model indicated by overall accuracy, average accuracy, and kappa coefficient. The accuracy of the target domain are the results of training and testing on the target city, which demonstrate the best achievable results. The numbers are the mean value of ten independent experiments with the standard deviation following the "$\pm$" symbol.

1. train with data of Munich, and test on data of Moscow
2. train with data of Moscow, and test on data of Munich

For each scenario, three experiments are carried out. First, the baseline for comparison is the ResNet-18 (He et al., 2016) which is trained on the source city and tested on the target city. Second, the mean teacher model is trained on the source city and tested on the target city. The student and teacher networks in the mean teacher model have the same structure as the ResNet-18. At last, the setting is to train and test a ResNet-18 structure with data of the target city so that it demonstrate the best achievable results. In evaluation, all experiments are carried out for ten times to provide statistical robust mean accuracy.

For the training, the Adam gradient descent (Kingma, Ba) was applied with the learning rate of 1e-3. Every training procedure lasts for 100 epochs. The number of batches is 100.

3.2 Discussion

Statistical outcomes. Table 2 shows the classification accuracy of the experiments in a statistical manner. For training with data of Munich, the mean teacher model has a similar performance with the baseline algorithm in terms of accuracy. For training with data of Moscow, the mean teacher model improves the overall accuracy, the average accuracy, and the kappa coefficient by 5%, 2.4%, and 0.05 comparing to the baseline experiment. This is a considerable improvement. However, it is also noticeable that the standard deviation of the overall accuracy and the kappa coefficient are much larger than the baseline algorithm. It means that the mean teacher model is not stable in terms of those two indicators. By comparing to training and testing in the target domain, there exist more than 20% potential of the overall and average accuracy to be improved. It also illustrates the difficulties of the cross-city classification challenge.

Individual outcome. Table 3 demonstrates the classification outcomes of every repetition of the experiments. Considering the best results of all four experiments (marked in blue), the mean teacher model exhibits superior performance for both cross-city scenarios by a considerable margin. Meanwhile, the worst results of all four experiments (marked in red) suggests that the mean teacher model could also perform worse than the baseline. Therefore, it concludes the mean teacher model is not stable for the task described in this paper.

Producer accuracy. Table 4 provides the number of training sample, the number of testing samples, and the mean producer accuracy. This table demonstrates the impact of imbalanced number of samples. For the compact low-rise, the scattered trees, and the bush, scrub, their samples are limited in both cities. The mean producer accuracy of these classes are so low that it is impossible to classify them. On the other hand, for the low plants, the dense trees, the large low-rise which have a large number of samples, the producer accuracy are relately high. Therefore, the sample balance has a major impact. Table 4 also demonstrates that the adapting difficulty is directional. For example, it is a easy task to recognize dense trees when adapting from Munich to Moscow, but it is hard on the other way around.

Confusion matrix. Figure 2 provides mean confusion matrices for the four cross-city experiments. For testing on Moscow, the most obvious confusion happens to the classes of the open high-rise, the open mid-rise, the sparsely built, and the dense trees. The trend is that the algorithms prefer to classify those samples as the dense trees. The mean teacher model even has a stronger preference than the baseline model. This trend should

