Domain Adaption for Fine-Grained Urban Village Extraction From Satellite Images

Qian Shi, *Member, IEEE*, Mengxi Liu, Xiaoping Liu, Penghua Liu, Pengyuan Zhang, Jinxing Yang, and Xia Li

***Abstract*— Urban villages (UVs) are distinctive products formed in the process of rapid urbanization. The fine-grained mapping of UVs from satellite images has always been a considerable challenge because of the complex urban structures and the insufficiency of labeled samples. In this letter, we propose using the domain adaptation strategy to tackle the domain shift problem by employing adversarial learning to tune the semantic segmentation network so as to adaptively obtain similar outputs for input images from different domains. The proposed method was coupled with several segmentation networks, including U-Net, RefineNet, and DeepLab v3+, and the results show that domain adaptation can significantly improve the pixel-level mapping of UVs.**

***Index Terms*— Adversarial learning, domain adaptation, satellite images, semantic segmentation, urban village (UV).**

I. Introduction

THE megacities have witnessed an unprecedented increase in urban population due to a large amount of rural–urban migration [1]. Most of these migrants prefer living in villages with poor living environments called "urban villages" (UVs). Because of insufficient infrastructures and high crime threats, the UVs are hindering the process of urban modernization and in urgent need of redevelopment. Therefore, to monitor the real-time change, UV mapping by utilizing remote sensing imagery has gained tremendous attention [2].

In recent years, with the availability of very high-resolution images, more details of objects can be observed from satellite images, which also bring new problems, such as high intraclass variance and low interclass variance of surface objects. Accordingly, more sophisticated image classification methods are needed to deal with these problems. UVs are presented as clustered buildings with irregular arrangements in appearance. In order to accurately extract UVs, hand-engineering methods based on texture and structure features, such as the gray-level co-occurrence matrix, were used extensively in UV mapping [3], which were recently replaced by data-based approaches, such as deep learning. One of the superiorities of deep learning is the ability to learn hierarchical representations automatically [4], and recent advances in image recognition and object detection [5], [6] have proven its competitive performance over methods using hand-crafted features. In the remote sensing field, deep learning models based on convolutional neural networks (CNNs) have been widely used in land-use classification and feature extraction [7], [8]. However, the CNNs were limited to the classification of fixed-size images, which is far from meeting the actual needs of producing fine-grained pixelwise thematic maps. Therefore, fully convolutional network-based models were proposed to recover the spatial details through the encoder–decoder architecture, in which the encoder extracts hierarchical semantic features and the decoder recovers details at pixel level by utilizing transposed convolutions [9].

In order to meet the demands of large-scale UV mapping, the semantic knowledge should be transferred to new regions. However, the distribution of the appearance of UVs may be different in different cities, and even weather and illumination conditions will break the independently identically distribution assumption on the training and testing models. Hence, pretrained models in one city may not work well in another city, especially on the different acquisition times. Therefore, domain adaptation is required to minimize the difference between source and target domains [10].

A popular way for domain adaptation in deep learning is fine-tuning [11], which slightly adjusts the weights of pretrained models on the source domain by backpropagation according to the newly labeled samples from the target domain. However, it is expensive to obtain enough manually labeled samples to achieve large-scale mapping. In this case, unsupervised transfer learning methods which need no training samples in the target domain are suitable for cross-regional mapping. The key point of unsupervised transfer learning is to align the domain shift between the source and target domains [12].

Domain adaptation methods based on adversarial learning find an ingenious way to minimize cross-domain discrepancy with a discriminator and a generator [13]. The discriminator is a network designed to determine the origin of a distribution, while the generator aims to fool the discriminator by generating similar representations of samples from both

Manuscript received April 30, 2019; revised October 9, 2019; accepted October 11, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0604402, in part by the Key National Natural Science Foundation of China under Grant 41531176, in part by the National Natural Science Foundation of China under Grant 41871318, and in part by the National Nature Science Foundation of China under Grant 61601522. *(Corresponding author: Penghua Liu.)*

Q. Shi, M. Liu, X. Liu, P. Liu, and P. Zhang are with the Guangdong Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: shixi5@mail.sysu.edu.cn; liumx23@mail2.sysu.edu.cn; liuxp3@mail.sysu.edu.cn; liuph3@mail2.sysu.edu.cn; zhangpy9@mail2.sysu.edu.cn).

