Depth-Assisted ResiDualGAN for Cross-Domain Aerial Images Semantic Segmentation

Yang Zhao®, Peng Guo, Han Gao, and Xiuwan Chen

Abstract—Unsupervised domain adaptation (UDA) is an approach to minimizing the domain gap. Generative methods are common approaches to minimizing the domain gap of aerial images, which improves the performance of the downstream tasks, for example, cross-domain semantic segmentation. For aerial images, the digital surface model (DSM) is usually available in both the source domain and the target domain. Depth information in DSM brings external information to generative models. However, little research utilizes it. In this letter, depth-assisted ResiDualGAN (DRDG) is proposed where depth supervised loss (DSL) and depth cycle consistency loss (DCCL) are used to bring depth information into the generative model. Experimental results show that DRDG reaches state-of-the-art accuracy between generative methods in cross-domain semantic segmentation tasks. Source code is available at https://github.com/miemiemyanga/ResiDualGAN-DRDG.

Index Terms—Aerial images, semantic segmentation, unsupervised domain adaptation (UDA).

I. INTRODUCTION

SEMANTIC segmentation is a classification task that gives a category to every pixel of an image. A pixel-level annotated dataset is imperative for training a deep-learning semantic segmentation model. However, the pixel-level annotation of images is laborious, time-consuming, and expensive [1], [2]. Most of the datasets used in the practical operation are unannotated, which is impossible for training a deep-learning model. A common idea is to train a model from annotated datasets and then apply it to the unannotated dataset. Nevertheless, because of the domain gap resulting from distinct data distribution between different datasets, the performance of the model trained in an annotated dataset will greatly decline when transferred to another dataset [3]. For remote-sensing (RS) images, the problem of the domain gap is significantly magnified because of various imaging sensors, imaging time, weather conditions, image resolution, and so on [1], [4].

Unsupervised domain adaptation (UDA) provides an approach for minimizing the impact of the domain gap, making it possible to transfer a deep-learning model to an unannotated dataset. Methods of UDA for semantic segmentation can be roughly divided into several categories: adversarial generative methods [3], adversarial discriminative methods [5], and self-training methods [6]. For RS images, adversarial generative methods, which are aimed to generate new annotated images based on generative adversarial networks (GANs), show significant advantages compared with other methods [2], [4], [7]. ResiDualGAN (RDG) [4] is an adversarial generative method for UDA of RS images. Focusing on the unique feature of scale discrepancy and real-to-real image translation of RS images, RDG reaches state-of-the-art performance in cross-domain semantic segmentation. In this letter, RDG is selected as the foundation of our model.

Depth information is utilized for improving the effect of UDA in some recent computer vision research. The motivation is to utilize the depth information as an external assistant that gives the deep-learning model other information to improve the performance. Several works have been done by combining deep information with the generative method [8], adversarial discriminative method [9], [10], and self-training method [11], which demonstrate the effectiveness of depth information in UDA. Though improvements have been made, the defect of these methods is still nonnegligible. Depth information is usually inaccessible for the street view datasets used in these methods. An additional depth estimation model is used to give an approximate depth estimation for the street view image [11], which is both inaccurate and compute-intensive.

The digital surface model (DSM) is an elevation model that reflects the depth information of the ground surface. DSM can be easily obtained when producing an aerial image dataset by the photogrammetric method. As a result, utilizing DSM as assisted information for UDA of aerial images is reasonable. However, little research has been made on utilizing DSM to assist the UDA procedure of aerial images. Wittich and
Rottensteiner [12] work regards the DSM as an additional channel of input images. At the training phase, the concatenated images and DSMs are fed forward into the adversarial model and then conduct the normal UDA procedure. Particularly, at the testing phase, the segmentation model got by Wittich’s method also needs an additional DSM band as an input when deployed on devices. The requirement of additional DSMs at test time constricts the application scope of the segmentation model such as real-time segmentation when we cannot get the DSM instantly from the optical camera.

In this work, we adapt the depth information to the UDA task of semantic segmentation of aerial images Fig. 1. Depth-assisted RDG (DRDG) is proposed by combining DSM with RDG where the depth supervised loss (DSL) and depth cycle consistency loss (DCCL) are used to force RDG to learn external information. Our final segmentation model needs only optical images as inputs at test time and is better in performance compared to other methods. The experimental results show that DRDG reaches state-of-the-art performance in open datasets. Our contribution can be summarized as follows.

