UGCNet: An Unsupervised Semantic Segmentation Network Embedded With Geometry Consistency for Remote-Sensing Images

Danpei Zhao, *Member, IEEE*, Bo Yuan, Yue Gao, Xinhu Qi, and Zhenwei Shi, *Member, IEEE*

***Abstract*— In remote-sensing image (RSI) semantic segmentation, the dependence on large-scale and pixel-level annotated data has been a critical factor restricting its development. In this letter, we propose an unsupervised semantic segmentation network embedded with geometry consistency (UGCNet) for RSIs, which imports the adversarial-generative learning strategy into a semantic segmentation network. The proposed UGCNet can be trained on a source-domain dataset and achieve accurate segmentation results on a different target-domain dataset. Furthermore, for refining the remote-sensing target geometric representation such as densely distributed buildings, we propose a geometry-consistency (GC) constraint that can be embedded in both image-domain adaptation process and semantic segmentation network. Therefore, our model could achieve cross-domain semantic segmentation with target geometric property preservation. The experimental results on Massachusetts and Inria buildings datasets prove that the proposed unsupervised UGCNet could achieve a very comparable segmentation accuracy with the fully supervised model, which validates the effectiveness of the proposed method.**

***Index Terms*— Generative-adversarial learning, geometry consistency (GC), remote-sensing images (RSIs), semantic segmentation, unsupervised.**

I. Introduction

SEMANTIC segmentation aims to assign a label to every single pixel in the image. Due to the rapid development of deep learning in recent years, numerous semantic segmentation algorithms like [1], [2] for remote-sensing images (RSIs) are proposed. However, most of them extremely rely on large-scale pixel-level annotated datasets, which leads to a narrow application situation. During the last couple of years, a number of domain-adaptation methods are applied in image-to-image (I2I) translation [3]–[5] and semantic segmentation [6]–[11], providing a board prospect for unsupervised semantic segmentation.

Based on previous research, one rational strategy for unsupervised semantic segmentation task is to utilize image domain adaptation. In line with generative adversarial learning, the image domain adaptation aims at transferring the model learned on a labeled source domain to a target domain. For example, [6], [7], [9] propose to decouple the image to pixel level and representation level and conduct the translation. Luo *et al.* [10] design a category-level adversaries for semantic-consistent domain adaptation. Recently, [11], [12] introduces self-supervision strategy to the domain-adaptation phase and [13] designs an entropy loss to penalize low-confident predictions on target domain. Lv *et al.* [14] emphasize domain-invariant features by constructing pivot information. However, the above-mentioned methods tried to optimize the adaptation from pixel or representation level but ignored the image-level features. In RSIs, a very common phenomenon is that the objects have distinct geometric properties, and simple geometric transformations do not change the images' semantic structure. Here, the semantic structure refers to the information that distinguishes different staff/object classes, which can be easily perceived by humans regardless of trivial geometric transformations such as vertical flipping and rotation. In this letter, we propose an unsupervised semantic segmentation network embedded with geometry consistency (UGCNet) for RSIs that preserves image geometric structures during the adaptation process.

The UGCNet can be decomposed into a cross-domain adaptation network (CAN) and a geometry-consistent segmentation network (GSN). In the CAN, the goal is to learn a mapping function that only changes the image style without distorting the semantic structures. In the GSN, a pixel-level segmentation network is trained on the transferred source-domain images, and its model can be applied on the target-domain images. Compared with fully supervised semantic segmentation models, the UGCNet alleviates the challenge of obtaining sufficient training data. While different from unsupervised approaches like [6], [7], our UGCNet pays more attention on the target geometric structures. For densely distributed buildings with various scales, the proposed method could maintain clear boundary of independent targets. Our contributions are summarized as follows.

1) Through the adaptation between source-domain images and target-domain images, we propose a novel UGCNet for RSIs. Our model achieves very comparable segmentation accuracy with the fully-supervised approach on the building extraction task.
2) We propose a geometry-consistency (GC) constraint that could be embedded in both image translation and

Manuscript received February 2, 2021; revised July 22, 2021; accepted November 18, 2021. Date of publication November 22, 2021; date of current version January 7, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC1510905 and in part by the Air Force Equipment Pre-Research Project under Grant 303020401. *(Corresponding author: Danpei Zhao.)*

Danpei Zhao, Bo Yuan, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: zhaodanpei@buaa.edu.cn; yuanbobuaa@buaa.edu.cn; shizhenwei@buaa.edu.cn).

