ROTATION CONSISTENCY-PRESERVED GENERATIVE ADVERSARIAL NETWORKS FOR CROSS-DOMAIN AERIAL IMAGE SEMANTIC SEGMENTATION

Te Shi, *Student Member, IEEE,* Yansheng Li*, *Member, IEEE,* Yongjun Zhang*, *Member, IEEE*

School of Remote Sensing and Information Engineering, Wuhan University, China

IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554606

ABSTRACT

Due to its wide applications, aerial image semantic segmentation attracts increasing research interest in recent years. As well known, deep semantic segmentation network (DSSN) has been widely used to deal with aerial image segmentation and achieves spectacular success. However, when applying the DSSN trained with the labeled aerial images (i.e., the source domain) to predict the aerial images acquired with different acquisition conditions (i.e., the target domain), the performance often dramatically degrades. To alleviate the negative influence of cross-domain data shift, this paper proposes a domain adaptation approach to deal with cross-domain aerial image semantic segmentation. More precisely, this paper proposes a novel rotation consistency-preserved generative adversarial network (RCP-GAN) to carry out domain adaptation for mapping aerial images in the source domain to the target domain. Furthermore, the mapped aerial imageries with labels are used 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 effectively address the domain shift problem and outperform the state-of-the-art methods with a large margin.

***Index Terms*—** Rotation consistency-preserved generative adversarial network (RCP-GAN), cross-domain aerial image semantic segmentation, domain adaptation, unsupervised style transfer.

1. INTRODUCTION

Compared with satellite images, aerial images often provide sufficient information about the ground objects. Pixel-level classification (i.e., semantic segmentation) of aerial images is a fundamental research task in the remote sensing community that has great significance in infrastructure planning, landcover classification, and urban object detection.

This work was supported by the National Natural Science Foundation of China under grants 42030102 and 41971284. (Corresponding authors: Yansheng Li and Yongjun Zhang)

As well known, deep semantic segmentation network (DSSN) has achieved spectacular breakthroughs in aerial image semantic segmentation [1]. However, its superior performance highly depends on the abundant labeled samples. As aerial images present massive complex structures, it is time-consuming and laborious to annotate pixel-level labels for oversized aerial images. To alleviate the annotation labor, one may borrow some similar data (not the target data itself) with labels to train deep network. However, when applying the deep learning model trained with the labeled aerial images from the source domain to predict the aerial images acquired with different acquisition conditions (i.e., the target domain), the performance often critically degrade. This phenomenon is called domain shift that is caused by the distribution gap between different aerial image domains. Distribution gaps between different aerial datasets are mainly due to the diverse data acquisition conditions including imaging sensors, varied geospatial regions, ground sampling distances and arbitrary shooting angles, the images often present many distinct characteristics such as variety of imaging mode, multi-scale of objects and variety of color saturation. Varied data acquired regions may result in significant distinctions in architectural style and urban layout between data sets. Thus, a reasonable solution is to perform style transfer, i.e., to transfer the images from the source domain to the style of the target domain. Then, the transferred images with the corresponding labels of images from the source domain are used to train the DSSN. As a consequence, this strategy naturally reduces the effect of domain shift.

The existing style transfer algorithms are mainly constructed based on generative adversarial networks (GANs) and can be roughly divided into two categories: methods that require paired images [2, 3] and methods that do not require paired images [4, 5]. For the former, the strictly paired images are taken as one kind of supervision constraint. For instance, conditional GAN [3], which requires the pairs of corresponding images, is first proposed to address image-to-image translation. But this kind of framework does not harmonize with the goal of generalizing cross-domain aerial image semantic segmentation. The latter methods are not conditioned on the paired images, which is conducive to promoting the cross-domain aerial image semantic segmentation task. However, these methods do not consider the distinct

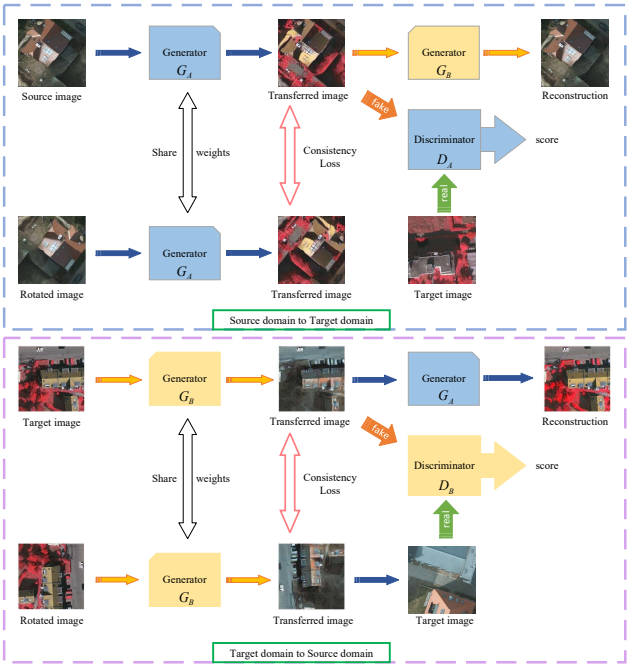

Fig. 1: Our proposed RCP-GAN.

characteristics in aerial images as aforementioned. With this consideration, we propose rotation consistency-preserved generative adversarial network (RCP-GAN) to carry out unsupervised style transfer of aerial images for cross-domain aerial image semantic segmentation. The main contributions of this paper include the following two aspects:

- By fully exploiting the unsupervised constraints, this paper proposes a novel RCP-GAN model for unsupervised style transfer of aerial images to mitigate the effect of domain shift.
- As a flexible framework, the proposed RCP-GAN can be easily extended to address kinds of cross-domain aerial image semantic segmentation cases including: (I) variation of geographic location; (II) variation of both geographic location and imaging mode.

