An Unsupervised Domain Adaption Framework for Aerial Image Semantic Segmentation Based on Curriculum Learning

Lingyan Ran[a,b,c*], Cheng Ji[b,c], Shizhou Zhang[b,c], Xiaoqiang Zhang[d], Yanning Zhang[b,c]
[a] Ningbo Institute of Northwestern Polytechnical University, 218 Qingyi Road, Ningbo, 315103, China
[b] School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China
[c] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an, China
[d] School of Information Engineering, Southwest University of Science and Technology, Mianyang, China
{lran*,szzhang,ynzhang}@nwpu.edu.cn, {chji_nwpu}@foxmail.com, {xqzhang}@swust.edu.cn

2022 7th International Conference on Image, Vision and Computing (ICIVC) | 978-1-6654-6734-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIVC55077.2022.9886060

***Abstract*—With the development of deep learning, semantic segmentation has made breakthrough progress, but supervised learning requires a large amount of data with pixel-level annotation. However, for remote sensing data, it is difficult to obtain large-scale pixel-level datasets. There is visual differences between the data of different geospatial regions inevitably. In particular, this difference is often referred to as a "domain gap" and can lead to significant performance degradation. The unsupervised domain adaptive method can effectively solve the above problems, by making the most of existing source domain annotated data, without re-annotating the target dataset, better semantic segmentation results can be obtained on the target dataset. In this paper, we propose a novel unsupervised domain adaptive framework based on curriculum learning (UDA-CL), and a class-aware pseudo-label filtering strategy to dynamically learn the class information during training. Comprehensive experiments show that this method achieves the encouraging semantic segmentation performance on aerial image datasets.**

***Index Terms*—aerial image semantic segmentation, domain adaption, curriculum learning, unsupervised learning**

I. Introduction

Semantic segmentation is one of the traditional tasks in computer vision. The general purpose of semantic segmentation is to assign pixel-level semantic labels by generalizing a large number of densely labeled images [1]–[3]. Along with the development of the field of remote sensing, remote sensing satellites can acquire a large amount of remote sensing image data. Effective semantic segmentation of remote sensing images can classify ground objects at pixel level, which is widely used in road network extraction [4], [5] and land cover [6]–[8] , etc. It is of great significance in updating basic geographic data, autonomous agriculture, intelligent transportation, urban planning and sustainable development, and has a wide range of practical value. There are two challenges in semantic segmentation of remote sensing images: high resolution and large scale variance, which requires huge human resources and time to label; Moreover, there are great differences in topography and architectural style in different regions, and the segmentation effect of trained models is often unsatisfactory when applied to different geographical space regions. For example, in urban and rural areas, land cover is completely different in class distribution, object scale and pixel spectrum.

This work is supported in part by the National Natural Science Foundation of China (No. 61902322, No. 62101453), the Ningbo Natural Science Foundation (No. 202003N4368), the GHfund B (No. 20210702), the College Students Innovative Entrepreneurial Training PlanProgram (No. S202110699670), the Doctoral Fund of Southwest University of Science and Technology (No. 19zx7123).

Unsupervised domain adaptive method [9]–[11] can solve this problem better. Using annotated source domain data as much as possible, better semantic segmentation results can be obtained on unseen target data sets without re-annotating the target datasets. Unsupervised domain adaptation assumes that no part of the test data is labeled and the goal is to generate high-quality segmentation even when there is a large domain shift between the training image and the test image. In this case, in order to improve the generalization ability of CNN, one of the simplest and most commonly used methods is to enrich the training data by using a variety of data enhancement technologies such as gamma correction, random contrast change, etc. In addition, the adversarial feature alignment method [12]–[15] uses generative adversarial networks (GAN) [16], [17] to minimize the distance between feature representation of source domain and target domain, where discriminators can be used at multiple levels. In addition, the method based on image style transfer [7], [18], [19] is to transform the style of the source domain image to the target domain under the condition of preserving the image content, so as to use the label of the source domain image for training. Most of these methods are also implemented by generative adversarial networks.