Yue Gao and Xinhu Qi are with Space Star Technology Company Ltd., Beijing 100094, China (e-mail: bjlguniversity@163.com; heavenww@163.com).

Digital Object Identifier 10.1109/LGRS.2021.3129776

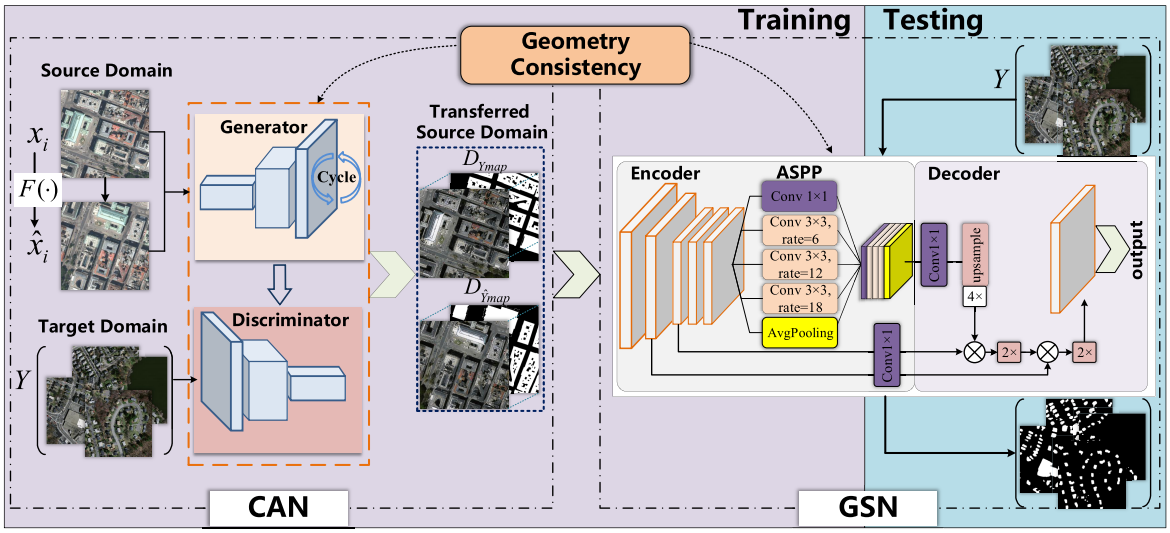

Fig. 1. Pipeline of the proposed UGCNet. It consists of two in-line modules: the CAN and the GSN. The geometry-consistency (GC) constraint is embedded in both CAN and GSN. In the training phase, both the CAN and GSN are trained. While in the testing stage, the target-domain's input only goes through the trained GSN model to output segmentation map.

semantic segmentation networks, which simultaneously preserves semantic structures of source-domain images and improves semantic segmentation performance on target domain.

II. Methodology

The architecture of the proposed UGCNet is shown in Fig. 1. It can be decomposed into two main components, of which the CAN is for image domain adaptation and the GSN is for pixel-level segmentation. Given the source domain images with pixel-level annotations and unlabeled target domain images, the CAN transfers images from the source domain to the target domain in an adversarial manner. The GSN is trained on the transferred source-domain data and could be applied on the target domain. The proposed GC constraint module is embedded in both CAN and GSN by inserting specific loss components, respectively.

A. Geometry-Consistency Constraint

To start with, let $\mathcal{X}$ and $\mathcal{Y}$ be two domains with unpaired training examples $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$, where x_i and y_i are drawn from the marginal distributions P_X and P_Y, where X and Y are two random variables associated with $\mathcal{X}$ and $\mathcal{Y}$, respectively. $F(\cdot)$ is a predefined transformation function. $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$ are obtained by applying $F(\cdot)$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively. G_{XY} and $G_{\hat{X}\hat{Y}}$ are the translators which target the adaptation tasks from $\mathcal{X}$ to $\mathcal{Y}$ and $\hat{\mathcal{X}}$ to $\hat{\mathcal{Y}}$. D_Y and $D_{\hat{Y}}$ are two adversarial discriminators in domains $\mathcal{Y}$ and $\hat{\mathcal{Y}}$.