1) A new GAN-based model, DRDG, is proposed for the UDA task of semantic segmentation of aerial images. Experimental results show that DRDG is the state-of-the-art adversarial generative UDA method for aerial images in open datasets.
2) DSM, which can be easily accessed for aerial images, is introduced to the DRDG as assisted information to improve the performance of the generative model, which illustrates the effectiveness of DSM for UDA of aerial images.

II. METHODOLOGY

Consider a source domain $S$ and a target domain $T$, where only images in the $S$ are annotated, while DSM can be obtained in both $S$ and $T$. Formally, we define $X_S \in \mathbb{R}^{H_S \times W_S \times 3}$ as the image set of $S$, $Y_S \in \{1, \ldots, C\}^{H_S \times W_S}$ is the corresponding semantic segmentation label set, where $C$ is the number of categories, $Z_S \in \{Z_{\text{min}}, Z_{\text{max}}\}^{H_S \times W_S \times 1}$ is the DSM image set. Similarly, the image set in $T$ can be defined as $X_T \in \mathbb{R}^{H_T \times W_T \times 3}$, $Z_T \in \{Z_{\text{min}}, Z_{\text{max}}\}^{H_T \times W_T \times 1}$ is the DSM image set of $T$. Labels are inaccessible in $T$.

The objective of the proposed method is that, given the annotated $S$ and unannotated $T$, train a model $f_T$ that performs a semantic segmentation task on $T$. To this end, DRDG is designed to generate target-domain stylized images $x_S \rightarrow T$ from the source-domain image $x_S$ that is annotated (Fig. 2). In other words, DRDG minimizes the pixel-level domain gap between the source domain and the target domain. Specifically, depth information is involved in DRDG by combining two constraints, DSL and DCCL, to further improve the performance of RDG. In this section, we will introduce RDG first and then describe the DSL and DCCL. Finally, a semantic segmentation model $f_T$ can be trained on the annotated target-domain stylized image $x_S \rightarrow T$.

A. ResiDualGAN

The RDG is an adversarial generative method for UDA of RS images, which shows prominent performance on cross-domain RS image semantic segmentation tasks. The function of RDG is to perform image-to-image translation that, in our case, transfers $x_S$ to the style of $x_T$ which minimizes the discrepancy between $S$ and $T$. Given $x_S$, $x_S \rightarrow T$ can be obtained by the generator of RDG, $\text{ResiG}_S \rightarrow T$. The procedure can be written as follows:

$$x_S \rightarrow T = \text{ResiG}_S \rightarrow T(x_S) = \text{Resize}_{S \rightarrow T}(D_{\text{resi}}^{S \rightarrow T}(E_{S \rightarrow T}(x_S)) + x_S) \quad (1)$$

where $\text{Resize}_{S \rightarrow T} : \mathbb{R}^{H_S \times W_S \times 3} \rightarrow \mathbb{R}^{H_T \times W_T \times 3}$ is a resize function that resizes images of the source domain to the size of the
target domain, implemented as a bilinear interpolation in this letter, $E_{S \rightarrow T}$ is the encoder part of ResiG$_{S \rightarrow T}$ and $D^\text{resi}_{S \rightarrow T}$ is the decoder.

$x_{S \rightarrow T}$ is then passed through a discriminator $D_T$, which is designed to try to distinguish whether images are generated by the generator or the initial images of the target domain. On the contrary, the objective of the generator is to generate target-stylized images to try to fool the discriminator. The opposite optimization direction between the generator and discriminator brings about an adversarial loss $L_{\text{adv}}(S, T)$. In RDG, the adversarial loss can be written as follows:

$$L_{\text{adv}}(S, T) = \|E_{S \rightarrow T}(D_T(x_T)) - E_{S \rightarrow X_S}(D_T(ResiG_{S \rightarrow T}(x_S)))\|.$$  

(2)

ResiG$_{T \rightarrow S}$, and $D_S$ is the inversion of the procedure above where another generator ResiG$_{T \rightarrow S}$ is implemented to reconstruct $x_{S \rightarrow T}$ to the $x_S$, and $D_S$ is used to differentiate whether images are generated by ResiG$_{T \rightarrow S}$ or original images in the source domain. The reconstruction of $x_S$ will bring into a reconstruction loss $L_{\text{cyc}}$ represented as follows:

$$L_{\text{cyc}}(S, T) = \|E_{S \rightarrow X_S}(ResiG_{T \rightarrow S}(ResiG_{S \rightarrow T}(x_S))) - x_S\|_1.$$  

(3)

where $\| \cdot \|_1$ is the L1 normalization.