J. Yang is with the School of Geographical Sciences, Guangzhou University, Guangzhou 510006, China (e-mail: yangjx11@gzhu.edu.cn).

X. Li is with the School of Geographic Sciences, East China Normal University, Shanghai 200241, China, and also with the Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200241, China (e-mail: lixia@geo.ecnu.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2019.2947473

domains. Unlike traditional domain adaptation methods, which only adopt high-level semantic features from both domains to build an effective pipeline, adversarial domain adaptation exploits not only global knowledge but also local spatial layout to measure the distance between the source and target domains [14]. For example, even though there is domain shift in original space, the output segmentation result may be similar.

In this letter, we introduce an adversarial domain adaptation method for semantic segmentation to overcome the domain shift and mapping cross-regional UVs at pixel scale by using several state-of-the-art (SOTA) semantic segmentation techniques, which provides an efficient and low-cost method for obtaining large-scale mapping of UVs. Based on adversarial learning, the network also contains a discriminator and a generator. In order to minimize the adversarial loss, the proposed segmentation model aims to confuse the discriminator by generating more similar label outputs for both source and target images.

Related works of semantic segmentation and domain adaptation are presented in Section II. Section III introduces the domain adaptation method used in this letter. In Section IV, we show the experimental settings and results. At last, we make a conclusion in Section V.

II. Related Works

A. Semantic Segmentation

Early semantic segmentation methods are based on low-dimensional or hand-crafted features combined with shallow architectures, which fail to extract the intrinsic features behind images and are time-consuming. Deep CNNs, due to their excellent performance in computer vision, now dominate the semantic segmentation field. The proposal of a fully convolutional network initiated the end-to-end way of dense predictions [15]. However, because of the information loss in the encoding architecture, the accuracy could not meet the practical requirement. Later studies tackled the aforementioned problems by training upsampling filters (decoders) or applying atrous convolutions and global/pyramid pooling to retain a large receptive field, and thus obtained more accurate dense predictions [16]. Apart from limitations in architectures, requirements for massive dense annotations also hinder models from achieving better results. Manual annotations are substantially time-consuming and may cause ambiguity. Consequently, weakly and semisupervised methods were proposed to reduce the difficulty in ground truth labeling. Other approaches, e.g., relating annotated synthetic data sets with real-world data sets [17], may cause poor performance due to the inconsistency in reflecting the real world using artificial images. Hence, more labor-saving and reliable techniques, such as domain adaptation, need to be explored.

B. Domain Adaptation

Domain adaptation aims at tackling domain shift problems by optimizing a learner across different domains with different distributions. This problem was usually resolved by either matching marginal distributions [18] or conditional distributions [19]. Maximum mean discrepancy (MMD) was a loss function to measure distance of distributions, which is a traditional and widely used way for domain adaptation [20]. Recent studies focused on deep architectures because they have recognized advantages in capturing the essential yet invariant features from data. But deep networks can neither be transferred to data with different distributions nor eliminate cross-domain discrepancy, and latest studies have further developed the domain adaptation techniques by using a generative adversarial network [21]. An excellent work is adversarial discriminative domain adaptation [10], which generalized a framework for domain adaptation and introduced an adversarial strategy without the need of annotations from the target domain. Later, a conditional domain adaptation network that adopts conditioning strategies to improve discriminability and transferability has significantly improved the accuracy. Though previous work has successfully transferred knowledge across domains at the pixel level [22] and further adopted adversarial strategy to semantic segmentation in a fully convolutional network [23], [24], pixel-level domain adaptation has not been extensively explored and is still a challenging field due to difficulties in feature representations.