This paper is closer to the algorithm based on pseudo-label generation. A lot of work adopts the method of self-training on pseudo-label [20]–[22]. High quality pseudo-labels are sought through self-training to achieve category prediction with high reliability. Most methods compute the label "offline" beforehand, then use it to update the model and repeat the process for several rounds. More recent frameworks that follow this strategy rely on adversarial training, style transfer, or both.

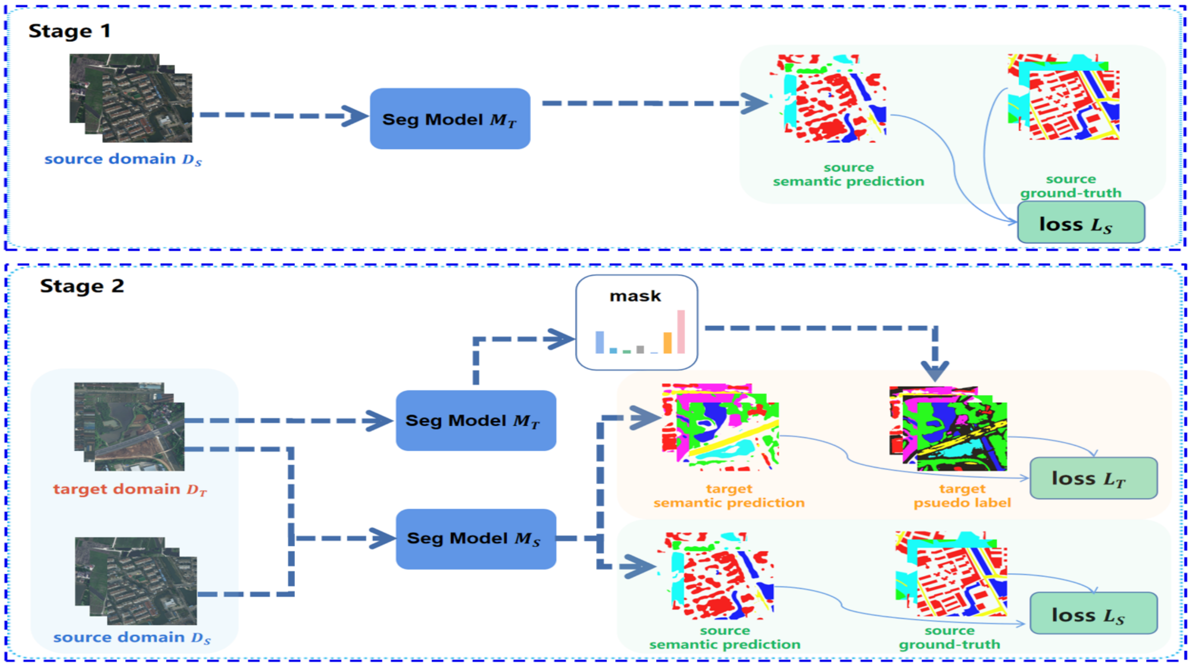


Fig. 1. Overview of the UDA-CL network architecture. The training process consists of two steps. In step 1, the teacher model M_T is pre-trained using source domain data and reused next with fixed parameters. In Step 2, source domain data and target domain data are simultaneously put into the student model M_S (consistent with the teacher model) for training. We use teacher network to generate pseudo labels for the target domain: according to the proportion of predicted pixels of each category, we set the confidence threshold of linear growth to generate masks, so as to obtain pseudo labels of the region with higher confidence. In the training process, the pixel number of pseudo labels is gradually increased to realize the learning from easy to difficult.

The main contributions of this paper are: 1) it presents a simple and effective unsupervised domain adaptive framework with curriculum learning (UDA-CL). The framework adopts the idea of curriculum learning and the method of pseudo label generation to realize the adaptive semantic segmentation of remote sensing image domain of urban and rural areas; 2) it realizes the pseudo label of target domain data from easy to difficult through dynamic modification to achieve stable and effective training.

