UNSUPERVISED STYLE TRANSFER VIA DUALGAN FOR CROSS-DOMAIN AERIAL IMAGE CLASSIFICATION

Yansheng Li [1], *Te Shi* [1], *Wei Chen* [1], *Yongjun Zhang* [1], *Zhibin Wang* [2], *and Hao Li* [2]

[1]School of Remote Sensing and Information Engineering, Wuhan University, China
[2]Alibaba Group, China

ABSTRACT

Due to its wide applications, aerial image classification, which is also called semantic segmentation of aerial imagery, attracts increasing research interest in recent years. Until now, deep semantic segmentation network (DSSN) has been widely adopted to address aerial image classification and achieves tremendous success. However, the superior performance of DSSN highly depends on massive targeted data with labels. When DSSN is trained on data from the source domain but tested on data from the target domain, the performance of DSSN is often very limited due to the data shift between source and target domains. To alleviate the disadvantage influence of data shift, this paper proposes a domain adaptation approach via unsupervised style transfer to cope with cross-domain aerial image classification. More specifically, this paper innovatively recommends DualGAN to conduct unsupervised style transfer for mapping aerial images in the source domain to the target domain. The mapped aerial imagery with labels is adopted to train DSSN, which is further used to classify aerial imagery in the target domain. To verify the validity of the presented approach, we give two cross-domain experimental settings including: (I) variation of geographic location; (II) variation of both geographic location and imaging mode. Extensive experiments under two typical cross-domain settings show that our proposed method can obviously outperform the state-of-the-art methods.

Index Terms— Cross-domain aerial image classification, domain adaptation, unsupervised style transfer, DualGAN.

1. INTRODUCTION

Due to its wide usage in disaster rescue, crop assessment, intelligent traffic, and so forth, aerial image classification attracts more and more research interest. More specifically, aerial image classification works for assigning a land-cover type from a predefined set (e.g., building, car, tree, and so on) to each pixel in the image. Although it has been widely exploited, aerial image classification is still an open problem and needs much more exploration around how to decrease the supervison dependency of labeled data.

As is well known, deep semantic segmentation network (DSSN) has achieved tremendous success in aerial image classification under the basic premise that a targeted aerial image dataset with accurate labels is available. However, if DSSN is trained on labeled images from dataset A (i.e., source domain), but is directly deployed to classify the images from dataset B (i.e., target domain), the classification performance often dramatically decreases due to the domain shift between source and target domains. One naive solution is to annotate the images in the target domain and train DSSN on it. In reality, collecting a large-scale aerial image dataset with pixel-wise annotations is time-consuming and expensive. For example, pixel-wise annotation of the natural Cityscapes image will take 90 minutes on average [1]. Compared with natural images, aerial images present more complex structure. So it can be imagined that the labeling process is much more difficult than that of natural images. One potential solution is to train a deep model with the existing labeled data from the source domain, and then try to transfer the model to the data from the target domain.

In the field of computer vision, there have been many methods for cross-domain image classification. Tasi et al. [2] proposed an adversarial learning method for domain adaptation in the context of semantic segmentation. They adopt adversarial learning in the output space and improve the performance. Yonghao et al. [3] utilized the self-ensembling attention network to extract attention-aware features for domain adaptation. On the surface, aerial image classification is similar to natural image classification. Actually, compared with natural image classification, aerial image classification has to address many further challenges because of the dense structure and arbitrary orientation of geospatial objects. Hence, domain adaptation for aerial image classification deserves much more special exploration.

In the remote sensing community, the pioneers in [4] proposed an unsupervised domain adaptation method to address cross-domain aerial image classification where CycleGAN is

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program; the National Natural Science Foundation of China under grants 41971284 and 41601352; the China Postdoctoral Science Foundation under grants 2016M590716 and 2017T100581; the Hubei Provincial Natural Science Foundation of China under grant 2018CFB501. (Corresponding author: Yansheng Li)

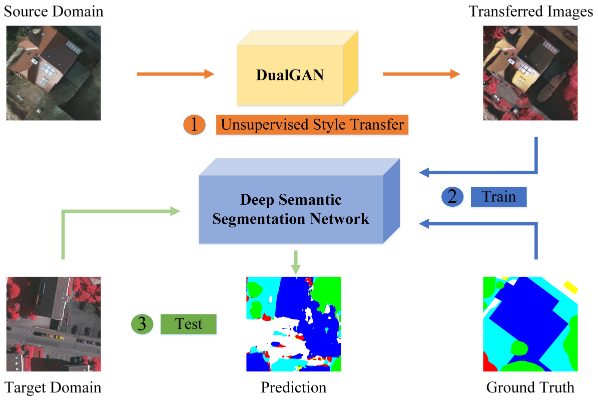


Fig. 1: Flowchart of the proposed UST-DG.