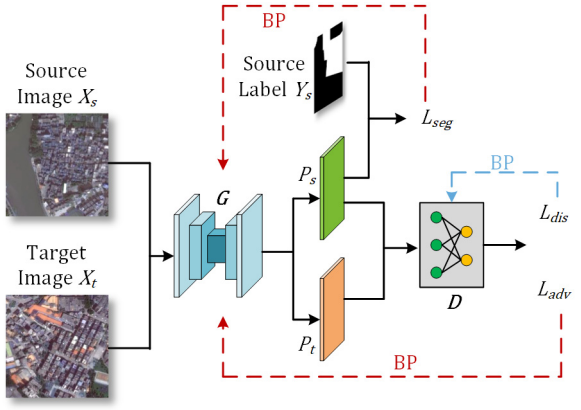

Fig. 1. Workflow of the proposed domain adaptation method for UV extraction. Red dashed lines: backpropagation stage when training the generator, i.e., the segmentation model. Blue dashed line: backpropagation stage when training the discriminator. More details about L_{seg}, L_{dis}, and L_{adv} are presented in Sections III-A and III-B. (BP: backpropagation.)

III. Methodology

In this section, a brief overview of the adopted algorithm and the model details is provided first. Then, a comprehensive summary of the optimization process is presented.

We adopt an innovative domain adaptation algorithm in this letter. The algorithm mainly comprises two modules: a segmentation network and a discriminator. We first train the segmentation network G with annotated source images and labels (i.e., X_s and Y_s). Then, we feed the pretrained network with unannotated target images (i.e., X_t) to obtain their segmentation predictions (i.e., P_t). To realize domain adaptation, predictions from both domains (i.e., P_s and P_t) are taken as inputs for discriminator D to distinguish whether they come from the source or target domain. G is then optimized to generate similar segmentation results by minimizing the adversarial loss. Fig. 1 shows the basic idea of the algorithm.

A. Discriminator

The discriminator follows the architecture used in [18] with five convolutional layers, except for the utilization of

fully convolutional layers to preserve spatial information. The first four convolutional layers are each followed by a leaky ReLU with a coefficient of 0.2 of the leaky part. The last convolutional layer is followed by an up-sampling layer to rescale the predictions to the size of input images. No batch-normalization techniques are used because the training is based on a small batch size. The objective for the discriminator is to train a two-class classifier. Suppose $P = G(X) \in R^{h*w*c}$ denotes the segmentation results of both domains, where X denotes the input images of G from either the source or target domain, and h, w, and c denote the image height, width, and the number of classes, respectively. We then forward P to train the discriminator D using a binary cross-entropy loss, which is presented as

$$L_{\text{dis}} = -\sum_{h,w} y\log(D(P)) + (1-y)\log(D(P)). \tag{1}$$

B. Segmentation Network

Recent studies have suggested the significance of selecting a good base model for experiments to achieve better performance. Consequently, we adopt three latest eminent segmentation models as our base model, including U-Net [25], RefineNet [26], and DeepLab v3+ [16].

Because high-level feature representations extracted from convolutional layers contain global visual information, they are regularly employed in domain adaptation for image classification tasks. However, in semantic segmentation tasks, too complex representations from high-level features have proved not to be the best choice for adaptation.

Segmentation maps, though commonly assumed to be low dimensional and less representative, contain very rich information such as scene layout and context. Intuitively, segmentation results from diverse domains should share notable similarities in both global and local scales. Therefore, we choose segmentation predictions for domain adaptation via adversarial learning.

The segmentation loss in the source training procedure is defined as (2), where Y_s represents the ground truth of source images

$$L_{\text{seg}}(X_s) = -\sum_{h,w}\sum_{c} Y_s\log(P_s). \tag{2}$$

After training the segmentation network, we forward target images to G and obtain their predictions P_t, and the adversarial loss can then be defined as (3). By maximizing the probability for target prediction being recognized as the source, the discriminator can then be fooled, and thus, we reach our goal

$$L_{\text{adv}}(X_t) = -\sum_{h,w}\log(D(P_t)). \tag{3}$$

C. Network Training

Finally, combining the two losses illustrated earlier, the adaptation task can be formulated as (4), where γ is the weight to balance the two losses

$$L(X_s, X_t) = L_{\text{seg}}(X_s) + \gamma\, L_{\text{adv}}(X_t). \tag{4}$$

The general goal of optimization follows the criterion defined in (5), indicating a minimization process of segmentation loss and a maximization process of target predictions being recognized as the source:

$$\max_{D}\min_{G} L(X_s, X_t). \tag{5}$$

In the course of experimental exploration, we find that training G and D simultaneously is more effective than training them individually. Therefore, the training process can be summarized as follows. We first use source samples to train the segmentation network with a goal to minimize L_{seg}. Additionally, we generate segmentation predictions from both domains and feed them to the discriminator to minimize L_{dis}. Meanwhile, we compute the adversarial loss L_{adv} and use it to confuse the discriminator. In this way, the segmentation model can be adapted to the target domain by adversarial learning.

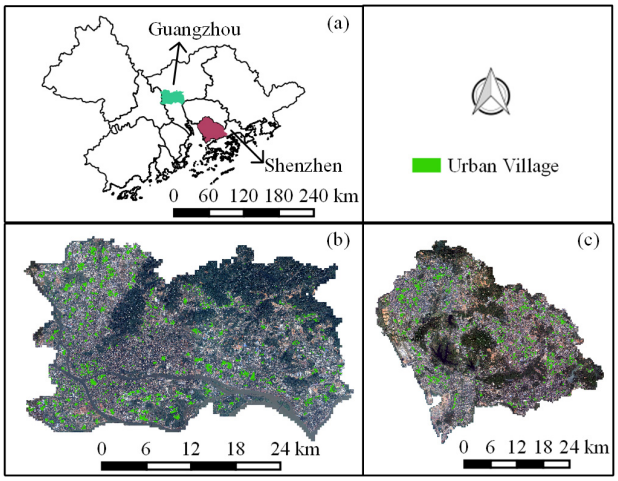

Fig. 2. Study areas. (a) Locations of Guangzhou and Shenzhen in PRD. (b) Study area in Guangzhou. (c) Study area in Shenzhen.

IV. Experiments and Analysis

A. Data Sets

As shown in Fig. 2, in our experiments, Guangzhou and Shenzhen were selected as study areas. Although these two metropolises are both located in the Pearl River Delta (PRD), different historical backgrounds have led to their differences in economic, cultural, and even urban appearance. The only data applied in the experiments were three-band (R, G, and B) Google Earth images with a spatial resolution of about 1 m. There may be huge differences between images due to diverse acquisition time and illumination conditions of satellite. We randomly cropped the satellite images and corresponding pixelwise labels into image patches with a size of 256 × 256, which were suitable for the size of UV at 1-m spatial resolution and convenient for feature extraction in the CNN model without too much GPU occupation. All pixel values were normalized between 0 and 1. Ultimately, the sample set of Guangzhou had around 14 000 sample patches and that of Shenzhen had 20 000. In the experimental process, 70% of the sample sets were used for training, and the rest 30% were used for validation.

B. Experimental Setups

To overcome the domain shift, we first trained a base model on the source-domain data set, which could be adapted to the

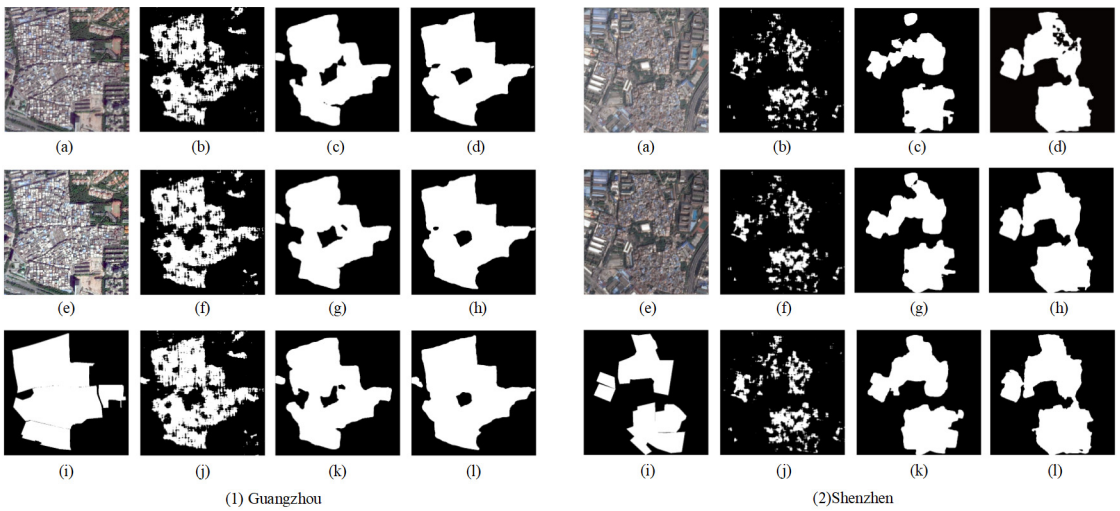

Fig. 3. Examples. (a) Image. (b) U-Net. (c) RefineNet. (d) DeepLab v3+. (e) Image + MMD. (f) U-Net + MMD. (g) RefineNet + MMD. (h) DeepLab v3+ + MMD. (i) Ground truth. (j) U-Net + domain adaptation. (k) RefineNet + domain adaptation. (l) DeepLab v3++ domain adaptation.

target domain using a few unlabeled target images. Due to the lack of labeled samples, two groups of parallel experiments were conducted, taking one of Guangzhou and Shenzhen as the source domain and the other one as the target domain to make full use of our existing data. To prove the validity of the above domain adaptation algorithm, we also compared the adapted model with the traditional MMD-based method, as well as the base model without any transfer tricks.

Furthermore, three widely used segmentation models, namely, U-Net, RefineNet, and DeepLab v3+, were introduced for model comparison in order to evaluate the capabilities of different segmentation networks on UV extraction and their suitability on the domain adaptation algorithm. Brief descriptions of these three models are as follows.

1) *U-Net:* U-Net was first proposed by Ronneberger *et al.* [25] for biomedical image segmentation. It adopted an encoder–decoder architecture consisting of a contracting path for downsampling and a symmetric expanding path to capture contexts and gradually recover details. It has been widely used in many specific applications for its remarkable performance.
2) *RefineNet:* RefineNet is a multipath refinement network proposed by Lin *et al.* [26] with the purpose to refine object details through multiple cascaded residual connections. We used a 4-cascaded 1-scale RefineNet in the experiments, which adopted ResNet-101 as the encoder.
3) *DeepLab v3+:* DeepLab v3+ is an excellent network newly proposed by Zhang *et al.* [14]. It considers the advantages of both the decoder–encoder architecture and atrous spatial pyramid pooling (ASPP). DeepLab v3+ applies an improved xception-41 as an encoder to capture contexts and an ASPP module to aggregate multiscale contexts. It has gained a new SOTA performance in benchmarks, including VOC 2012.

In summary, six base segmentation models were trained for our experiments, all of which adopted a batch size of two images and the Adam optimizer with a learning rate of 0.0001 to make the model converge quickly. Each model was trained for 80 epochs while the learning rate decreased at a rate of 0.9 per epoch. To avoid overfitting, an L_2 regularization was employed to convolutions whose weight decay equals to 0.0001. Each base model was retrained in a domain adaptation network. The strategy to optimize the discriminator is the same as above. After 1000 iterations of adversarial training, the adapted models applicable to the target domain were obtained. Each iteration requires two images each from the source and target domains. All experiments were conducted on a Nvidia GTX 1080 GPU to accelerate model training and prediction.

TABLE I
METRICS IN GUANGZHOU AND SHENZHEN

Study Area		**Guangzhou**		**Shenzhen**	
Strategy	***Model***	***F1(%)***	***IoU(%)***	***F1(%)***	***IoU(%)***
Direct Predicting	U-Net	52.93	35.99	47.28	31.27
	RefineNet	77.28	62.97	70.67	54.64
	DeepLab v3+	**82.18**	**70.09**	**78.04**	**63.99**
Maximum Mean Discrepancy	U-Net	54.52	37.48	50.32	33.62
	RefineNet	79.73	66.28	71.83	56.04
	DeepLab v3+	**83.13**	**71.13**	**80.57**	**67.46**
Domain Adaptation	U-Net	55.57	38.47	50.99	34.21
	RefineNet	86.65	70.42	73.67	58.31
	DeepLab v3+	**84.32**	**72.68**	**81.22**	**68.53**

C. Results and Analysis

The performance of each model is evaluated using F1-score and IoU. The most usual two metrics to evaluate the effect of the binary classification model are precision and recall. However, they can only reflect one aspect of the model. F1-score is the weighted average of precision and recall, considering both aspects of the model. IoU refers to the overlap rate between detection results and the ground truth.