Here, we first take a review of the cycle consistency [3]. Through the translators $G_{XY} \circ G_{YX} : X \to Y \to X$ and $G_{YX} \circ G_{XY} : Y \to X \to Y$, the examples x and y in domains $\mathcal{X}$ and $\mathcal{Y}$ should satisfy $x \approx G_{YX}(G_{XY}(x))$ and $y \approx G_{XY}(G_{YX}(y))$. Cycle consistency is implemented by a bidirectional reconstruction process that requires G_{XY} and G_{YX} to be jointly learned. Motivated by [5], the implementation of our GC constraint is based on a fact that image

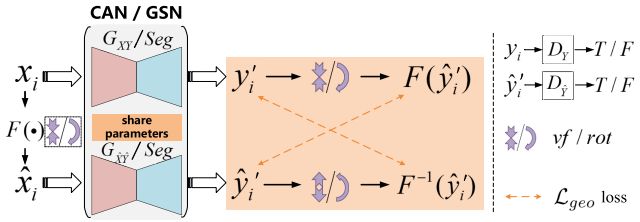

Fig. 2. GC constraint.

semantic structures can be preserved even after simple geometric transformations. Fig. 2 illustrates how the GC constraint applied in the CAN/GSN. For example, in the CAN, given a $F(\cdot)$, the GC constraint can be expressed as $F(G_{XY}(x)) \approx G_{\hat{X}\hat{Y}}(F(x))$ and $F^{-1}(G_{\hat{X}\hat{Y}}(F(x))) \approx G_{XY}(x)$, where $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$. Similarly, in the GSN, the GC constraint can be expressed as $F(\text{Seg}(x)) \approx \text{Seg}(F(x))$ and $F^{-1}(\text{Seg}(F(x))) \approx \text{Seg}(x)$, where $\text{Seg}(\cdot)$ represents a segmenter. In this letter, we employ two representative geometric transformations, that is, vertical flipping (*vf*) and 90° clockwise rotation (*rot*), to execute the GC constraint.

B. Cross-Domain Adaptation Network (CAN)

The CAN transfers the images from one domain to another to change appearance while preserving source semantic content as much as possible. It contains the following components: G_{XY}, $G_{\hat{X}\hat{Y}}$, D_Y, $D_{\hat{Y}}$, and geometric translator $F(\cdot)$. Assuming that $\mathcal{X}$ represents the source-domain dataset and $\mathcal{Y}$ is the target-domain dataset, then $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. The CAN aims to learn two mappings $G_{XY}(x)$ and $G_{\hat{X}\hat{Y}}(\hat{x})$, where $\hat{x} = F(x)$. As shown in Fig. 2, given a predefined $F(\cdot)$, we feed the images $x \in X$ and $\hat{x} = F(x)$ into the G_{XY} and $G_{\hat{X}\hat{Y}}$, respectively. Following the GC constraint, the outputs $y' = G_{XY}(x)$ and $\hat{y}' = G_{\hat{X}\hat{Y}}(\hat{x})$ ought to satisfy $\hat{y}' \approx F(y')$ and $y' \approx F^{-1}(\hat{y}')$. Considering both $F(\cdot)$ and the inverse geometric transformation function $F^{-1}(\cdot)$, the GC

loss $\mathcal{L}_{\text{geo}}(G_{XY}, G_{\hat{X}\hat{Y}}, X, Y)$ in the CAN is defined as

$$\begin{aligned}\mathcal{L}_{geo}&\left(G_{XY}, G_{\hat{X}\hat{Y}}, X, Y\right)\\ &= \mathbb{E}_{x\sim P_X}\left[\left|\left|G_{XY}(x) - F^{-1}\left(G_{\hat{X}\hat{Y}}(F(\hat{x}))\right)\right|\right|_1\right]\\ &\quad+\mathbb{E}_{x\sim P_X}\left[\left|\left|G_{\hat{X}\hat{Y}}(F(\hat{x})) - F(G_{XY}(x))\right|\right|_1\right]. \end{aligned} \tag{1}$$