B. Depth-Assisted RDG

Based on RDG, in this letter, DSL and DCCS are involved in the generators DRDG$_{S \rightarrow T}$ and DRDG$_{T \rightarrow S}$ to constrain the image translation procedure. The function of DRDG$_{S \rightarrow T}$ is similar to ResiG$_{S \rightarrow T}$, which is designed to perform image translation between domains. The difference is that DSM is utilized as assisted information to improve the depth correctness of image translation.

1) Depth Supervised Loss: DSL is a regression loss of DSM. The DSL network shares the encoder $E_{S \rightarrow T}$ with RDG generator ResiG$_{S \rightarrow T}$ and utilizes another decoder $D^\text{resi}_{S \rightarrow T}$ to generate a DSM prediction $\hat{z}_S = D^\text{resi}_{S \rightarrow T}(E_{S \rightarrow T}(x_S))$. Following [11], a Berhu loss, which is a reverse version of Huber loss, is utilized as the measure between the prediction $\hat{z}_S$ and real DSM $z_S$:

$$L_{\text{DSL}}(S) = E_{\{z_S, \hat{z}_S\}}(\text{Berhu}(z_S, \hat{z}_S))$$  

(4)

where $\hat{Z}_S$ is the predicted depth image set of $S$. The Berhu loss is

$$\text{Berhu}(z_S, \hat{z}_S) = \begin{cases} \frac{|\hat{z}_S - z_S|}{L}, & |\hat{z}_S - z_S| \leq L \\ (\hat{z}_S - z_S)^2 + L^2, & |\hat{z}_S - z_S| > L \end{cases}$$  

(5)

where $L = 0.2 \max(|\hat{z}_S - z_S|)$.

2) Depth Cycle Consistency Loss: In the process of image translation using RDG without depth information, only a cycle loss is utilized to constrain the pixel-level cycle consistency between $S$ and $T$. As a result, images may be incorrectly translated because of the single constraints. For example, the roof may be translated as grass from $S$ to $T$ and then the incorrectly translated grass will be incorrectly translated to the roof from $T$ to $S$, where the cycle loss is minimized but the translation is bidirectionally wrong. To mitigate the bidirectionally incorrect problem, DCCS is utilized to constrain the depth cycle consistency in the image translation process. Specifically, for target stylized images $x_{S \rightarrow T}$, the corresponding depth prediction $\hat{z}_{S \rightarrow T}$ can be generated by DRDG$_{T \rightarrow S}$. As well as the DSL, a supervised depth loss is applied to $\hat{z}_{S \rightarrow T}$ and $z_S$, which forces the depth information immutable during the image translation process. A Berhu loss is used in DCCL

$$L_{\text{DCCL}}(S, T) = E_{\{z_S, \hat{z}_S, \hat{z}_{S \rightarrow T}\}}(\text{Berhu}(z_S, \hat{z}_{S \rightarrow T}))$$  

(6)

where we omit the nearest resizing operation for $z_S$ for simplifying the expression.

3) Total Loss: In the end, the total loss of the image translation process can be written as

$$L_{\text{total}} = \lambda_{\text{adv}}(L_{\text{adv}}(S, T) + L_{\text{adv}}(T, S)) + \lambda_{\text{cyc}}(L_{\text{cyc}}(S, T) + L_{\text{cyc}}(T, S)) + \lambda_{\text{DSL}}(L_{\text{DSL}}(S) + L_{\text{DSL}}(T)) + \lambda_{\text{DCCL}}(L_{\text{DCCL}}(S, T) + L_{\text{DCCL}}(T, S))$$  

(7)

where $\lambda_{\text{adv}}, \lambda_{\text{cyc}}, \lambda_{\text{DSL}}$, and $\lambda_{\text{DCCL}}$ are hyperparameters. By minimizing $L_{\text{total}}$, a well-constrained DRDG$_{S \rightarrow T}$ can be obtained to perform the image translation process to generate $x_{S \rightarrow T}$.