II. Related work

A. Semantic segmentation

Semantic segmentation is a challenging visual task that aims at obtaining pixel-wise category predictions. The emergence of full convolutional neural network (FCN) [23] greatly improves semantic segmentation performance, but it ignores context information. In order to achieve higher resolution prediction, [24], [25] further applies deconvolution layer to CNN with good performance. On the other hand, in order to learn long-range context dependence better, researchers proposed dilated convolution [26]–[28], spatial pyramid pooling [29], attention mechanism [30]–[32], and other methods to increase the receptive field of convolution layers.

B. Domain adaptation for Semantic segmentation

With the rapid improvement of semantic segmentation network performance, people committed to apply deep learning method to the remote sensing image analysis. Semantic segmentation of remote sensing images faces several challenges, such as lack of training data and pixel-level accuracy requirements. Although the number of remote sensing images is very large, there is a lack of training data of pixel annotations. The topography and landform of different regions in remote sensing images will be different, for example, the architectural style and vegetation type of urban and rural will be greatly different. Unsupervised domain adaptive can effectively solve the problem of large differences in data fields. The tasks of source domain and target domain are the same, but there are differences in data distribution. At present, there are methods such as feature alignment based on adversarial training [12]–[15], image style transfer [7], [18], [19] and pseudo-label generation based on self-training [20]–[22].

C. Curriculum learning

Curriculum learning (CL) is a popular frontier direction in recent years. Bengio [33] first proposed the concept of Curriculum learning [34], which is a training strategy that imitates the learning process of human beings and advocates that the model should start learning from easy samples and gradually advance to complex samples and knowledge. In this paper, pixels with high confidence in the prediction map are relatively easier to learn, while pixels with low confidence are more difficult for the model. From easy to difficult, pixels are gradually added into the model for training, so as to gradually achieve better and more stable segmentation effects.

III. METHOD

In this section, we elaborate the proposed UDA-CL framework for aerial image semantic segmentation.

A. Overall Framework

Given a set of labeled data in source domain $\mathbf{D}_s = \{\mathbf{X}_s, \mathbf{Y}_s\}$ and unlabelled data in target domain $\mathbf{D}_t = \{\mathbf{X}_t\}$, where $\mathbf{X}_s$ is source domain image with its corresponding label $\mathbf{Y}_s$, $\mathbf{X}_t$ is target domain image, the goal of unsupervised domain adaptation is to use labeled source domain data in $\mathbf{D}_s$ and unlabeled target domain data in $\mathbf{D}_t$ to train a model, which will perform well on the unseen test data in the target domain. In our work, the two domain datasets share the same label space.

As shown in Figure 1, our UDA-CL framework consists of two stages. Stage 1 performs a teacher model training procedure on source domain dataset. And the teacher model, named as $\mathbf{M_T}$, stays fixed afterwards. On stage 2, both the source and target domain data are fed into the student model. Note that $\mathbf{M_T}$ in Stage 2 is used for two aspects. One is for the initialization of the student model, and the other is to generate the pseudo labels for the target domain dataset.

Firstly, we use source data $\mathbf{X}_s$ and its corresponding ground-truth $\mathbf{Y}_s$ to warm up the model, and save the pre-trained weights as $\mathbf{M}_T$. Then the pseudo labels $\hat{\mathbf{Y}}_t$ are produced on target domain, data from both domains are put into the network for training in the following stage.

We define the loss for source domain data, denoted by $\mathcal{L}_s$, as a standard pixel-level cross-entropy loss to measure the ground truth Y_s. While the loss for target domain is denoted by $\mathcal{L}_t$, which uses probability map $\mathbf{P}_t^{(h,w,c)}$ and its pseudo labels $\hat{\mathbf{Y}}_t^{(h,w,c)}$.

$$\mathcal{L}_s = -\sum_{h,w}\sum_{c}\mathbf{Y}_s^{(h,w,c)} log \mathbf{P}_s^{(h,w,c)}. \tag{1}$$

$$\mathcal{L}_t = -\sum_{h,w}\sum_{c}\hat{\mathbf{Y}}_t^{(h,w,c)} log \mathbf{P}_t^{(h,w,c)}. \tag{2}$$

Our objective is to minimize overall loss $\mathcal{L}$, formulated as:

$$\mathcal{L} = \mathcal{L}_s + \lambda * \mathcal{L}_t, \tag{3}$$

where λ is a hyper-parameter that adjusts the contribution of unlabeled information.

