DOMAIN-AGNOSTIC DOMAIN ADAPTION FOR BUILDING FOOTPRINT EXTRACTION

Fahong Zhang[1], *Yilei Shi*[2], *Xiao Xiang Zhu*[1,3]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2]Chair of Remote Sensing Technology (LMF), Technical University of Munich, Munich, Germany
[3] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883996

ABSTRACT

For global range satellite imaging mission, images captured from different areas may have large distribution biases due to different illuminations, shooting angles and atmospheric conditions. A straightforward idea to mitigate this problem is to categorize the images into different domains according the cities they belong to, and apply domain adaptation approaches. However, categorization by cities becomes unreasonable with the increase of the city number, and the emergence of inter-city similarity and intra-city discrepancy.

With such consideration, this paper proposes a novel domain adaptation method named domain-agnostic domain adaptation (DADA) to reduce the distribution biases without explicitly defining the domain each image belongs to. To implement this, we augment the images to the styles of different domains by Generative Adversarial Networks (GAN) and contrastive learning to increase the generalizability of downstream tasks. Experiments on Planetscope building footprint extraction datasets verify the effectiveness of our method.

***Index Terms*—** Domain Adaptation, Generative Adversarial Networks, Contrastive Learning.

1. INTRODUCTION

With the rapid developments of satellite imaging techniques and the increase of spacing missions, massive amounts of world-wide remote sensing data can be achieved with less efforts. This enables the emergence of global range remote sensing applications such as object detection [1], land cover classification [2] and build footprint extraction [3]. Among earlier practices, people notice the generalization inferiority when applying models trained on source cities to unseen target cities, i.e., the domain shift problem. The underlying reason could be the source and target data distributions are biased due to different illuminations, shooting angles and atmospheric conditions.

To solve such problem, domain adaptation methods are adopted to increase the generalizability of downstream networks [4, 5]. Previous methods treat each city as a domain and perform single-source or multi-source domain adaptation to stylize images to the appearances of different domains. However, with the increase of the city number (e.g., 10^2 - 10^3), there could be different cities with similar appearances. Besides, image patches from the same city could be of different styles when the city image are composited by two or more different flights. Theses kinds of inter-city similarities and intra-city discrepancies violate the assumption that each city image corresponds to a domain.

To mitigate this problem, we propose a domain-agnostic domain adaptation (DADA) approach without explicitly defining the domain, but seek to exploit the spatial neighboring relation among image patches, and contrastively learn to model the similarity between images. The contributions of this paper can be summarized as follows:

- We highlight and study the inter-city similarity and intra-city discrepancy problem occurred when applying domain adaptation methods on large-scale global range appliactions.
- We propose a domain-agnostic domain adaptation (DADA) methods to solve such problem without explicitly defining the domain each image belongs to. In DADA, a contrastively learning and adversarial learning-based framework is built to generate images with different styles, and further improve the generalizability.
- Experiments on Planetscope building footprint datasets demonstrate the effectiveness of the proposed methods both qualitatively and quantitatively.

2. METHODOLOGY

The overall architecture of DADA is illustrated in Fig. 1. To evaluate the effectiveness of DADA, we set building footprint extraction as the downstream task, yet one can easily extend it to other tasks.

2.1. Problem Formulation

This section formulates the domain adaptation setting for building footprint extraction. First, the source domain data