After the image translation is finished, a cross-entropy loss is used to train the semantic segmentation model $f_T$

$$L_{\text{seg}} = E_{(x_{S \rightarrow T}, y_S)}(CE(f_T(x_{S \rightarrow T}, y_S))$$  

(8)

where cross-entropy loss $CE$ is

$$CE(f, x, y) = -\sum_{c=1}^{C} y_c \cdot \log(f(x)).$$  

(9)

C. Network Settings

The architecture of DRDG follows the settings of RDG [4], which is a U-shaped full convolution network. The convolution kernel size is 4, the stride is 2 and the padding is 1. Channels of convolution layers are set as {64, 128, 256, 512, 512, 512}. The discrepancy between DRDG and RDG is the depth decoder in DRDG. The depth decoder is identical to the RDG decoder except for the channel of the last layer is set as 1, followed by a softmax activation to predict a depth image. The discriminator is a full convolution network with the same kernel with DRDG and channels of {64, 128, 256, 512, 512, 1}. DeepLabV3 is selected as our baseline of semantic segmentation. Based on thorough experiments, the hyperparameters ($\lambda_{\text{adv}}, \lambda_{\text{cyc}}, \lambda_{\text{DSL}}, \lambda_{\text{DCCL}}$) are set as (5, 10, 2, 1), respectively.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Datasets

Two open-source RS datasets: Potsdam and Vaihingen are used for validating the proposed method [14]. Potsdam is set as the source domain, while Vaihingen is set as the target domain in the following experiments. Both of the images are produced into true orthophotos (TOPs), with annotations for six ground classes: clutter/background, impervious surfaces, car, tree, low vegetation, and building. We select the IR-R-G bands mode of Potsdam as Potsdam infrared red-red-green (IRRG), which consists of 38 TOPs in which every TOP contains $6000 \times 6000$ pixels, with a resolution of 5 cm. We select the IR-R-G bands mode of Vaihingen which consists
of 33 TOPs and every TOP contains 2000 × 2000 pixels, with a resolution of 9 cm. To compare the proposed method with other methods more conveniently, we exploit the second, fifth, seventh, eighth, 13th, 20th, 22nd, and 24th TOPs in Vaihingen as the validation dataset, while others for the test dataset, which follows other works settings. Following the settings of RDG, we also clip images of PotsdamIRRG into the size of 896 × 896 and images of Vaihingen into the size of 512 × 512.

DSMs of PotsdamIRRG and Vaihingen are provided in the raw dataset. We follow the above processes and normalize values of DSMs into the range of 0–1.

B. Experimental Results

To facilitate comparing with different methods, IoU and F1_score are employed as metrics in this letter. For every class in six different ground classes, the formulation of IoU can be written as

\[
\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \quad (10)
\]

where A is the ground truth, while B is predictions. After calculations of IoU for six classes, mIoU can be obtained which is the mean of IoU for every class. And the F1_score can be written as

\[
\text{F1_score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (11)
\]

Table I shows the quantitative results of the proposed DRDG. The final result is the average of three results at the different random seeds. Five recent methods are selected for comparison, that is, Benjdira et al. [13], AdaptSegNet [5], MUCSS [1], Wittich and Rottensteiner [12], and RDG [4]. It is worth noting that, Wittich’s method needs additional DSMs as inputs at the testing phase, while our method and other methods need only three-band optical images as inputs. Quantitative results show that our method DRDG reaches the SOTA results between all methods by the mIoU of 53.26 and F1 of 65.75 in the experiment of PotsdamIRRG to Vaihingen, which surpasses RDG by 2.9% and 1.3%, respectively. Significantly, DRDG reaches the mIoU of 77.67 and F1 of 87.41 in the building class, which surpasses RDG by 4.2% and 2.4%, respectively. As shown in Fig. 1, the outline of buildings is distinguishable, which may contribute to improving the accuracy of the building class through the constrain of DSL and DCCL.

Compared with Wittich and Rottensteiner [12] method, our method is also better in mIoU and overall F1. However, Wittich’s method shows better results in the building class surpassing our method by 3.23% of IoU and by 1.97% of F1. It is plausible that Wittich’s method needs an additional DSM band as input at the test phase, while our model does not. The DSM gives a direct guide to the segmentation model to clearly distinguish the building class. It should be noted that it is difficult for an optical camera to get accurate DSMs in real-time during the testing phase. Consequently, our model is easier to deploy on common devices with only optical cameras compared with Wittich’s method.