In the CAN, we employ the same discriminator and generator as CycleGAN [3]. The generator is a standard encoder–decoder, where the encoder contains two convolutional layers with stride 2 and 9 residual blocks. The decoder contains two deconvolutional layers also with stride 2. The discriminator distinguishes images at the patch level following [15]. And for parameter reduction, G_{XY} and $G_{\hat{X}\hat{Y}}$ share all the parameters. In the transformed domains $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$, the adversarial loss $\mathcal{L}_{\text{gan}}(G_{\hat{X}\hat{Y}}, D_{\hat{Y}}, \hat{X}, \hat{Y})$ has the same form with $\mathcal{L}_{\text{gan}}(G_{XY}, D_Y, X, Y)$. By combining the GC constraint with the standard adversarial constraint and cycle constraint [3], the full objective for CAN has the following form:

$$\begin{aligned}\mathcal{L}_{\text{CAN}} &= \mathcal{L}_{\text{gan}}(G_{XY}, D_Y, X, Y) + \mathcal{L}_{\text{gan}}\left(G_{\hat{X}\hat{Y}}, D_{\hat{Y}}, \hat{X}, \hat{Y}\right)\\ &\quad+\gamma\left[\mathcal{L}_{\text{cyc}}(G_{XY}, G_{YX}, X, Y)+\mathcal{L}_{\text{cyc}}\left(G_{\hat{X}\hat{Y}}, G_{\hat{Y}\hat{X}}, \hat{X}, \hat{Y}\right)\right]\\ &\quad+\lambda\mathcal{L}_{\text{geo}}\left(G_{XY}, G_{\hat{X}\hat{Y}}, X, Y\right)\end{aligned} \tag{2}$$

where γ and λ are to weigh the contribution of $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{\text{geo}}$ during the training process, respectively.

C. Geometry-Consistent Segmentation Network (GSN)

The purpose of the GSN is to train a segmentation model on the transferred source domain which can be applied on the unlabeled target domain. Closely connected with CAN, the GSN consists of an encoder that contains an fully convolutional network (FCN), atrous spatial pyramid pooling (ASPP) [16], and a decoder. Functionally, the encoder captures multi-level image features, and the decoder recovers the spatial information from low-resolution feature map and output the segmentation map.

However, FCN usually has poor performance when multi-scale objects coexist. Previous researches [1], [17] have demonstrated that multi-scale feature fusion is of great benefit to improve segmentation efficiency. Here, we extend ASPP to perform multi-rate dilated convolutions for extracting multi-scale features in spatial pyramid form. As the input of the GSN, the transferred source-domain images are first put in the encoder to extract the features computed by deep residual networks. We set output stride as 16 as the ratio of input image spatial resolution to the final output resolution. In the decoder, we first use a 1×1 convolution to reduce channels, after which the encoder features are bilinearly upsampled by $4\times$. Then the features are concatenated with the corresponding low-level features from the network backbone that have the same spatial resolution. We use a few 3×3 convolutions to refine the concatenate features and use two $2\times$ bilinear upsamplings after the concatenation to gradually recover the segmentation map that has the same resolution with the input.

In the GSN, we use cross-entropy (CE) loss and the GC constraint loss to supervise the network training. Similar to the CAN, the input images are y' and $\hat{y}'$, the segmentation outputs $o = \text{Seg}(y')$ and $\hat{o} = \text{Seg}(\hat{y}')$ should also satisfy $o \approx F^{-1}(\hat{o})$ and $\hat{o} \approx F(o)$. Hence, the GC loss $\mathcal{L}_{\text{geo}}(o, \hat{o})$ in the GSN is set as

$$\mathcal{L}_{\text{geo}}(o, \hat{o}) = \text{smooth}_{L1}\left(o, F^{-1}(\hat{o})\right) + \text{smooth}_{L1}(\hat{o}, F(o)) \tag{3}$$

where

$$\text{smooth}_{L1}(x) = \begin{cases}0.5x^2, & \text{if } x \geq 0\\ |x| - 0.5, & \text{otherwise}\end{cases}$$

is less sensitive to outliers than $L2$ loss and can prevent exploding gradients. x is the element-wise discrepancy between the prediction and ground truth. The full objective for the GSN is defined as

$$\mathcal{L}_{\text{GSN}} = \beta_1\mathcal{L}_{\text{CE}}\left(o, x_{gt}\right) + \beta_2\mathcal{L}_{\text{CE}}\left(\hat{o}, \hat{x}_{gt}\right) + \delta\mathcal{L}_{\text{geo}}(o, \hat{o}) \tag{4}$$

where x_{gt} represents the ground truth of source-domain images and $\hat{x}_{gt}$ is the geometric transformations of x_{gt}. β_1 and β_2 are two constant coefficients. δ is a trade-off hyperparameter to weigh the contribution of the GC constraint.