Baseline ResNet-18 model, Train on Munich, Test on Moscow												
Experiments	1	2	3	4	5	6	7	8	9	10	Mean	STD
Overall accuracy	0.5427	0.5165	0.5162	0.5256	0.5037	0.545	0.5496	0.511	0.495	0.516	0.52213	0.01737
Average accuracy	0.339	0.3266	0.3284	0.3322	0.3105	0.3309	0.3343	0.3206	0.3183	0.319	0.32598	0.00827
Kappa coefficient	0.4788	0.4524	0.4514	0.4596	0.4316	0.4783	0.4848	0.4469	0.4278	0.4497	0.45613	0.01846
Mean teacher model, Train on Munich, Test on Moscow												
Experiments	1	2	3	4	5	6	7	8	9	10	Mean	STD
Overall accuracy	0.4954	0.5527	0.5362	0.5183	0.5346	0.4977	0.5273	0.5289	0.5398	0.6068	0.53377	0.02969
Average accuracy	0.2985	0.3278	0.3147	0.3171	0.3355	0.3002	0.3342	0.3315	0.3387	0.3437	0.32419	0.01504
Kappa coefficient	0.4148	0.4774	0.4629	0.4415	0.4643	0.4208	0.455	0.4532	0.4717	0.537	0.45986	0.03223
Baseline model ResNet-18, Train on Moscow, Test on Munich												
Experiments	1	2	3	4	5	6	7	8	9	10	Mean	STD
Overall accuracy	0.5956	0.5317	0.5693	0.5873	0.5594	0.5882	0.5648	0.5718	0.6058	0.6024	0.57763	0.02155
Average accuracy	0.4009	0.3519	0.3743	0.406	0.3792	0.3927	0.3609	0.3621	0.4051	0.412	0.38451	0.02057
Kappa coefficient	0.5392	0.4701	0.5089	0.5303	0.4973	0.5332	0.5055	0.5173	0.5509	0.5475	0.52002	0.02389
Mean teacher model, Train on Moscow, Test on Munich												
Experiments	1	2	3	4	5	6	7	8	9	10	Mean	STD
Overall accuracy	0.6972	0.5952	0.4992	0.6952	0.6024	0.5673	0.5995	0.7005	0.6412	0.6734	0.62711	0.06293
Average accuracy	0.4434	0.4022	0.363	0.4545	0.4024	0.3684	0.3773	0.4259	0.4306	0.4166	0.40843	0.02987
Kappa coefficient	0.6474	0.5433	0.4299	0.6494	0.542	0.5076	0.5421	0.6513	0.58	0.6223	0.57153	0.06893

Table 3. Overall accuracy, average accuracy, and kappa coefficient of all repetitions of each experiment setting are shown in this table. The worst and best repetition in terms of overall accuracy are marked in red and blue, respectively.

Class code			**2**	**3**	**4**	**5**	**6**	**8**	**9**	**10**
Class name			Compact mid-rise	Compact low-rise	Open high-rise	Open mid-rise	Open low-rise	Large low-rise	Sparsely built	Heavy industry
Number of samples	Moscow		330	12	861	429	117	334	403	229
	Munich		336	3	24	359	765	727	138	63
	Train	Test	Mean Producer Accuracy							
Baseline	Munich	Moscow	0.2797	0	0.1545	0.1235	0.0684	0.7006	0.2903	0.0786
Mean teacher	Munich	Moscow	0.2606	0	0.1870	0.0326	0.0513	0.7036	0.2308	0.0480
Target domain	Moscow	Moscow	0.8600	0	0.8394	0.5916	0.4932	0.8413	0.7074	0.2748
Baseline	Moscow	Munich	0.0060	0	0.2083	0.2368	0.4301	0.5832	0.2536	0.2222
Mean teacher	Moscow	Munich	0.0030	0	0.0833	0.0808	0.6065	0.5598	0.2899	0.3810
Target domain	Munich	Munich	0.9571	0	0.1583	0.5572	0.8603	0.8547	0.5797	0.1468
Class code			**11**	**12**	**13**	**14**	**15**	**17**		
Class name			Dense trees	Scattered trees	Bush, scrub	Low plants	Bare rock or paved	Water		
	Train	Test	Mean Producer Accuracy							
Number of samples	Moscow		808	66	2	861	12	338		
	Munich		775	48	35	775	9	775		
Baseline	Munich	Moscow	0.9653	0	0	0.8606	0.0833	0.9822		
Mean teacher	Munich	Moscow	0.9963	0.0152	0	0.9477	0.0833	0.9882		
Target domain	Moscow	Moscow	0.9896	0.0818	0	0.9490	0.1167	0.9959		
Baseline	Moscow	Munich	0.4852	0.0208	0	0.9523	1.0	1.0		
Mean teacher	Moscow	Munich	0.6839	0.0833	0	0.9548	1.0	1.0		
Target domain	Munich	Munich	0.9601	0.1334	0.9778	0.9451	0.8000	1.0		

Table 4. The producer accuracy of the baseline network and the mean teacher model. The numbers are the mean value of ten independent experiments.