B. Class-balanced Label Sampling

Pseudo-labels $\hat{\mathbf{Y}}_t$ are generated according to the confidence of the model prediction. We set different thresholds for each category, and select the pixels whose confidence is higher than the threshold of this category for annotation, and ignore the rest pixels.

$$\hat{\mathbf{Y}}_t^{i,j} = \begin{cases} \underset{c}{argmax}\, \mathbf{P}_t^{i,j}, & \mathbf{P}_t^{i,j} > \tau_c \\ ignore, & \text{otherwise} \end{cases} \tag{4}$$

There is an obvious class imbalance problem in semantic segmentation tasks. Some categories have very few pixels and only appear in a small number of data samples. For this "long tail" phenomenon, we adopt class-balanced sampling, and set the threshold for each category respectively, so as to achieve the consistency of the class distribution of the selected samples and the training set as far as possible:

$$N_c = \sum_{h,w}\hat{\mathbf{Y}}_t^{(h,w,c)} \tag{5}$$

We can obtain the class distribution of the target domain, where N_c is the total number of pixels predicted to be class c in the target domain, and N_t is the total number of pixels in the target domain.

$$\sigma_c = N_c / N_t \tag{6}$$

We sort the predicted confidence from high to low, and select the top N_c pixels in class c as the pseudo label to sample for subsequent training. The confidence threshold of the top N_c-th pixel is τ_c.

C. Curriculum Learning Strategy

We believe that the model is easier to learn the pseudo labels with higher confidence, while the pseudo labels with lower confidence are more difficult to learn. Based on this, we dynamically divided the difficulty of label for training, adding the pseudo labels that are easy to learn at the beginning, and gradually adding the pseudo-labels that are relatively difficult during the training process. If self-pace is completely adopted, the model is likely to be greatly affected by pseudo label noise in the training process. In this case, the pre-trained model $\mathbf{M}_S$ is used to generate pseudo labels $\hat{\mathbf{Y}}_t$ at one time with different thresholds τ_c, so that the training process is more stable and perform better.

The whole process of the proposed framework is detailed in Algorithm 1. In the process of target domain data training, the number of pseudo labels increases linearly.

Algorithm 1: UDA-CL: unsupervised domain adaption framework

Input: Labeled source domain training set $\mathbf{D}_s$, unlabeled target domain training set $\mathbf{D}_t$, semantic segmentation model $\mathbf{M}_T, \mathbf{M}_S$,
Output: Fine trained model $\mathbf{M}_S$.
Step 1:
Train teacher model $\mathbf{M}_T$ on $\mathbf{D}_s$ with $\mathcal{L}_s$;
Step 2:
Initialize student model $\mathbf{M}_S \leftarrow \mathbf{M}_T$;
Predict pseudo label $\hat{\mathbf{Y}}_t$ on $\mathbf{D}_t$ with $\mathbf{M}_T$;
for $k_i = k_0$ to K **do**
  τ_c = $N_c * k_i$;
  Obtain $\hat{\mathbf{Y}}_t$ by using τ_c to mask $\mathbf{P}_t$
  for minibatch $\{\mathbf{x}_{s,i}, \mathbf{x}_{t,i}\} \subset \{\mathbf{D}_s, \mathbf{D}_t\}$ **do**
    Train $\mathbf{M}_S$ on $\{\mathbf{x}_{s,i}, \mathbf{x}_{t,i}\}$ with loss $\mathcal{L}_s$, $\mathcal{L}_t$
  end for
end for
Return $\mathbf{M}_S$.