D. Joint Training

The proposed UGCNet supports joint training. Given a labeled source-domain dataset and an unlabeled target-domain dataset, the CAN first transfers the source-domain images with generative-adversarial learning manner. Then the transferred source-domain images are used as the input of the GSN to train the segmentation model and its model could be performed on the target-domain images. Comprehensively, the final loss for the UGCNet is

$$\mathcal{L}_{\text{UGC}} = \alpha_1\mathcal{L}_{\text{CAN}} + \alpha_2\mathcal{L}_{\text{GSN}} \tag{5}$$

where α_1 and α_2 are two coefficients. The pseudocode of our algorithm is shown in Algorithm 1.

III. Experiments

A. Datasets and Evaluation Metrics

To validate the performance of the proposed UGCNet, we test it on Massachusetts buildings dataset (Massachusetts)

Algorithm 1 Training Process of the UGCNet

Input:
Source-domain images X and labels X_{gt}, Target-domain images Y;
Suppose $x \in X, x_{gt} \in X_{gt}, y \in Y$;
Output:
Predicted labels of the target domain Y: O_y;
while iteration is effective:
  $F(x) \to \hat{x}$, $F(x_{gt}) \to \hat{x}_{gt}$ {forward prop}
  $G_{XY}(x) \to y'$, $G_{\hat{X}\hat{Y}}(\hat{x}) \to \hat{y}'$ {forward prop}
  $D_Y(\{y', y\}) \to D_{Y\text{map}}$, $D_{\hat{Y}}(\{\hat{y}', \hat{y}\}) \to D_{\hat{Y}\text{map}}$ {forward prop}
  Minimize $\mathcal{L}_{\text{CAN}}$ in (2). {backward prop}
  Pair $(\{y', x_{gt}\})$, $(\{\hat{y}', \hat{x}_{gt}\})$ to train the GSN.
  $\text{GSN}(y') \to o, \text{GSN}(\hat{y}') \to \hat{o}$ {forward prop}
  Minimize $\mathcal{L}_{\text{GSN}}$ in (4). {backward prop}
end **while**