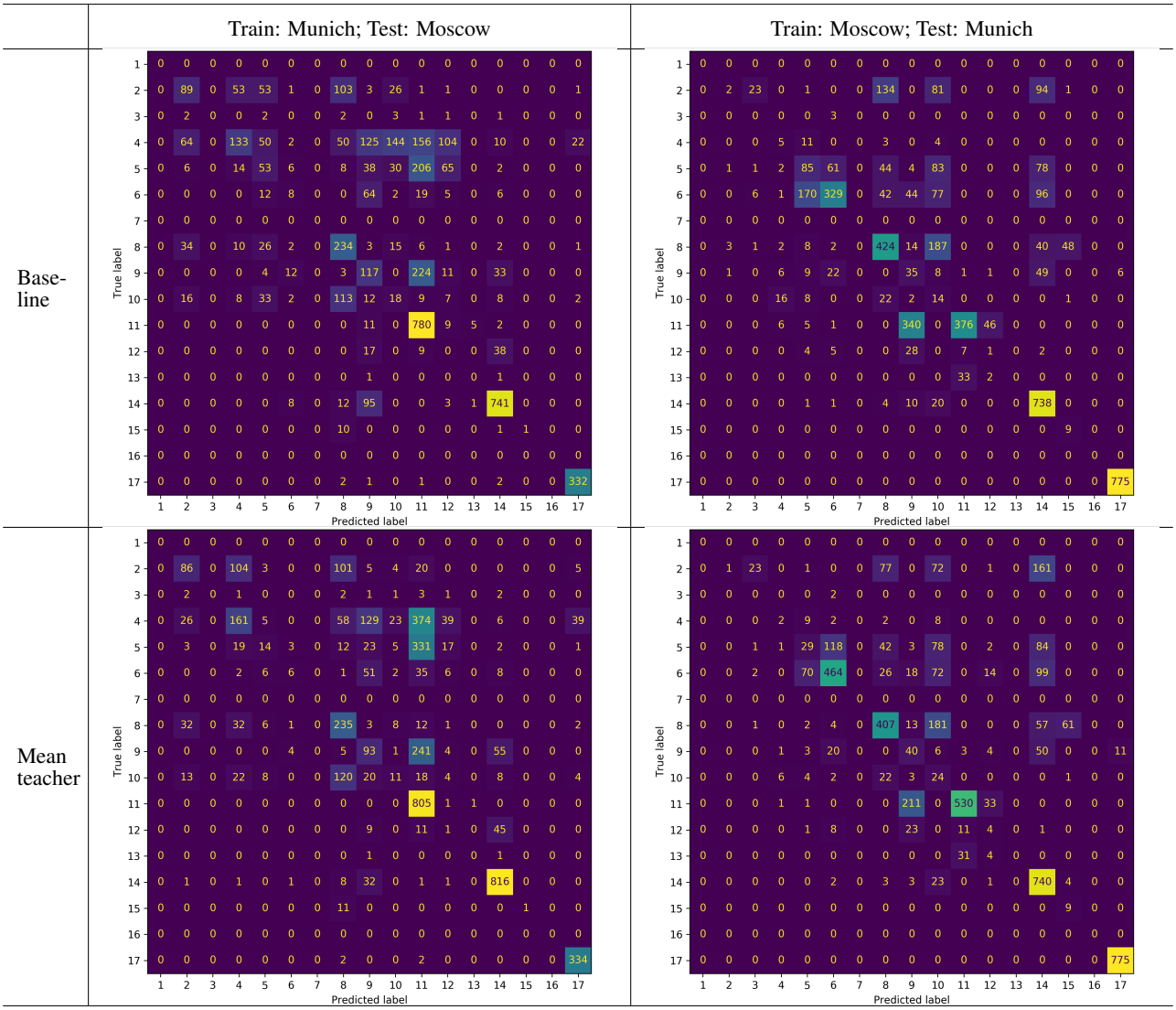

Figure 2. This figure shows the confusion matrices achieved by applying the baseline and the mean teacher model on the two cross-city scenarios. Each confusion matrix is a mean confusion matrix of ten repeated experiments, and the numbers in the confusion matrix are rounded into Integer. From left to right, top to bottom: the confusion matrix of the baseline which is trained with data of Munich and tested on data of Moscow; the confusion matrix of the baseline which is trained with data of Moscow and tested on data of Munich; the confusion matrix of the mean teacher model which is trained with data of Munich and tested on data of Moscow; the confusion matrix of the mean teacher model which is trained with data of Moscow and tested on data of Munich.