The rest of this paper is organized as follows. Section 2 depicts the methodology. Section 3 specifically gives the experimental results. Section 4 gives a conclusion of this paper.

2. METHODOLOGY

In this section, we first describe the problem to be coped with. Next, we show our proposed RCP-GAN model which is an unsupervised way to do style transfer. Then, DSSN is trained with the transferred image dataset.

To facilitate clarifying the methodology, we first formulate the involved data. Let S denote the source dataset and T denote the target dataset. 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 RCP-GAN is designed to make images of the source domain mimic the style of the target domain.

2.1. RCP-GAN for unsupervised style transfer of aerial images

Similarly, our proposed RCP-GAN employs two GANs, the primal GAN $\{G_A, D_A\}$ and a dual GAN $\{G_B, D_B\}$. As shown in Fig. 1, 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:

$$l_A^d(s,t) = D_A(G_A(s,z)) - D_A(t) \quad (1)$$

$$l_B^d(s,t) = D_B(G_B(t,z')) - D_B(s) \quad (2)$$

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

Considering the distinct characteristic of arbitrary shooting angles in aerial images, the rotation consistency constraint is introduced into our GAN model. Specifically, rotation transformation φ (random rotation of 90 degrees, 180 degrees, 270 degrees) is performed on the image s and t, and then we obtain $s' = \varphi(s)$ and $t' = \varphi(t)$. s' is fed into G_A and the output $G_A(s', z)$ is obtained. To compute the pixel-level consistency of two outputs, we have to perform the inverse transform to put every pixel to the original location. We denote inverse transforms of the random rotation as φ^{-1}. Thus, we can obtain the inverse transformed outputs $\overline{G_A(s', z)} = \varphi^{-1}(G_A(s', z))$, and the rotation consistency loss can be computed. The consistency loss term often uses the mean squared error, which encourages the pixel-level consistency of the output under different random rotation transforms. The loss function can be described as Eq. (3). The rotation consistency loss of unsupervised style transfer from target domain to source domain is similar to this.

$$l_{\text{con}}^{S \to T} = \ell_{MSE}\left(G_A(s', z), \overline{G_A(s', z)}\right) \quad (3)$$

where $s \in S$ and $t \in T$. ℓ_{MSE} represents mean squared error loss.

So, the total loss function can be defined as Eq. (4).

$$\begin{aligned} l^g(s,t) &= \lambda_S||s - G_A(G_B(t,z'),z)|| \\ &+ \lambda_T||t - G_B(G_A(s,z),z')|| \\ &+ l_{con}^{S\to T} + l_{con}^{T\to S} \\ &- D_B(G_B(t,z')) - D_A(G_A(s,z)) \end{aligned} \tag{4}$$

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

2.2. Cross-domain aerial image semantic segmentation with RCP-GAN

After transferring the source dataset to style of the target dataset by RCP-GAN in section 2.1, we use the transferred dataset with origin labels to train a DSSN, where we employ Deeplab v3+ in our implementation. Finally, the DSSN can work well on the target domain to do semantic segmentation.

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. 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 $\times$ 512 and feed them into the network.

In details, we give two cross-domain experimental settings including: (I) variation of geographic location; (II) variation of both geographic location and imaging mode. We use $accuracy$, $precision$, $recall$, $F1-score$ and $mIoU$ to evaluate the performance of these models.

3.2. Implementation details

Our proposed RCP-GAN is constructed with the same network architecture for G_A and G_B. The generator is configured with equal number of down sampling and up sampling layers. In addition, we configure the generator with skip connections between mirrored down sampling and up sampling layers, making it a U-shaped net. For discriminators, we employ the Patch-GAN architecture. The patch size at which the discriminator operates is fixed at 70×70, and the image resolutions were mostly 256×256, same as pix2pix [3].

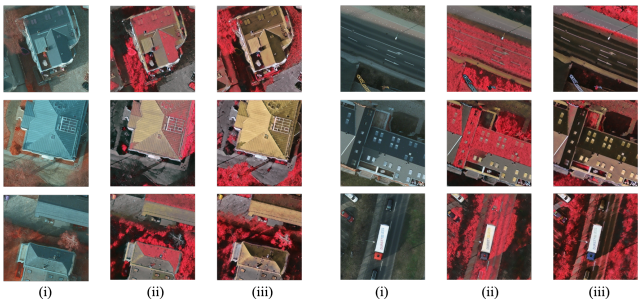

Fig. 2: Unsupervised style transfer results. (a) From Potsdam IR-R-G to Vaihingen IR-R-G. (b) From Potsdam R-G-B to Vaihingen IR-R-G.

