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 to 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 is 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 applications.
- 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 $x_{\text{anc}}$, a positive patch $x_{\text{pos}}$, and a negative patch $x_{\text{neg}}$ will be sampled, in which $x_{\text{anc}}$ and $x_{\text{pos}}$ 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 $x_{\text{anc}}$, $x_{\text{pos}}$, $x_{\text{neg}}$, the fake negative image $x_{\text{anc}}$ and the fake anchor image $x_{\text{neg}}$ simultaneously to learn to perform style transfer.

are given as $S = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H \times W \times 3}$ denotes the image patch and $y_i$ its label, indicating whether each pixel corresponds to the building area or not. The target domain data are given as $T = \{(x_i, y_i)\}_{i=1}^{N}$. In contrary to $S$, the target domain labels $y_i$ 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 $x_i$ as $\Omega_{x_i}$. With such formulation, the building footprint extraction problem can be formulated as:

$$\min_h \sum_{(x, y) \in S} L_{\text{seg}}(x, G(x, \tilde{x}), y).$$  

(1)

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

$$L_{\text{seg}} = L_{\text{cse}}(h(x), y) + L_{\text{cse}}(h(G(x, \tilde{x})), y),$$  

(2)

where $L_{\text{cse}}$ 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 $L_{\text{cse}}$. The semantic segmentation networks are trained on the original images $x$ as well as the stylized images $G(x, \tilde{x})$. Since $\tilde{x}$ can be sampled from the target domain, the model’s generalizability can be improved by training on target-style source data $G(x, \tilde{x})$.

The generator $G$ is trained in an adversarial manner:

$$\min_{G, D} \sum_{(x, y) \in S \cup T} L_{\text{gen}} + L_{\text{dis}},$$  

(3)

where $D$ is the discriminator, $\pi$ is a sampling function that will be introduced in Section 2.2, and $L_{\text{gen}}$ and $L_{\text{dis}}$ 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 $x_{\text{anc}}$, a positive patch $x_{\text{pos}}$, and a negative patch $x_{\text{neg}}$ will be sampled from the union of source and target domain $S \cup T$ at each time: $(x_{\text{anc}}, x_{\text{pos}}, x_{\text{neg}}) = \pi(S \cup T)$. Here all of these three patches are randomly sampled from $S \cup T$, under the only restriction that $x_{\text{anc}}$ and $x_{\text{pos}}$ are neighboring patches, i.e., $x_{\text{pos}} \in \Omega_{x_{\text{anc}}}$.

2.3. Generator Loss $L_{\text{gen}}$

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

$$L_{\text{gen}} = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{cyc}} + \lambda_3 L_{\text{adv}}.$$  

(4)

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

$$L_{\text{rec}} = \sum_{x \in \{x_{\text{anc}}, x_{\text{neg}}\}} \|G_{\text{dec}}(G_{\text{enc}}(x)) - x\|_1,$$  

(5)

where $G_{\text{enc}}(\cdot)$ and $G_{\text{dec}}(\cdot, \cdot)$ are the encoder and decoder of the generator $G$. $G_{\text{dec}}(z_1, z_2)$ will first normalize $z_1$ with the style of $z_2$ by adaptive instance normalization (AdaIn) [7], and then decode $z_1$ to the size of the original image.
The adversarial loss for the generator is defined as:
\[
\mathcal{L}_{adv}^g = (D(\tilde{x}_{anc}) - 1)^2 + (D(\tilde{x}_{neg}) - 1)^2
+ \max(0, S(\tilde{x}_{anc}, x_{neg}) - S(\tilde{x}_{anc}, \tilde{x}_{neg}) + \alpha)
+ \max(0, S(\tilde{x}_{anc}, \tilde{x}_{neg}) - S(x_{anc}, \tilde{x}_{neg}) + \alpha).
\]  
(6)

Here \(\tilde{x}_{anc}\) (or \(\tilde{x}_{neg}\)) is the fake image generated from \(x_{anc}\) (or \(x_{neg}\)) with the style of \(x_{neg}\) (or \(x_{anc}\)) as illustrated in Fig. 1:
\[
\begin{align*}
\tilde{x}_{anc} &= G_{dec}(G_{enc}(x_{anc}), G_{enc}(x_{neg})), \\
\tilde{x}_{neg} &= G_{dec}(G_{enc}(x_{neg}), G_{enc}(x_{anc})).
\end{align*}
\]  
(7)

\(D(x)\) is the prediction of the discriminator, evaluating the probability that \(x\) is a real image rather than a fake one. \(S(x_1, x_2)\) is the similarity of \(x_1\) and \(x_2\) 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 \(\tilde{x}_{anc}\) and \(\til| Table 1: Number of patches for training domain and testing domain cities.

<table>
<thead>
<tr>
<th>Train City</th>
<th>Munich</th>
<th>Moscow</th>
<th>Paris</th>
<th>Rome</th>
<th>Zurich</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>2,836</td>
<td>2,981</td>
<td>2,997</td>
<td>2,869</td>
<td>2,312</td>
</tr>
<tr>
<td>Test City</td>
<td>Yaounde</td>
<td>Djibouti</td>
<td>Niamey</td>
<td>Thamaga</td>
<td>Daressalaam</td>
</tr>
<tr>
<td>#</td>
<td>853</td>
<td>283</td>
<td>361</td>
<td>141</td>
<td>2,228</td>
</tr>
</tbody>
</table>

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 \(\times\) 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 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 \(\times\) 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 \(\alpha\) in Eq. (6) and Eq. (10) is set to 0.3. The loss weight \(\lambda_1, \lambda_2\) and \(\lambda_3\) 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.

3. EXPERIMENTS

3.1. Datasets

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>mIoU</th>
<th>F1</th>
<th>OA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>20.0</td>
<td>32.4</td>
<td>74.9</td>
</tr>
<tr>
<td>Hist. Equ.</td>
<td>29.6</td>
<td>44.2</td>
<td>77.4</td>
</tr>
<tr>
<td>DADA</td>
<td>30.6</td>
<td>45.5</td>
<td>79.2</td>
</tr>
</tbody>
</table>

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.

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