be a result of two reasons. First, those classes all have a large green coverage which makes them similar. Second, the number of dense tree samples is larger than the number of samples of other classes which leads the algorithms learn preference. For testing on Munich, a similar phenomenon can be found among the compact mid-rise, open mid-rise, open low-rise, and the low plants. The preference goes toward the low plants which have the largest number of samples. And for the case of testing on Munich, the mean teacher model is able to reduce the confusion between the sparsely built and the dense tree by comparing to the baseline model.

4. CONCLUSION

This paper investigates the cross-city classification problem where the classification algorithm is trained on data set of a city and is deployed on a data set of a different city. The cross-city scenario is a fundamental set up for a global task, yet is more challenge than the conventional ones whose training and testing data are located in the same region. This paper attempts to adapt an end-to-end unsupervised domain adaptation model, the mean teacher model, to solve the cross-city problem. The mean teacher model is implemented to be trained on the data of Munich and be tested on the data of Moscow for the local climate zone classification. The cities for training and testing were switched for an extra experiment. For comparison, the baseline model is a network of ResNet-18. Each of the experiments were repeated for ten times to provide statistical outcomes which are reliable for analysis. This work summarizes three findings from the experiments: (1) the mean teacher model has a potential to be a solution to the domain adaptation problem in remote sensing because of accuracy improvements have been found; (2) the mean teacher model is unstable according the standard deviation of accuracy resulted from repeated experiments; (3) the sample imbalance cross classes and cross source-target domain could be problematic in the domain adaptation problem of remote sensing.

Based on the findings of this work, the future work will be: (1) the mean teacher model should be tested on a large data set; (2) a strategy should be developed to overcome the impact of imbalanced samples, e.g. data augmentation; (3) the mean teacher model should be modified for remote sensing tasks.

References

Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., Stewart, I., 2015. Mapping local climate zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information*, 4(1), 199–219.

Demuzere, M., Bechtel, B., Mills, G., 2019. Global transferability of local climate zone models. *Urban climate*, 27, 46–63.

Fang, B., Kou, R., Pan, L., Chen, P., 2019. Category-Sensitive Domain Adaptation for Land Cover Mapping in Aerial Scenes. *Remote Sensing*, 11(22), 2631.

French, G., Mackiewicz, M., Fisher, M., 2017. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, W., Su, F., Huang, X., 2019. Unsupervised Adversarial Domain Adaptation Network for Semantic Segmentation. *IEEE Geoscience and Remote Sensing Letters*.

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A. et al., 2019. Outlining where humans live–The World Settlement Footprint 2015. *arXiv preprint arXiv:1910.12707*.

Melchiorri, M., Pesaresi, M., Florczyk, A. J., Corbane, C., Kemper, T., 2019. Principles and Applications of the Global Human Settlement Layer as Baseline for the Land Use Efficiency Indicator—SDG 11.3. 1. *ISPRS International Journal of Geo-Information*, 8(2), 96.

Paganini, M., Petiteville, I., Ward, S., Dyke, G., Steventon, M., Harry, J., Kerblat, F., CEOS, 2018. Satellite Earth Observations in Support of the Sustainable Development Goals. *European Space Agency*.

Pan, S. J., Yang, Q., 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345–1359.

Schmitt, M., Hughes, L. H., Qiu, C., Zhu, X. X., 2019. AGGREGATING CLOUD-FREE SENTINEL-2 IMAGES WITH GOOGLE EARTH ENGINE. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4.

Stewart, I. D., Oke, T. R., 2012. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12), 1879–1900.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 1195–1204.

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237, 111322.

Yoo, C., Han, D., Im, J., Bechtel, B., 2019. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 157, 155–170.

Zhu, X. X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Häberle, M., Hua, Y., Huang, R. et al., 2019. So2Sat LCZ42: A benchmark dataset for global local climate zones classification. *arXiv preprint arXiv:1912.12171*.