utilized in the domain adaptation process. To address cross-domain aerial image classification, we propose an unsupervised style transfer approach via DualGAN (UST-DG). In our proposed UST-DG, DualGAN is innovatively recommended to conduct unsupervised style stransfer due to its transfer superiority compared with CycleGAN. More specifically, we firstly deploy DualGAN [5] to conduct unsupervised style transfer from the source aerial image dataset to the target aerial image dataset. Then, train a DSSN on the transferred arial image dataset with labels. The trained DSSN using the transferred images can fluently understand the images from the target aerial image dataset.

To verify the validity of the presented approach, we give two typical cross-domain experimental settings including: (I) variation of geographic location; (II) variation of both geographic location and imaging mode. Extensive experiments show that our proposed UST-DG can outperform the state-of-the-art methods, remarkably.

2. METHODOLOGY

In this section, we describe the problem setting and show the framework of our UST-DG algorithm. Specifically, we introduce the DualGAN which is adopted to perform unsupervised style transfer, and the DeepLab v3 plus [6] which is a typical DSSN. After training the DSSN with the transferred dataset, the DSSN can be applicable to work on the target dataset.

In the unsupervised domain adaptation task, a well annotated dataset S from the source domain and unlabeled dataset T from target domain are given. The proposed method aims to use the paired source domain images to train a model and then apply it to predict the label for the target dataset. The unsupervised image to image style transfer procedure done by DualGAN is designed to make images of the source domain mimic the style of the target domain, which will reduce the domain shift between the source images and target images. The flowchart of the framework is depicted in Fig. 1. Our proposed algorithm consists of three steps. The first step is to train a DualGAN network and transfer the source domain images to the style of the target domain, whose output is a transferred dataset conserves the structures representation of the source dataset but simulates the global style of the target dataset. The second step is to train a DSSN with the translated dataset associated with the source labels. This step helps the model learn the patterns of the target dataset and converge to a better generalization ability of image structure on the target dataset. Finally, the DSSN is capable of working on the target dataset.

2.1. Unsupervised style transfer via DualGAN

DualGAN employs two GANs, the primal GAN $\{G_A, D_A\}$ and a dual GAN $\{G_B, D_B\}$, which map a sample from the source (target) domain to the target (source) domain and generate samples that are indistinguishable from samples in the target (source) domain, respectively.

As shown in Fig. 2, image $s \in S$ is converted to domain T by G_A. Then, D_A is used to measure how well the translation $G_A(s, z)$ fits in T, where z is random noise to perform data augmentation and so is z'. $G_A(s, z)$ is then converted back to domain S by G_B, which outputs $G_B(G_A(s, z), z')$ as the reconstruction of s. Similarly, $t \in T$ is translated to S as $G_B(t, z')$ and then reconstructed as $G_A(G_B(t, z'), z)$. The discriminator D_A is trained with t as positive samples and $G_A(s, z)$ as negative examples, which means it give samples from t a high score but gives samples from $G_A(s, z)$ a low score. Meanwhile, D_B is trained in the same way. Generators G_A and G_B are optimized to emulate "fake" outputs to confuse the corresponding discriminators D_A and D_B, as well as to minimize the reconstruction losses $||s - G_A(G_B(t, z'), z)||$ and $||t - G_B(G_A(s, z), z')||$.

The corresponding loss functions used in D_A and D_B are defined as:

$$l^d_A(s, t) = D_A(G_A(s, z)) - D_A(t) \tag{1}$$

$$l^d_B(s, t) = D_B(G_B(t, z')) - D_B(s) \tag{2}$$

where $s \in S$ and $t \in T$.

The same loss function is used for both generators G_A and G_B as they share the same objective.

$$\begin{aligned} l^g(s, t) = {} & \lambda_S||s - G_A(G_B(t, z'), z)|| \\ & + \lambda_T||t - G_B(G_A(s, z), z')|| \\ & - D_B(G_B(t, z')) - D_A(G_A(s, z)) \end{aligned} \tag{3}$$

where $s \in S$, $t \in T$ and λ_S, λ_T are two constant parameters depending on the specific task.