We set the proportion of the pixel number of the ini pseudo-label to the pixel number of all target domain datε k_0, and proceed in a cycle: when the loss function of mc training is stable, the ratio of the pseudo-label is increasec k_i, and the pseudo-label is updated according to the curı ratio, and the training continues until maximum proport K is reached. In this process, the proportion of pseudo-la pixels increases linearly, and the proportion of each increm is consistent.

IV. Experiments

A. Experimental Settings

Dataset: LoveDA [35] dataset encompasses both urban and rural domains, the urban dataset is composed of 1156 images for training, 677 images for validation and 820 for testing, and the rural dataset is composed of 1366 images for training, 992 images for validation and 976 for testing. The spatial resolution is 0.3 m, with red, green, and blue bands.

Model: We implement the semantic segmentation model with DeepLabv2 and employ ResNet-50 as the backbone, which is pre-trained on ImageNet. The Adam optimizer was used for the discriminator with the momentum of 0.9 and 0.99. The number of training iterations was set to 10k, with a batchsize of 16. Each batch consists of eight source domain images and eight target domain images randomly extracted from the datasets. The threshold setting for filtering pseudo-label pixels increases by 5% each time from 20% of the number of pixels in the training set in the target domain to the maximum 50%.

B. Comparisons with State-of-the-Arts

As is shown in Table II, the Oracle setting obtains the best overall performances. Compared with the adversarial training method, the self-training method address the problem of class imbalance with pseudo label generation, and achieves better performance. Our method achieves the highest overall mIoU score in rural $\rightarrow$ urban experiments, and 0.45% mIoU higher than CBST [36]. Table III shows the performance on reverse domains. Due to the inconsistent category distribution, IAST [37] has the highest accuracy in urban $\rightarrow$ rural experiments, our method is 2.22% mIoU higher than CBST, and the categories with few pixels like Building, Road, and Barren achieved higher score than IAST. Figure 2 is the loss curve of CBST and UDA-CL in training. As illustrated, the loss of UDA-CL fluctuates less in the early stage and is more stable. We believe that the reason lies in the fact that the pseudo-label generated by our teacher network is more stable. Sample results are visualized in Figure 3.

C. Ablation Studies

TABLE I
Ablation Study On Rural $\rightarrow$ Urban Test.

Method	mIoU(%)	$\triangle$
w/o CL	39.68	-2.09
w/o sample	35.46	-6.31
UDA-CL	41.77	-

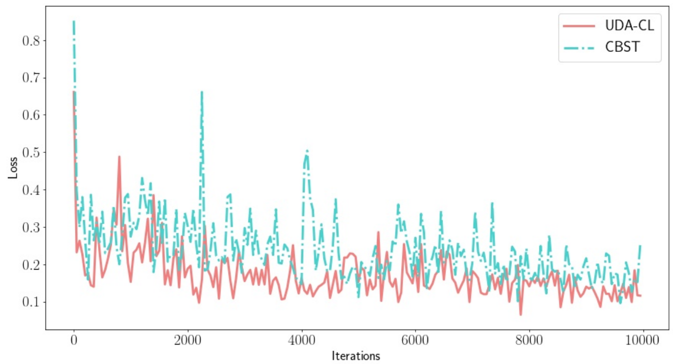


Fig. 2. CBST [36] and UDA-CL loss function curves.

In this part, we conducted ablation experiments on whether to use curriculum learning and whether to sample high confidence pixel values, and the results in table I showed that mIoU decreased by 2.09% in the training method of one-time generation of pseudo-labels in the target domain without CL. mIoU is reduced by 6.31% by randomly selecting pixels to generate pseudo-labels instead of high confidence pixel selection.

V. Conclusion

This paper addresses the label consuming problem when manipulating domain adaption for aerial images. We managed to produce high confidence pseudo labels with the curriculum learning method on large amount of unlabeled target domain images. Experimental results on the publicly available LoveDA dataset confirms the efficiency of the proposed framework. In the upcoming works, better performance seems promising with advanced CL variants.