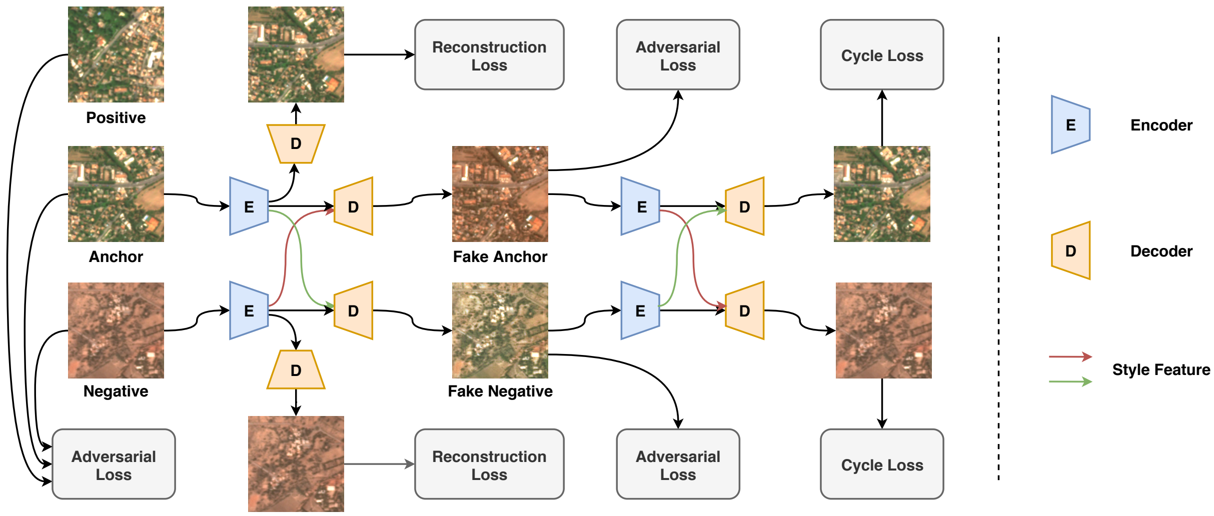

Fig. 1: Illustration of DADA. During the training phase, An anchor patch $\mathbf{x}_{anc}$, a positive patch $\mathbf{x}_{pos}$ and a negative patch $\mathbf{x}_{neg}$ will be sampled, in which $\mathbf{x}_{anc}$ and $\mathbf{x}_{pos}$ are neighboring patches. Inspired by previous arts [6, 4], a reconstruction loss and a cycle consistency loss are applied to maintain the structural information. Besides, a novel triplet adversarial loss is applied on $\mathbf{x}_{anc}$, $\mathbf{x}_{pos}$, $\mathbf{x}_{neg}$, the fake negative image $\tilde{\mathbf{x}}_{anc}$ and the fake anchor image $\tilde{\mathbf{x}}_{neg}$ simultaneously to learn to perform style transfer.

are given as $\mathcal{S} = \{(\mathbf{x}_i^s, \mathbf{y}_i^s\}_{i=1}^{N_s}$, where $\mathbf{x}_i^s \in \mathbb{R}^{H\times W\times 3}$ denotes the image patch and $\mathbf{y}_i^s \in \mathbb{R}^{H\times W}$ its label, indicating whether each pixel corresponds to the building area or not. The target domain data are given as $\mathcal{T} = \{(\mathbf{x}_i^t, \mathbf{y}_i^t\}_{i=1}^{N_t}$. In contrary to $\mathcal{S}$, the target domain labels $\mathbf{y}_i^t$ is only available during evaluation. During the acquisition of each image patch, we assume the coordinate information is recorded, which allow us to access the neighboring patches of each patch. Here we denote the neighboring patches of $\mathbf{x}_i$ as $\Omega_{\mathbf{x}_i}$. With such formulation, the building footprint extraction problem can be formulated as:

$$\min_{h} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{S}} \mathcal{L}_{seg}(\mathbf{x}, G(\mathbf{x}, \tilde{\mathbf{x}}), \mathbf{y}). \tag{1}$$

Here h denotes the segmentation networks, $G(\mathbf{x}, \tilde{\mathbf{x}})$ the fake image generated by the generator depicted in Fig. 1, and $\tilde{\mathbf{x}}$ a randomly sampled image patch that provide the style information. The segmentation loss $\mathcal{L}_{seg}$ is further defined as:

$$\mathcal{L}_{seg} = \mathcal{L}_{cse}(h(\mathbf{x}), \mathbf{y}) + \mathcal{L}_{cse}(h(G(\mathbf{x}, \tilde{\mathbf{x}})), \mathbf{y}), \tag{2}$$

where $\mathcal{L}_{cse}$ is the cross entropy loss. To balance the importance of building and non-building area, we use a class weight of $5:1$ when calculating $\mathcal{L}_{cse}$. The semantic segmentation networks are trained on the original images $\mathbf{x}$ as well as the stylized images $G(\mathbf{x}, \tilde{\mathbf{x}})$. Since $\tilde{\mathbf{x}}$ can be sampled from the target domain, the model's generalizability can be improved by training on target-style source data $G(\mathbf{x}, \tilde{\mathbf{x}})$.