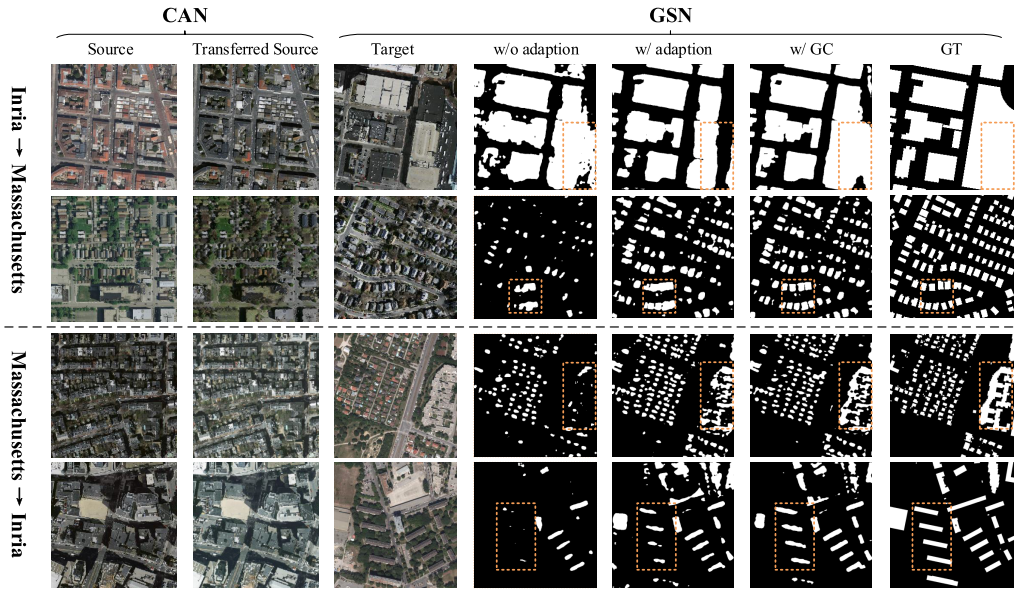

Fig. 3. Visualization results of the proposed UGCNet. The first two columns display the original and transferred source-domain images, respectively. The third column exhibits the target-domain test images. The fourth column indicates the GSN trains on the original source-domain images and tests on the target-domain images. The fifth column represents the GSN trains on the transferred source-domain images and tests on the target-domain images. The sixth column shows the segmentation results by using the GC constraint, which also corresponds with the output segmentation map of the proposed UGCNet.

and Inria aerial image labeling dataset (Inria). Both of them contain only two categories: buildings and others. Numerically, the Massachusetts dataset consists of 151 aerial images of the Boston area, with 1-m resolution and 1500 × 1500 pixels. The Inria dataset consists of 180 images with 0.3-m resolution and 5000 × 5000 pixels. In comparison, the Inria covers 810 km^2 dissimilar urban settlements, ranging from densely populated areas to alpine towns, while the Massachusetts covers 340 km^2. Perceptually, pixel values of buildings in the Massachusetts are lower than those in the Inria in the most cases. In addition, an average omission noise level of 5% is applied in the Massachusetts as partial areas in some samples are covered with solid color. While the Inria covers urban areas with no noisy tags. Considering the GPU capacity, we randomly cut each sample to several 512 × 512 pixels sub-images and totally obtain 5010 pieces (4110 for training and 900 for testing) from the Massachusetts and 3600 pieces (2800 for training and 800 for testing) from the Inria. We apply pixel accuracy (PA) and intersection over union (IoU) as the evaluation metrics.

B. Implementation Details

In the CAN, we set the training batch size as 8 and testing batch size as 1. The learning rate is fixed in the initial 100 epochs and linearly decays over the following epochs. In the GSN, the baseline of the encoder is ResNet-101 by removing its fully connected layers. All input samples are resized to 512 × 512 × 3. In the encoder, the output stride is 16. The initial learning rate is multiplied by $(1-(\text{iter})/((\text{max\_iter})))^{\text{power}}$ with power = 0.9. We set $\gamma = 10$, $\lambda = 20$, $\delta = 0.05$, $\beta_1 = \beta_2 = 0.1$, $\alpha_1 = 1, \alpha_2 = 0$ at the first 200 epochs and $\alpha_1 = \alpha_2 = 1$ for the following training stage. All experiments are performed on four NVIDIA 2080Ti GPUs.

TABLE I

CROSS-DOMAIN SEMANTIC SEGMENTATION PERFORMANCE AND ABLATION ANALYSIS OF THE PROPOSED UGCNET

Source→Target	Supervision			Evaluation	
	baseline	+*rot*	+*vf*	PA (%)	mIoU (%)
fully supervised	✓			*94.62*	*77.60±0.04*
w/o adaptation	✓			*89.98*	*69.77±0.15*
Inria	✓			90.96	73.61±0.11
	✓	✓		**92.35**	75.49±0.08
↓	✓		✓	92.27	75.53±0.07
Massachusetts	✓	✓	✓	92.22	**76.54**±0.05
fully supervised	✓			*89.38*	*76.50±0.03*
w/o adaptation	✓			*81.08*	*63.01±0.12*
Massachusetts	✓			84.48	67.02±0.08
	✓	✓		86.62	70.28±0.06
↓	✓		✓	85.69	70.54±0.03
Inria	✓	✓	✓	**86.75**	**70.58**±0.04

"w/o adaptation" indicates GSN directly trains on source-domain train-set and tests on target-domain test-set. "fully supervised" means GSN both trains and tests on target-domain data set. "baseline" refers to GSN trains without geometry-consistency constraint. "*rot*" and "*vf*" refer to two kinds of geometry-consistency constraints. "w/ adaptation" indicates GSN trains on the transferred source-domain train-set and tests on target-domain test-set. "w/ GC" means GSN trains with the geometry-consistency constraint.