References

[1] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Computer Vision amp; Pattern Recognition*, 2014.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. *IEEE*, 2016.

[3] B. Zhou, Z. Hang, Francesco Xavier Puig Fernandez, S. Fidler, and A. Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] C. Lacoste, X. Descombes, and J. Zerubia. Road network extraction in remote sensing by a markov object process. In *International Conference on Image Processing*, 2003.

[5] Y. Chen, W. Li, and L Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. 2017.

[6] L. Bruzzone and R. Cossu. A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. *IEEE Transactions on Geoscience and Remote Sensing*, 40(9):1984–1996, 2002.

[7] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–16, 2020.

[8] C. Persello and L. Bruzzone. Active learning for domain adaptation in the supervised classification of remote sensing images. *IEEE Transactions on Geoscience Remote Sensing*, 50(11):4468–4483, 2012.

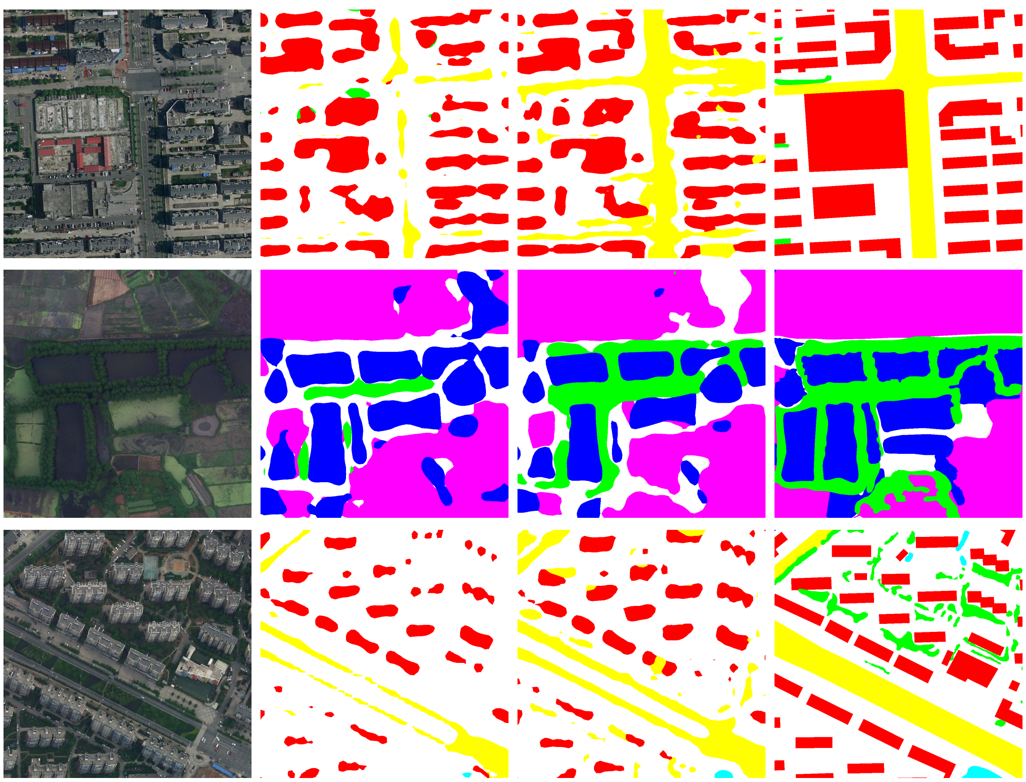
[9] L. Bruzzone and M. Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE Trans Pattern Anal Mach Intell*, 32(5):770–787, 2010.

TABLE II
UNSUPERVISED DOMAIN ADAPTATION RESULTS OBTAINED ON THE TEST SET OF THE LOVEDA DATASET, RURAL → URBAN.