The generator G is trained in an adversarial manner:

$$\min_{G,D} \sum_{\pi(\mathcal{S}\cup\mathcal{T})} \mathcal{L}_{gen} + \mathcal{L}_{dis}, \tag{3}$$

where D is the discriminator, π is a sampling function that will be introduced in Section 2.2, and $\mathcal{L}_{gen}$ and $\mathcal{L}_{dis}$ are loss functions that will be presented in Section 2.3 and 2.4.

2.2. Triplet Sampling

The generator networks G is trained based on a triplet sampling strategy. More specifically, an anchor patch $\mathbf{x}_{anc}$, a positive patch $\mathbf{x}_{pos}$ and a negative patch $\mathbf{x}_{neg}$ will be sampled from the union of source and target domain $\mathcal{S}\cup\mathcal{T}$ at each time: $(\mathbf{x}_{anc}, \mathbf{x}_{pos}, \mathbf{x}_{neg}) = \pi(\mathcal{S}\cup\mathcal{T})$. Here all of these three patches are randomly sampled from $\mathcal{S}\cup\mathcal{T}$, under the only restriction that $\mathbf{x}_{anc}$ and $\mathbf{x}_{pos}$ are neighboring patches, i.e., $\mathbf{x}_{pos} \in \Omega_{\mathbf{x}_{anc}}$.

2.3. Generator Loss $\mathcal{L}_{gen}$

The generator loss $\mathcal{L}_{gen}$ consists of self-reconstruction loss $\mathcal{L}_{rec}$, cycle loss $\mathcal{L}_{cyc}$ and adversarial loss $\mathcal{L}_{adv}$.

$$\mathcal{L}_{gen} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{cyc} + \lambda_3\mathcal{L}_{adv}. \tag{4}$$

$\mathcal{L}_{rec}$ is applied to ensure the features extracted by the encoder can be used to reconstruct the original image.

$$\mathcal{L}_{rec} = \sum_{\mathbf{x}\in\{\mathbf{x}_{anc},\mathbf{x}_{neg}\}} \left\| G_{dec}(G_{enc}(\mathbf{x}), G_{enc}(\mathbf{x})) - \mathbf{x} \right\|_1, \tag{5}$$

where $G_{enc}(\cdot)$ and $G_{dec}(\cdot,\cdot)$ are the encoder and decoder of the generator G. $G_{dec}(\mathbf{z}_1, \mathbf{z}_2)$ will first normalize $\mathbf{z}_1$ with the style of $\mathbf{z}_2$ by adaptive instance normalization (AdaIn) [7], and then decode $\mathbf{z}_1$ to the size of the original image.

The adversarial loss for the generator is defined as:

$$\begin{aligned}\mathcal{L}^g_{adv} = &(D(\tilde{\mathbf{x}}_{anc}) - 1)^2 + (D(\tilde{\mathbf{x}}_{neg}) - 1)^2 \\ &+ max(0, S(\mathbf{x}_{anc}, \mathbf{x}_{neg}) - S(\tilde{\mathbf{x}}_{anc}, \mathbf{x}_{neg}) + \alpha) \\ &+ max(0, S(\mathbf{x}_{anc}, \mathbf{x}_{neg}) - S(\mathbf{x}_{anc}, \tilde{\mathbf{x}}_{neg}) + \alpha).\end{aligned} \tag{6}$$

Here $\tilde{\mathbf{x}}_{anc}$ (or $\tilde{\mathbf{x}}_{neg}$) is the fake image generated from $\mathbf{x}_{anc}$ (or $\mathbf{x}_{neg}$) with the style of $\mathbf{x}_{neg}$ (or $\mathbf{x}_{anc}$) as illustrated in Fig. 1:

$$\begin{aligned}\tilde{\mathbf{x}}_{anc} &= G_{dec}(G_{enc}(\mathbf{x}_{anc}), G_{enc}(\mathbf{x}_{neg})), \\ \tilde{\mathbf{x}}_{neg} &= G_{dec}(G_{enc}(\mathbf{x}_{neg}), G_{enc}(\mathbf{x}_{anc})).\end{aligned} \tag{7}$$

$D(\mathbf{x})$ is the prediction of the discriminator, evaluating the probability that $\mathbf{x}$ is a real image rather than a fake one. $S(\mathbf{x}_1, \mathbf{x}_2)$ is the similarity of $\mathbf{x}_1$ and $\mathbf{x}_2$ measured by the discriminator according to their feature-level cosine similarity. The first two terms are utilized to cheat the discriminator to take the generated images as the real ones. The last two are triplet losses that are used to improve the stylization quality of $\tilde{\mathbf{x}}_{anc}$ and $\tilde{\mathbf{x}}_{neg}$, which incorporates a contrastive learning mechanism.