2.2. Learning a deep semantic segmentation network

Based on the aforementioned unsupervised DualGAN in section 2.1, the images in the source aerial image dataset are automatically transferred to approximate the style of the target aerial image dataset, which benefits minimizing the influence

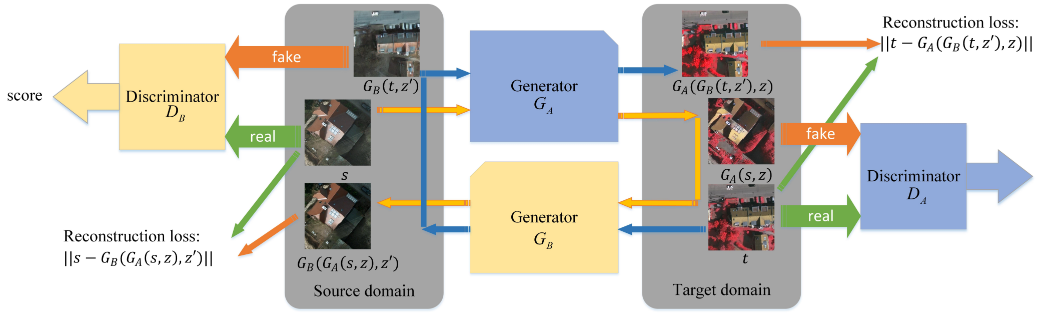


Fig. 2: Architecture of DualGAN for unsupervised style transfer.

of data shift between different domains. In addition, we use the transferred images with labels to train the DSSN. In our implementation, DSSN is implemented by DeepLab v3 plus as DeepLab v3 plus is the state-of-the-art semantic segmentation network and often achieves the best performance in the natural semantic segmentation field.

2.3. Classifying the images from the target domain

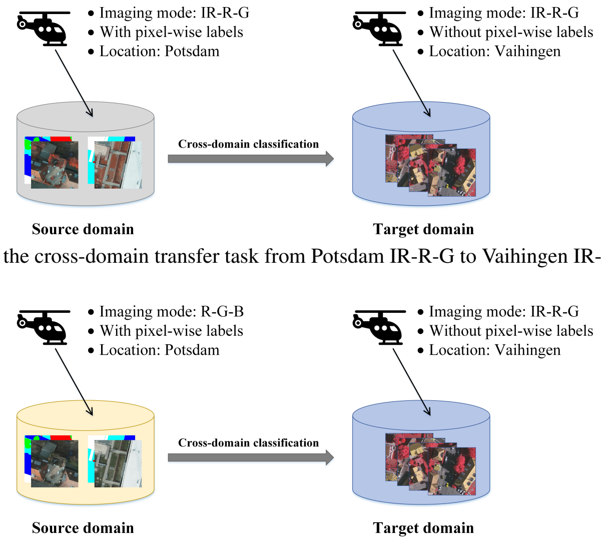
The trained DSSN, in section 2.2, is utilized to classify images from the target aerial image dataset. As the trained data (i.e., the transferred images with labels) is highly similar to the data from the target domain, the trained DSSN can can fluently understand the images from the target image dataset. More experiments can refer to the experimental section.

3. EXPERIMENTAL RESULTS

3.1. Experimental settings and evaluation metrics

To verify our methodology, we conduct experiments by Potsdam and Vaihingen datasets which belong to the ISPRS 2D semantic segmentation benchmark dataset [7]. All images in both datasets are provided with their semantic labels, including six classes of ground objects: building, tree, car, impervious surfaces, low vegetation, and clutter/background. The Potsdam dataset contain 3 different imaging modes: IRRG: 3 channels (IR-R-G), RGB: 3 channels (R-G-B), RGBIR: 4 channels (R-G-B-IR), we use the first two kinds. The Vaihingen dataset contains only one imaging mode: IRRG: 3 channels (IR-R-G). To lift the computational efficiency, we crop the images and their corresponding labels into patches with a size of 512 × 512 and feed them into the network.

In details, we give two cross-domain experimental settings including: (I) variation of geographic location, shown as Fig. 3(a); (II) variation of both geographic location and imaging mode, shown as Fig. 3(b). Similar to [2, 3, 4], we use $accuracy$, $precision$, $recall$, $F1 - score$ and $mIoU$ to evaluate the performance of these models.



(a) the cross-domain transfer task from Potsdam IR-R-G to Vaihingen IR-R-G.

(b) the cross-domain transfer task from Potsdam R-G-B to Vaihingen IR-R-G.

Fig. 3: Two different cross-domain classification tasks.