Fig. 3 is the qualitative results of DRDG, which also shows the superiority of our method.

C. Ablation Study

Table II shows the ablation study of DRDG. We performed the experiments of removing DCCL (w/o DCCL), removing DSL (w/o DSL), and removing both DCCL and DSL (RDG). Experimental results show that both DCCL and DSL contribute to the improvements of DRDG. Significantly, if only DCCL is involved, the mIoU and F1 decline to 50.33 and 63.39, respectively. It is plausible that DCCL is designed to constrain the depth consistency but to bring depth information into the model. If only DCCL is used without DSL, the generators merely learn how to predict DSMs of translated images while we cannot ensure the semantic correctness of translated images. If there are some semantic errors in the translated images, the generators learn the wrong information from this situation. As a result, the performance will decline if only DCCL is involved. As a comparison, because DSL
always brings correct information into models, the mIoU and F1 slightly improve by 1.18% and 0.63%, respectively, if only DSL is involved. When combining the DSL and DCCL, the generators learn information from DSMs by DSL and constrain semantic correctness by DCCL while translating images. $\lambda_{DCCL}$ is set smaller than $\lambda_{DSL}$ to ensure DSL brings correct depth information into the model.

### IV. Conclusion

In this letter, a depth-assisted GAN-based framework, DRDG, is proposed for aerial image cross-domain semantic segmentation. The DSM information is introduced into the GAN process by two constraints: DSL and DCCL, which force GAN models to maintain the DSM correctness during generating images. The experimental results show the priority of the proposed method, which reaches the state-of-the-art performance of the generative method in the PotsdamIRRG to Vaihingen cross-domain segmentation task.

There is still a lot of work to be done in our letter. The first is to resolve the instability of the GAN-based method. Although RDG has made a progress on stability, however, the results still fluctuate greatly at different random initialization. The second is to find an effective way to directly combine the DSM with the segmentation model. Training the model through an end-to-end approach instead of a multistage framework may yield a more stable result. The last is to utilize self-training methods to further improve performance. DSMs are accessible in both the source domain and the target domain, which facilitates self-training strategies to initialize the weights of models.

#### ACKNOWLEDGMENT

The authors would like to thank the support from Professor Ge Li at Peking University and the computation resources from VirtAI Tech.

#### REFERENCES

[1] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, “Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 20–33, May 2021.

[2] S. Ji, D. Wang, and M. Luo, “Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3816–3828, May 2021.

[3] J. Hoffman et al., “CyCADA: Cycle-consistent adversarial domain adaptation,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[4] Y. Zhao, P. Guo, Z. Sun, X. Chen, and H. Gao, “ResiDualGAN: Residual DualGAN for cross-domain remote sensing images semantic segmentation,” 2022, arXiv:2201.11523.

[5] Y.-H. Tsai, W.-C. Hung, S. Schuller, K. Sohn, M.-H. Yang, and M. Chandraker, “Learning to adapt structured output space for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[6] Y. Zou, Z. Yu, B. V. K. Kumar, and J. Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.

[7] H. Gao, Y. Zhao, P. Guo, Z. Sun, X. Chen, and Y. Tang, “Cycle and self-supervised consistency training for adapting semantic segmentation of aerial images,” *Remote Sens.*, vol. 14, no. 7, p. 1527, Mar. 2022.

[8] K.-H. Lee, G. Ros, J. Li, and A. Guindon, “SPGAN: Privileged adversarial learning from simulation,” 2018, arXiv:1810.08756.

[9] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. P. Perez, “DADA: Depth-aware domain adaptation in semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7364–7373.

[10] S. Saha et al., “Learning to relate depth and semantics for unsupervised domain adaptation,” 2021, arXiv:2105.07830.

[11] Q. Wang, D. Dai, L. Hoyer, L. Van Gool, and O. Fink, “Domain adaptive semantic segmentation with self-supervised depth estimation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 8515–8525.

[12] D. Wittich and F. Rottensteiner, “Appearance based deep domain adaptation for the classification of aerial images,” *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 82–102, Oct. 2021.

[13] B. Benjdira, Y. Bazi, A. Koubba, and K. Ouni, “Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images,” *Remote Sens.*, vol. 11, no. 11, p. 1369, 2019.

[14] (2022). ISPRS WG II/4, ISPRS 2D Semantic Labeling Contest. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling