Unsupervised Domain Adaptation Augmented by Mutually Boosted Attention for Semantic Segmentation of VHR Remote Sensing Images

Xianping Ma, Xiaokang Zhang, Zhiguo Wang and Man-On Pun

Abstract—This work investigates Unsupervised Domain Adaptation (UDA)-based semantic segmentation of Very High Resolution (VHR) remote sensing images from different domains. Most existing UDA methods resort to Generative Adversarial Networks (GAN) to cope with the domain shift problem caused by the discrepancies across different domains. However, these GAN-based UDA methods directly align two domains in the appearance, latent, or output space based on convolutional neural networks, making them ineffective in exploiting long-range dependencies across the high-level feature maps derived from different domains. Unfortunately, such high-level features play an essential role in characterizing remote sensing images with complex content. To circumvent this obstacle, Mutually Boosted Attention Transformer (MBATrans) is proposed to capture cross-domain dependencies of semantic feature representations in this work. Compared with conventional UDA methods, MBATrans can significantly reduce domain discrepancies by capturing transferable features using global attention. More specifically, MBATrans uses a Mutually Boosted Attention (MBA) module to align cross-domain feature maps while enhancing domain-general features. Furthermore, a novel GAN-based network with improved discriminative capability is devised by integrating an additional discriminator to learn domain-specific features. Extensive experiments on two large-scale VHR remote sensing datasets, namely ISPRS Potsdam and Vaihingen, confirm the superior performance of the proposed MBATrans-Augmented-GAN (MBATA-GAN) architecture. The source code in this work is available at https://github.com/sstary/SSRS.

Index Terms—Unsupervised Domain Adaptation, remote sensing image, Mutually Boosted Attention, Generative Adversarial Networks

I. INTRODUCTION

Semantic segmentation is performed to categorize each pixel of an image into pre-defined classes, which is of great significance in many Remote Sensing (RS) applications, such as infrastructure planning, land planning, and urban area change detection. In the literature, many supervised semantic segmentation methods have been reported by exploiting fully labeled datasets [1, 2, 3, 4] while semi-supervised methods were explored on part of labeled datasets [5, 6]. However, these supervised and semi-supervised methods require manual labeling efforts in real-life applications. To circumvent this problem, Unsupervised Domain Adaption (UDA) methods were proposed by leveraging existing data and labels from one source domain to tackle segmentation tasks in a new target domain. In other words, the segmentation model in UDA is trained with labeled source images while being tasked to classify images from the target domain, which makes UDA particularly attractive as no new labels from the target domain are required. However, most UDA methods encounter the domain shift problem in which features between the source and the target domains are mismatched [7].

In the research area of Computer Vision (CV), many efforts [8, 9] have been devoted to alleviating the domain shift problem. Driven by the recent development of Deep Learning (DL) techniques, [10, 11, 12, 13] have achieved improved generalization and transferability byexploiting explanatory factors and sophisticated strategies to align feature distributions across domains, such as discrepancy-based [10], reconstruction-based [11] and adversarial-based [12, 13, 14, 15]. In particular, the seminal work [12] has inspired substantial research efforts on solving the domain shift problem using adversarial learning [16]. In contrast to the image classification tasks considered in [12, 14], semantic segmentation has to focus on the global visual information while encoding diverse visual information such as appearance, shape and context, which makes the alignment of high-level features more challenging. To address the UDA problem in semantic segmentation, AdasegNet [13] was proposed to adaptively learn a structured output space in which pixel-level classification tasks in both source and target domains share a significant amount of similarities in spatial layout and local context. Furthermore, [15, 17] explored the transferability and discriminability of features during domain adaptation. Despite their good performance in cross-domain semantic segmentation, these aforementioned domain adaptive methods were designed for natural image datasets. As
compared to natural images discussed in the UDA literature, remote sensing images exhibit much larger cross-domain discrepancies in terms of color, texture, spatial resolution, and scene content [18]. These discrepancies were attributed to the following two facts: First, the spectral heterogeneity of remote sensing images is higher than that of natural images acquired in the RGB bands [19, 20]. Furthermore, the structure information of remote sensing images is richer while the object scale changes more significantly [21]. As a result, the performance of these UDA methods is less satisfactory when directly applied to VHR remote sensing images of large variations in camera angles, image resolution, and so on. In particular, as remote sensing sensors have grown substantially in terms of both quantity and variety over the past decade, the availability of more high-quality VHR data has further widened the discrepancies between different domains.

Recently, some DL-based UDA methods have been reported for remote sensing tasks [22, 23, 24, 25, 26]. For instance, GANAI [22] was the first GAN-based UDA algorithm designed for semantic segmentation of aerial imagery whereas TriADA [23] introduced a class-aware self-training technology to enhance its discriminator. WTIC [27] further investigated the invariant semantic features across VHR remote sensing images and proposed a dynamical training strategy by incorporating multiple weakly-supervised constraints. Recently, subspace alignment [24, 28, 29] equipped with CNN-based shallow feature extractors was devised to explore and align the rich semantic vector subspace in which the complex image content can be further represented and analyzed. In particular, CCAGAN [18] employed a global and category-level adversarial loss to enhance local semantic consistency by employing negative transfer during domain adaptation and a Category-Certainty Attention (CCA) module used to process deep features. Despite the remarkable performance improvement demonstrated by these pioneering works, it remains very challenging to narrow the performance gap between these UDA models and their supervised counterparts in remote sensing. One possible reason is that most existing methods employ the Convolutional Neural Networks (CNN) as a generator to extract high-level features using a deep architecture before aligning the source and the target domains with a discriminator. However, it is well-known that the high-level features obtained by CNN suffer from the inductive bias of the local receptive fields and limited context modeling [30]. Furthermore, it is observed that the strategy of only aligning the source and the target domains using the CNN-based framework cannot fully utilize enriched high-level features in semantic subspace, which in turn limits the learning of transferable features.

To overcome the aforementioned challenges, some pioneering works were reported for various image classification tasks by exploiting the transformer architecture and multi-head self-attention [31, 32, 33]. More specifically, the cross-attention mechanism is designed to effectively exchange information of different scales [34] or different modalities [35] from a global perspective in a supervised manner. To take advantage of these attention mechanisms in UDA, CDTrans, a transformer-based structure with three weight-sharing branches was proposed in [36] to apply self-attention and cross-attention for source-to-target feature learning and cross-domain alignment. Built upon CDTrans, BCAT was proposed in [37] to perform the cross-attention computation in the source domain, which ingeniously combines self-attention and cross-attention while expanding them into a bidirectional four-branch architecture. By learning augmented feature representations for both the source and the target domains, BCAT demonstrated excellent performance on image classification tasks.

Inspired by CDTrans and BCAT, this work proposes Mutually Boosted Attention Transformer (MBATrans) by combining cross-attention and self-attention mechanisms to learn transferable features in UDA for semantic segmentation of VHR remote sensing. The proposed MBATrans can capture high-level intra- and inter-domain feature maps with rich semantic information by taking into account long-range dependencies among elements using its global receptive field. The term Mutually is coined for the fact that information from the source and the target domains is jointly processed by the proposed generator as shown in Fig. 1. In addition, the proposed MBATrans simultaneously exploits both self-attention and cross-attention in both the source and target domains. It is worth noting that the MBATrans mechanism proposed in this work utilizes both cross-attention and self-attention for UDA, which distincts MBATrans from our previous work in [35] that uses cross-attention only. Furthermore, the proposed MBATrans employs the same $Q, K, V$ matrices [33] for both self-attention and cross-attention computation. Finally, this work proposes a self-attention based transformer as discriminator to capture more discriminative information as discriminability often depends on more local and specific details [15]. The synergy of MBATrans and enhanced discriminators explicitly enforce the GAN-based method to learn domain-invariant and domain-specific representations simultaneously. It should be emphasized that such a design can be integrated into other domain adaptation frameworks in a straightforward manner.

The contributions of this work are fourfold as summarized in the following:

- MBATrans is proposed to align the high-level feature maps of the cross-domain VHR remote sensing images. By exploiting its global perception field to model long-range

![Fig. 1. Illustration of two UDA training approaches: (a) Alternating training: Models are first trained on data and labels from the source domain before being trained on data from the target domain in an iterative manner, which incurs high model volatility. (b) Cross-domain training: Models are trained with data and labels from the source domain and data from the target domain simultaneously.](image-url)
dependencies across different domains, the proposed segmentation network can learn more transferable features of ground objects. In the sequel, the segmentation network is also referred to as the generator, following the convention commonly used in GAN-based networks;

- To further utilize the highly aligned transferable features derived by MBATrans, a Multi-Discriminator Module (MDM) is developed and integrated into our proposed GAN-based network;

- Capitalizing on the MBATrans and discriminators, a novel GAN-based network called MBATrans-Augmented-GAN (MBATA-GAN) is proposed to overcome the domain shift problem by exploiting the enriched high-level features of VHR remote sensing images. Transferable and discriminative features across domains are learned by the fully enhanced GAN-based framework;

- Extensive experiments are performed on two VHR remote sensing datasets, namely Potsdam and Vaihingen, to confirm the superior performance achieved by MBATA-GAN.

The remainder of this paper is organized as follows. Sec. II introduces the architecture and modeling method of the proposed MBATrans and MBATA-GAN whereas Sec. III provides details on the experiment setup, datasets, comparing methods and analyses on the experimental results. Finally, the conclusion is given in Sec. IV.

II. METHODOLOGY

In this work, a transformer-based architecture equipped with both cross-attention and self-attention mechanisms is proposed to align high-level feature maps derived from CNN-based modules. To enhance the discriminative capability of the proposed architecture, multiple feature discriminators are integrated into the GAN framework. In the following, we will first provide an overview of the proposed MBATA-GAN before elaborating on its detailed components, including MBATrans and MDM. Finally, we devise the loss functions employed in the proposed architecture.

A. MBATA-GAN

Given an unsupervised domain adaptation problem, we denote by \( D_s = \{ (x_s, y_s) \}^{n_s} \) and \( D_t = \{ x_t \}^{n_t} \) a labeled source dataset and an unlabeled target dataset, respectively, where \( x_s, y_s \) and \( x_t \) are the sample in the source domain, its corresponding label and the sample in the target domain, respectively. Furthermore, \( n_s \) and \( n_t \) denote the number of samples in the source domain and the target domain, respectively. Due to the domain shift problem, the source and the target domains follow different marginal and conditional distributions, i.e., \( p_s(x_s) \neq p_t(x_t) \). However, it should be emphasized that both domains share an identical label space, i.e., \( Y_s = Y_t \) [37]. The goal of the UDA task is to train a generator that extracts high-level transferable features from the source domain \( D_s \) to help the classification performed in the target domain \( D_t \).

In this work, VHR remote sensing images of Red-Green-Blue (RGB) or InfraRed-Red-Green (IRRG) are employed as the input to the proposed MBATA-GAN. More specifically, we denote by \( X_s \in \mathbb{R}^{H \times W \times 3} \) and \( X_t \in \mathbb{R}^{H \times W \times 3} \) the training batches from the source and the target domains, respectively.
where $H$ is the height of the image and $W$ the width. As shown in Fig. 2, the proposed MBATA-GAN consists of a generator and multiple discriminators. The generator, i.e. the segmentation network, is comprised of three modules, namely a CNN-based feature extractor $\mathcal{F}$, an MBATrans module and two classifiers (Low-level and High-level). The multi-level strategy from AdasegNet [13] is adopted to better capture the lower-level features that are difficult to model as they are far away from the high-level output labels. Given input $\{X_s, X_t\}$ to the generator, the feature extractor first extracts multi-level features from each domain using the convolution operation. Note that the feature extractor in AdasegNet is a modified DeepLab-v2 [38] framework with ResNet-101 [39] model pre-trained on ImageNet [40]. We denote by $F^h_s \in \mathbb{R}^{D \times H \times W}$ and $F^h_t \in \mathbb{R}^{D \times H \times W}$ the resulting high-level feature maps with abstract semantic information for the source and the target domains, respectively. $F^f_s$ and $F^f_t$ are then input into the proposed MBATrans for feature alignment improvement by simultaneously exploiting cross-attention and self-attention. The outputs of the proposed MBATrans are high-level aligned feature maps denoted as $\tilde{F}_s$ and $\tilde{F}_t$. These feature maps are further processed by a high-level classifier. In contrast, the low-level feature maps from the feature extractor $F^f_s \in \mathbb{R}^{D' \times H' \times W'}$ and $F^f_t \in \mathbb{R}^{D' \times H' \times W'}$ are directly input into a low-level classifier without being processed by MBATrans. These low-level feature maps are rich in local detailed information, providing reliable support to accurate semantic prediction. Similar to AdasegNet, the Atrous Spatial Pyramid Pooling (ASPP) [38] equipped with an up-sampling layer is employed as both the high- and low-level classifiers. ASPP exploits multi-scale features by employing multiple parallel filters with different rates. Finally, the up-sampling layer along with the softmax output is applied to match the size of the input image. In our model, the numbers of input channels for the high-level and the low-level classifier are $D$ and $\frac{D}{2}$, respectively. This multi-level design can better preserve the characteristics of low-level features. The high-level aligned feature maps $\tilde{F}_s$ and $\tilde{F}_t$ and the outputs of the two classifiers denoted as $\{h_s, P^h_s\}$ and $\{h_t, P^h_t\}$ of size $\mathbb{R}^{H \times W \times C}$ are then fed into three discriminators, where $C$ is the number of categories. The segmentation loss $L_{seg}$ is derived by exploiting the source dataset with labels. In addition, a loss function $L_{global}$ is defined to reduce the discrepancy between the high-level aligned feature maps from both domains. Finally, $L_{global}$ is computed with the Maximum Mean Discrepancy [41] by utilizing the average pool of $\tilde{F}_s$ and $\tilde{F}_t$. More details about MBATrans will be elaborated in Sec. II-B.