Cycle loss $\mathcal{L}_{cyc}$ is first proposed in CycleGAN [8], which is used to ensure the stylized images $\tilde{\mathbf{x}}_{anc}$ and $\tilde{\mathbf{x}}_{neg}$ can maintain the structural information in $\mathbf{x}_{anc}$ and $\mathbf{x}_{neg}$:

$$\mathcal{L}_{cyc} = \left\| \mathbf{x}'_{anc} - \mathbf{x}_{anc} \right\|_1 + \left\| \mathbf{x}'_{neg} - \mathbf{x}_{neg} \right\|_1. \tag{8}$$

Here $\mathbf{x}'_{anc}$ and $\mathbf{x}'_{neg}$ are images reconstructed from $\tilde{\mathbf{x}}_{anc}$ and $\tilde{\mathbf{x}}_{neg}$:

$$\begin{aligned}\mathbf{x}'_{anc} &= G_{dec}(G_{enc}(\tilde{\mathbf{x}}_{anc}), G_{enc}(\tilde{\mathbf{x}}_{neg})), \\ \mathbf{x}'_{neg} &= G_{dec}(G_{enc}(\tilde{\mathbf{x}}_{neg}), G_{enc}(\tilde{\mathbf{x}}_{anc})).\end{aligned} \tag{9}$$

2.4. Discriminator Loss $\mathcal{L}_{dis}$

The discriminator loss $\mathcal{L}_{dis}$ is defined as:

$$\begin{aligned}\mathcal{L}_{dis} = &(D(\tilde{\mathbf{x}}_{anc}) - 0)^2 + (D(\tilde{\mathbf{x}}_{anc}) - 0)^2 + (D(\mathbf{x}_{anc}) - 1)^2 \\ &+ max(0, S(\mathbf{x}_{anc}, \mathbf{x}_{neg}) - S(\tilde{\mathbf{x}}_{anc}, \mathbf{x}_{neg}) + \alpha) \\ &+ max(0, S(\mathbf{x}_{anc}, \mathbf{x}_{neg}) - S(\mathbf{x}_{anc}, \tilde{\mathbf{x}}_{neg}) + \alpha).\end{aligned} \tag{10}$$

The first three terms are imposed to learn to discriminate generated images from the real ones to improve the generation quality. The last two are used to distinguish the fake stylized images from the real image of the target style, so as to improve the quality of stylization.

3. EXPERIMENTS

3.1. Datasets

We evaluate our method on Planetscope reflectance data. The data contain 3 RGB channels and a near infrared channel,

Train	City	Munich	Moscow	Paris	Rome	Zurich
	#	2,836	2,981	2,997	2,869	2,312
Test	City	Yaounde	Djibouti	Niamey	Thamaga	Daressalaam
	#	853	283	361	141	2,228

Table 1: Number of patches for training domain and testing domain cities.

where only the RGB channels are used in our experiments. The data are of resolution 3m, and are collected from 5 European cities including Munich, Rome, Moscow, Paris and Zurich, and 5 African cities including Daressalaam, Djibouti, Yaounde, Thamaga and Niamey. To evaluate under a relatively large domain shift, we select the 5 European cities as the training domain while the African cities as the testing domain. All the data are splitted to patches with size 256×256 with a overlap of 128 pixels. The number of patches for each city are listed in Table 1.

3.2. Implementation Details

To train DADA, we use a shallow structure with four network blocks, each contains a 2D convolution, a instance normalization, a max pooling and a ReLU layer. The number of channels for these blocks are 256, 128, 64, and 32 respectively. During the training phase, the batch size is set to 8. The network is trained by a SGD optimizer with Nesterov acceleration. The momentum and weight decay are set to 0.9 and 5×10^4, respectively. The initial learning rate is set to 0.01 and a polynomial learning rate decay with power 0.95 is applied. The training lasts for $160,000$ iterations. The hyperparameter α in Eq. (6) and Eq. (10) is set to 0.3. The loss weight λ_1, λ_2 and λ_3 in Eq. (4) are set to $10, 10$ and 1 respectively.

For the downstream building footprint extraction task, a semantic segmentation network is trained, where a Unet [9] architecture with a ResNet50 [10] backbone is used. The optimizer, learning rate and batch size setting are the same as above. The training lasts for $200,000$ iterations.