This work is implemented by Pytorch and trained on a single Nvidia TITAN RTX GPU with 24GB RAM. As an optimizer for the training, we used RMSprop optimizer with the initial learning rate set to 0.00005 and weight decay for RMSProp optimizer is set to 0.1. In our implementation, the batch size and epoch are set to 4 and 45, respectively. As for the DSSN, we use Deeplab v3+ and adopt the default settings.

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

3.3.1. Experimental results under the variation of geographic location

To confirm the effectiveness of our proposed RCP-GAN for aerial image unsupervised style transfer, we use Potsdam IR-R-G dataset as source domain and Vaihingen IR-R-G dataset serves as target domain. In addition, we compared RCP-GAN with DualGAN. The unsupervised style transfer results are shown in Fig. 2(a), where (i) are source images, (ii) and (iii) are transferred via DualGAN and our RCP-GAN, respectively. Further, we carry out the compariosn experiments with the relative cross-domain aerial images semantic segmentation methods. The Metrics of cross-domain semantic segmentation results are shown in Table 1. The visualization of the semantic segmentation results is shown in Fig. 3. 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.

3.3.2. 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. Fig. 2(b) shows the unsupervised style transfer results. The Metrics of cross-domain semantic segmentation results are shown in Table 2. The visualization of the semantic segmentation results is shown in Fig. 4. The experimental results are similar to the above, our method gains a higher performance, which further proves the effectiveness of our proposed model.

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
Deeplab v3+ without adaptation	b	0.404	0.473	0.510	0.491	0.253
SEANet [6]	c	0.612	0.552	0.562	0.557	0.377
AdaptSegNet [7]	d	0.596	0.552	0.534	0.523	0.352
UDA in [8]	e	0.326	0.177	0.179	0.155	0.092
Deeplab v3+ with DualGAN	f	0.661	0.579	0.635	0.606	0.416
Our RCP-GAN	g	**0.720**	**0.612**	**0.719**	**0.661**	**0.482**

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
Deeplab v3+ without adaptation	b	0.367	0.495	0.410	0.449	0.245
SEANet [6]	c	0.481	0.428	0.517	0.468	0.278
AdaptSegNet [7]	d	0.594	0.524	0.460	0.490	0.321
UDA in [8]	e	0.456	0.448	0.429	0.401	0.261
Deeplab v3+ with DualGAN	f	0.602	0.504	0.513	0.509	0.359
Our RCP-GAN	g	**0.683**	**0.538**	**0.586**	**0.561**	**0.407**

To sum up, our proposed RCP-GAN 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.

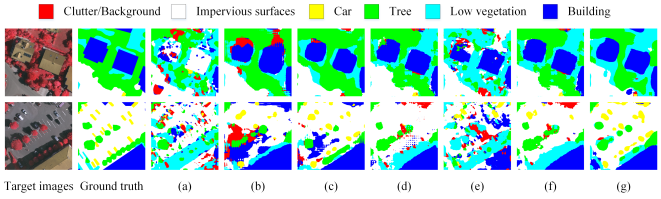

Fig. 3: Samples of semantic segmentation results on Potsdam IR-R-G to Vaihingen IR-R-G.

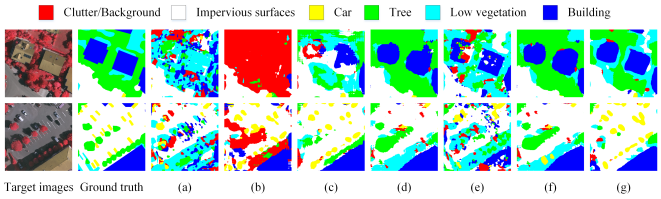

Fig. 4: Samples of semantic segmentation results on Potsdam R-G-B to Vaihingen IR-R-G.

4. CONCLUSION

To solve the negative influence domain shift in cross-domain semantic segmentation, this paper proposes a novel RCP-GAN to carry out unsupervised style transfer for mapping aerial images in the source domain to the target domain. Then, the mapped aerial imageries with labels are used to train DSSN, which is further used to classify aerial imagery in the target domain. Our proposed method can effectively address the domain shift problem in cross-domain semantic segmentation. In addition, it costs very little because it does not require extra 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 well address the domain shift problem in cross-domain aerial image semantic segmentation. In future work, we will improve our GAN model by considering more characteristics of aerial images.

5. REFERENCES

[1] Michele Volpi and Devis Tuia, "Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images," *ISPRS journal of photogrammetry and remote sensing*, vol. 144, pp. 48–60, 2018.

[2] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin, "Image analogies," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 327–340.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *CVPR*, 2017, pp. 2223–2232.

[5] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *ICCV*, 2017, pp. 2849–2857.

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

[7] 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 *CVPR*, 2018, pp. 7472–7481.

[8] 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.