C. Performance Analysis and Comparison

To validate the proposed method, we conduct a series of experiments on building extraction task of RSIs. We test our model on Inria→Massachusetts (training on transferred Inria train-set and testing on Massachusetts test-set) and Massachusetts→Inria, respectively. For reference, we also show the performance of w/o adaptation (e.g., directly training Inria train-set and testing on Massachusetts test-set) situation and the fully-supervised model (e.g., both training and testing on Inria). As shown in Table I, domain adaptation procedure

TABLE II

PERFORMANCE COMPARISON IN TERMS OF PER-CLASS IOUS AND MIOU (%). * INDICATES BOTH THE *rot* AND *vf* ARE USED

Source→Target	Method	building	others	mIoU
Inria	AdaptSegNet [9]	64.10	76.58	70.34
↓	CLAN [10]	68.42	82.54	75.48
Massachusetts	UGCNet*	**69.56**	**83.52**	**76.54**
Massachusetts	AdaptSegNet [9]	62.83	70.22	66.53
↓	CLAN [10]	65.84	74.14	69.99
Inria	UGCNet*	**66.38**	**74.78**	**70.58**

TABLE III

NETWORK EFFICIENCY OF THE GSN

Metric \ Backbone	ResNet-101	Xception	MobileNetv2
FLOPs	88.69G	82.76G	26.49G
Param.	59.34M	54.71M	5.82M
mIoU (w/o GC)	73.61	66.58	53.22
mIoU (w/ GC)	76.54(+**2.93**)	69.42(+**2.84**)	57.01(+**3.79**)

The metrics are obtained on Inria→Massachusetts cross-domain segmentation task.

in the CAN can effectively improve the segmentation performance on the target domain in both directions, for example, +3.84% (73.61% versus 69.77%) IoU improvement on Inria→Massachusetts. On the other hand, the GC constraints embedded in the GSN also improve the network performance, for example, +2.93% IoU on Inria→Massachusetts task. In addition, we present comparison of the UGCNet with several SOTA approaches including [9], [10] in Table II. We observe that UGCNet outperforms these models and achieves the highest mIoU. Fig. 3 displays the visualization results. The first two columns show the image translation results from the source domain to the target domain. In the target domain segmentation task, our model could maintain internal completeness and boundary of densely distributed individual target.

D. Network Efficiency Analysis of GSN

As shown in Table III, we survey the proposed GSN by using various widely used and efficient backbones including ResNet, Xception, and MobileNet. During the training process, we import the above-mentioned two kinds of GC constraints along with binary cross entropy as supervision. As a result, the model using deep residual ResNet-101 with dilated convolutions achieves the highest 76.54% mIoU, which is 2.93% higher than the model training without GC constraint. The testing results using Xception and MobileNetv2 also validate the effectiveness of GC constraint.

IV. CONCLUSION

In this letter, we propose an UGCNet for RSIs. By using generative-adversarial learning strategy, we adapt the source-domain images to the target-domain format to train the segmentation network for better segmentation accuracy in the target domain. To preserve object geometric representations in RSIs, we design a GC constraint for both domain adaptation and segmentation. Ultimately, we test the proposed UGCNet on Massachusetts and Inria buildings datasets. The experimental results demonstrate that the proposed domain adaptation strategy and the GC constraint could visibly improve the cross-domain semantic segmentation performance. We will further extend our model on more kinds of objects in RSIs.

REFERENCES

[1] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2612–2626, Aug. 2019.

[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.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[4] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, "TransGaGa: Geometry-aware unsupervised image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8004–8013.

[5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao, "Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2422–2431.

[6] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.

[7] L. Shi, Z. Wang, B. Pan, and Z. Shi, "An end-to-end network for remote sensing imagery semantic segmentation via joint pixel-and representation-level domain adaptation," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1896–1900, Nov. 2021.

[8] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6777–6786.

[9] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[10] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.

[11] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3764–3773.

[12] Z. Wang *et al.*, "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12632–12641.

[13] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2517–2526.

[14] F. Lv, T. Liang, X. Chen, and G. Lin, "Cross-domain semantic segmentation via domain-invariant interactive relation transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4333–4342.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.