Method	Type	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU(%)
Oracle	-	48.18	52.14	56.81	85.72	12.34	36.70	35.66	46.79
Source only	-	43.30	25.63	12.70	76.22	12.52	23.34	25.14	31.27
MCD [38]	-	43.60	15.37	11.98	79.07	14.13	33.08	23.47	31.53
AdaptSeg [39]	AT	42.35	23.73	15.61	81.95	13.62	28.70	22.05	32.68
FADA [40]	AT	43.89	12.62	12.76	80.37	12.70	32.76	24.79	31.41
CLAN [14]	AT	43.41	25.42	13.75	79.25	13.71	30.44	25.80	33.11
TransNorm [41]	AT	38.37	5.04	3.75	80.83	14.19	33.99	17.91	27.73
PyCDA [20]	ST	38.04	35.85	45.51	74.87	7.71	40.39	11.39	36.25
CBST [36]	ST	48.37	46.10	35.79	80.05	19.18	29.69	30.05	41.32
IAST [37]	ST	48.57	31.51	28.73	86.01	20.29	31.77	36.50	40.48
UDA-CL	ST	48.15	37.44	45.05	84.29	16.68	26.66	34.12	**41.77**

TABLE III
UNSUPERVISED DOMAIN ADAPTATION RESULTS OBTAINED ON THE TEST SET OF THE LOVEDA DATASET, URBAN → RURAL.

Method	Type	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU(%)
Oracle	-	37.18	52.74	43.74	65.89	11.47	45.78	62.91	45.67
Source only	-	24.16	37.02	32.56	49.42	14.00	29.34	35.65	31.74
MCD [38]	-	25.61	44.27	31.28	44.78	13.74	33.83	25.98	31.36
AdaptSeg [39]	AT	26.89	40.53	30.65	50.09	16.97	32.51	28.25	32.27
FADA [40]	AT	24.39	32.97	25.61	47.59	15.34	34.35	20.29	28.65
CLAN [14]	AT	22.93	44.78	25.99	46.81	10.54	37.21	24.45	30.39
TransNorm [41]	AT	19.39	36.30	22.04	36.68	14.00	40.62	3.30	24.62
PyCDA [20]	ST	12.36	38.11	20.45	57.16	18.32	36.71	41.90	32.14
CBST [36]	ST	25.06	44.02	23.79	50.48	8.33	39.16	49.65	34.36
IAST [37]	ST	29.97	49.48	28.29	64.49	2.13	33.36	61.37	38.44
UDA-CL	ST	28.55	49.69	35.74	53.52	4.96	31.36	52.26	36.58



(a) Images (b) CBST [36] (c) UDA-CL (d) GT

Fig. 3. Examples selected from CBST [36] and UDA-CL.

[10] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *IEEE*, 2018.

[11] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing*, 11(11):1369–, 2019.

[12] Long, Jonathan, Shelhamer, Evan, Darrell, and Trevor. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2017.

[13] T. H. Vu, H Jain, M. Bucher, M. Cord, and P Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[14] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[15] F. Pan, I. Shin, F. Rameau, S. Lee, and I. Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. 2020.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[17] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. *MIT Press*, 2015.

[18] J. Hoffman, E. Tzeng, T. Park, J. Y. Zhu, and T. DaRrell. Cycada: Cycle-consistent adversarial domain adaptation. 2017.

[19] M. Kim and H Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. 2020.

[20] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6758–6767, 2019.

[21] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 129(4):1106–1120, 2021.

[22] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Fang Wen, and Wenqiang Zhang. Dual path learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9082–9091, 2021.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[24] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015.

[25] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(12):2481–2495, 2017.

[26] Chen Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[27] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[28] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Springer, Cham*, 2018.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[30] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and J. Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[31] X. Li, Z. Zhong, J. Wu, Y. Yang, and H. Liu. Expectation-maximization attention networks for semantic segmentation. 2019.

[32] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[34] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. *IEEE Computer Society*, 2017.

[35] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[36] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

[37] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, pages 415–430. Springer, 2020.

[38] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[39] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[40] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European conference on computer vision*, pages 642–659. Springer, 2020.

[41] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. *Advances in neural information processing systems*, 32, 2019.