3.3. Evaluation Metrics

We evaluate the performance of DADA by three metrics, including mean Intersection over Union (mIoU), F1 score, and Overall Accuracy (OA) of the building area. The results are reported by averaging the results from each test domain city.

3.4. Qualitative Results

We visualize the images generated by DADA in Fig. 2. As can be observed, images in the same column generally have similar appearances, while those in different columns looked different, which indicates that DADA can perform style transfer well between any pair of image.

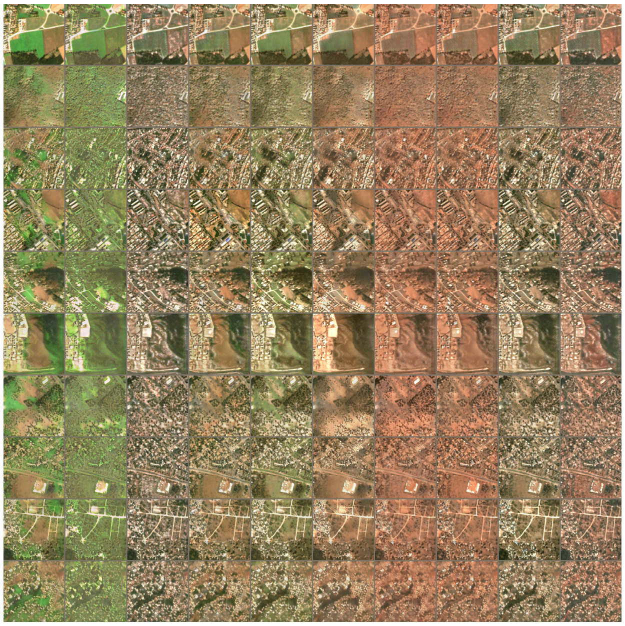

Fig. 2: Visualization of the images generated by DADA. Diagonal lines of the image matrix are the original images sampled from each city. The image located in row i and column j is generated from the original image in row i, with the style of the original image in row j.

Methods	mIoU	F1	OA
Baseline	20.0	32.4	74.9
Hist. Equ.	29.6	44.2	77.4
DADA	**30.6**	**45.5**	**79.2**

Table 2: Metrics (%) of different methods on the testing set. The best results are highlighted in **bold**. The results are reported by averaging the results from each test domain city.

3.5. Quantitative Results

We list the quantitative comparison results in Table 2. The results for a baseline method and a histogram equalization (Hist. Equ.) based method are reported for comparison. The baseline method here simply trains the semantic segmentation networks on the original image patches. Hist. Equ. trains and tests the networks on images normalized by histogram equalization. According to the results, both Hist. Equ. and DADA can improve the segmentation performance over the baseline. Besides, DADA can outperform Hist. Equ., which demonstrates its effectiveness.

4. CONCLUSION

This paper studies the domain shift problem occurred on different areas of the satellite images, and especially focuses on global range applications where it is hard to define the domain each image belongs to. We develop a novel domain-agnostic domain adaptation (DADA) method that can perform image-level style transfer between any pair of images without explicitly knowing the domain they come from. Comparative experiments on Planetscope datasets demonstrate the effectiveness of DADA both qualitatively and quantitatively.

5. ACKNOWLEDGEMENT

This work is supported by China Scholarship Council.

6. REFERENCES

[1] Gong Cheng and Junwei Han, “A survey on object detection in optical remote sensing images,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016.

[2] Zhitong Xiong, Yuan Yuan, and Qi Wang, “Ai-net: Attention inception neural networks for hyperspectral image classification,” in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 2647–2650.

[3] Yilei Shi, Qingyu Li, and Xiao Xiang Zhu, “Building segmentation through a gated graph convolutional neural network with deep structured feature embedding,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 184–197, 2020.

[4] Onur Tasar, SL Happy, Yuliya Tarabalka, and Pierre Alliez, “Semi2i: Semantically consistent image-to-image translation for domain adaptation of remote sensing data,” in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 1837–1840.

[5] Zhitong Xiong, Yuan Yuan, Nianhui Guo, and Qi Wang, “Variational context-deformable convnets for indoor scene parsing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Onur Tasar, Alain Giros, Yuliya Tarabalka, Pierre Alliez, and Sébastien Clerc, “Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1067–1081, 2020.

[7] Xun Huang and Serge Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.