Table I shows the qualitative results of our experiments obtained using different strategies. As can be seen from Table I, the accuracies of the transferred method were improved to a certain extent on both data sets. Take the results of Guangzhou data set for more specific analysis. When predicting without any transfer tricks, DeepLab v3+ gained

the highest F1 and IoU, i.e., 82.18% and 70.09%, respectively, which were increased to 83.13% and 71.13% after applying the MMD-based method, and 84.32% and 72.68% after the proposed adversarial domain adaptation method. RefineNet was slightly inferior to DeepLab v3+, but it showed great compatibility to both transferred methods. Compared with predicting directly, the F1 and IoU of RefineNet were increased by 2.45% and 3.31% with the MMD-based method, and they were significantly increased by 9.37% and 7.45% when using adversarial domain adaptation. Limited by the performance of the base model, U-Net-based models had lower accuracies. However, the IoU of U-Net was still improved by 1.04% after the MMD-based method, and the value reached 2.59% with the adversarial method. The results were very similar to Shenzhen data set, where there is a significant gain on evaluation metrics for all segmentation models after domain adaptation. Remarkably, compared to direct prediction, the highest F1 and IoU were improved from 78.04% and 63.99% to 81.22% and 68.53% after adversarial domain adaptation.

In summary, the fact that adversarial domain adaptation outperformed traditional MMD-based domain adaptation as well as direct prediction has well proved that adversarial domain adaptation is an effective way to overcome domain shift. It should also be noted that when adopting the same strategy, DeepLab v3+ obtained the highest accuracies among three semantic segmentation models, followed by RefineNet and U-Net.

Fig. 3 further demonstrates the visual segmentation results obtained by different segmentation models with respect to the ground truth. As Fig. 3 shows, adversarial domain adaptation can significantly improve the segmentation effect, which is mainly reflected in the fact that the results extracted by adversarial domain adaptation have the most complete morphology.

V. Conclusion

Domain adaptation provides a new effective way to improve the generalization and applicability of deep learning models. It shows significant potential in dealing with the domain shift caused by the spatial heterogeneity in different areas. In this letter, we employed a novel framework to map pixelwise UVs by integrating the domain adaptation strategy with several SOTA semantic segmentation models. The proposed method attempts to alleviate the domain shift of high-level representations through adversarial learning. Therefore, bidirectional domain adaptation was conducted in two fast-developing metropolises in PRD to validate the efficiency of the proposed model. The experimental results show that the domain adaption strategy helps to learn better representations for UV extraction in different cities without any labels in the target city. Furthermore, the results indicate that UVs are rather difficult to distinguish from other areas in a complex background.

References

[1] X. Liu *et al.*, "A future land use simulation model (FLUS) for simulating multiple land use scenarios by coupling human and natural effects," *Landscape Urban Planning*, vol. 168, pp. 94–116, Dec. 2017.

[2] X. Liu *et al.*, "High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth engine platform," *Remote Sens. Environ.*, vol. 209, pp. 227–239, May 2018.

[3] L. Li, Y. Chen, H. Gao, and D. Li, "Automatic recognition of village in remote sensing images by support vector machine using co-occurrence matrices," *Sensor Lett.*, vol. 10, nos. 1–2, pp. 523–528, 2012.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[5] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[6] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[7] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[8] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.

[9] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, 2019.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 7167–7176.

[11] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[12] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1306–1315.

[13] Q. Shi, B. Du, and L. Zhang, "Domain adaptation for remote sensing image classification: A low-rank reconstruction and instance weighting label propagation inspired algorithm," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5677–5689, Oct. 2015.

[14] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. CVPR*, 2017, pp. 2020–2030.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

[16] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018, pp. 801–818.

[17] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. ECCV*, 2016, pp. 102–118.

[18] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *Proc. ICML*, 2013, pp. 222–230.

[19] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *Proc. ICML*, 2013, pp. 819–827.

[20] M. Long, G. Ding, J. Wang, J. Sun, and P. S. Yu, "Transfer sparse coding for robust image representation," in *Proc. CVPR*, Jun. 2013, pp. 407–414.

[21] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[22] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. CVPR*, 2017, pp. 3722–3731.

[23] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," Dec. 2016, *arXiv:1612.02649*. [Online]. Available: https://arxiv.org/abs/1612.02649

[24] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. CVPR*, 2018, pp. 7472–7481.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[26] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. CVPR*, 2017, pp. 1925–1934.