3.2. Comparison results with the state-of-the-art methods

3.2.1. Experimental results under the variation of geographic location

To confirm the effectiveness of our proposed UST-DG on domain shift mainly caused by region variation, we use Potsdam IR-R-G dataset as source domain and Vaihingen IR-R-G dataset serves as target domain. The metrics of cross-domain classification results are shown in Table 1, where methods a to c are based on BiSeNet [8] framework and methods e to f are based on DeepLab framework. The visualization of the classification results is shown in Fig. 4. Through experiments, we can find that domain shift has a great impact on the accuracy of the model. It is shown that our proposed method obtains higher performance than other methods.

3.2.2. Experimental results under the variation of both geographic location and imaging mode

Furthermore, Potsdam R-G-B dataset serves as source domain and Vaihingen IR-R-G dataset serves as target domain in order to evaluate the effectiveness of our method on domain shift caused by variation of both geographic location and imaging mode. The metrics of cross-domain classification results are shown in Table 2. The visualization of the

Table 1: The cross-domain classification results from Potsdam IR-R-G to Vaihingen IR-R-G.

Method	Method id	Accuracy	Precision	Recall	F1-score	mIoU
BiSeNet without adaptation	*a*	0.518	0.501	0.454	0.438	0.245
UDA in [4]	*b*	0.326	0.177	0.179	0.155	0.092
BiSeNet + DualGAN	*c*	0.548	0.485	0.475	0.445	**0.279**
DeepLab v3 plus without adaptation	*d*	0.404	0.473	0.510	0.491	0.253
SEANet in [3]	*e*	0.612	0.552	0.562	0.557	0.377
Our proposed UST-DG	*f*	0.661	0.579	0.635	0.606	**0.416**

Table 2: The cross-domain classification results from Potsdam R-G-B to Vaihingen IR-R-G.

Method	Method id	Accuracy	Precision	Recall	F1-score	mIoU
BiSeNet without adaptation	*a*	0.415	0.311	0.325	0.287	0.167
UDA in [4]	*b*	0.456	0.448	0.429	0.401	0.261
BiSeNet + DualGAN	*c*	0.543	0.474	0.474	0.439	**0.283**
DeepLab v3 plus without adaptation	*d*	0.367	0.495	0.410	0.449	0.245
SEANet in [3]	*e*	0.481	0.428	0.517	0.468	0.278
Our proposed UST-DG	*f*	0.602	0.504	0.513	0.509	**0.359**

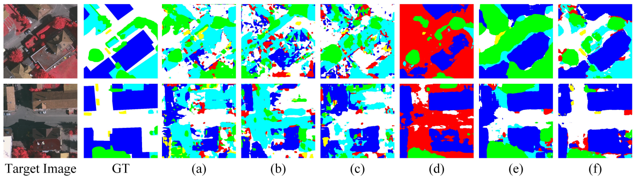


Fig. 4: Samples of classification results from Potsdam IR-R-G to Vaihingen IR-R-G.

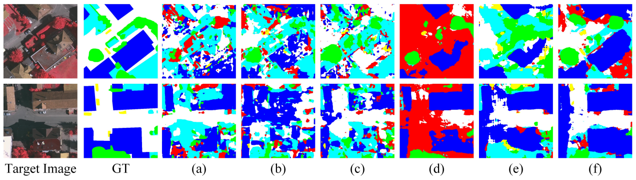


Fig. 5: Samples of classification results from Potsdam R-G-B to Vaihingen IR-R-G.

classification results is shown in Fig. 5. The experimental results are similar to the above, our proposed UST-DG gains a higher performance, which further proves the effectiveness of our proposed model.

To sum up, our proposed UST-DG has a good performance in dealing with both the domain shift mainly caused by the region variation and caused by the imaging mode variation. Our method shows strong robustness and great generalization capability.

4. CONCLUSION

In this work, we innovatively apply DualGAN to do style transfer with source dataset to the target dataset for unsupervised domain adaptation. Our proposed UST-DG method does not affect the ability of the segmentation model to classify classes not affected by domain shift. In addition, it costs very little because it does not require annotating data or other manual work. To verify our proposed approach, we give two cross-domain experimental settings including: (I) variation of geographic location; (II) variation of both geographic location and imaging mode. Extensive experiments under two typical cross-domain settings show that our proposed method can obviously outperform the state-of-the-art methods.

5. REFERENCES

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[2] 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*, 2018, pp. 7472–7481.

[3] Yonghao Xu, Bo Du, Lefei Zhang, Qian Zhang, Guoli Wang, and Liangpei Zhang, "Self-ensembling attention networks: Addressing domain shift for semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 5581–5588.

[4] Bilel Benjdira, Yakoub Bazi, Anis Koubaa, and Kais Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sensing*, vol. 11, no. 11, pp. 1369, 2019.

[5] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[7] Markus Gerke, "Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen)," 2014.

[8] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.