In the proposed Multi-Discriminator Module (MDM), the feature discriminator denoted as $D^f$ is designed to handle high-level aligned features $\tilde{F}_s$ and $\tilde{F}_t$. The outputs of $D^f$ are used to compute a cross-entropy loss $L^f_{ce}$ in the Generator training stage and an adversarial loss $L^f_{adv}$ in the MDM training stage. Furthermore, the outputs of two discriminators denoted as $D^h$ and $D^f$ in different training stages are used to compute a cross-entropy loss $L^h_{ce}$ and adversarial loss $L^h_{adv}$. $D^h$ and $D^f$ are built upon a self-attention based transformer to extract comprehensive discriminative information after dimensionality reduction achieved by CNN.

Based on these loss functions, the GAN-based network is optimized by the following max-min criterion:

$$\max_{G} \min_{\mathcal{D}} L(X_s, X_t),$$

where $\mathcal{D}$ and $\mathcal{G}$ denote the Generator and MDM, respectively. The ultimate goal is to minimize the segmentation loss while maximizing the probability that the target predictions come from the source domain. More details about the loss functions will be elaborated in Sec. II-D.

**B. MBATrans**

MBATrans is designed to alleviate the domain shift problem through modeling the global information simultaneously across different domain feature maps by exploiting the synergy of the cross-attention and self-attention mechanisms in the proposed Mutually Boosted Attention blocks. Specifically, the cross-domain feature interactions between each pair of elements are exploited to establish the similarities between domains.

Fig. 3. The structure of the proposed MBATrans.

As illustrated in Fig. 3, the proposed MBATrans consists of an embedding layer, $N$ blocks of the MBA modules and a reconstruction layer. Given the input feature maps $F^h_s$ and $F^h_t \in \mathbb{R}^{D \times H \times W}$ where $D = 2048$ is the embedding dimension of the last layer of $\mathcal{F}$, the embedding layer first flattens the feature maps into sequences of 2D patches. After a convolutional layer performs the reshaping operation with both kernel and stride sizes of two, the extracted features are tokenized to $F_s$ and $F_t \in \mathbb{R}^{Z \times D}$, where $Z = 256$ is the fixed size of 2D patches. Furthermore, tokens $F_{2s}$ and $F_{2t}$ are utilized to learn cross-domain features using the cross-attention mechanism. More specifically, $F_{2s}$ is designed to represent the features derived from the source domain with the target domain being the assisting domain. Similarly, $F_{2t}$ denotes the features derived from the target domain with the source domain being the assisting domain. In our experiments, $F_{2s}$ and $F_{2t}$ are initialized with values equal to $F_s$ and $F_t$, respectively. Next, the four tokens are fed into the MBA modules
followed by a Multi-Layer Reconstruction (MLP) as shown in Fig. 4. As previously discussed, the MBA module is designed for the task of domain adaptation by calculating both self-attention and cross-attention in only one unique transformer. To fully benefit from the multi-head attention, cross-domain tokens are split into four non-overlapping copies denoted as $F_{si} \in \mathbb{R}^{Z \times \frac{D}{4}}$ and $F_{ti} \in \mathbb{R}^{Z \times \frac{D}{4}}$ where $i = 1, 2, 3, 4$. Each head of the MBA module consists of three groups of linear matrices $W_{Q_i}, W_{K_i},$ and $W_{V_i} \in \mathbb{R}^{\frac{D}{2} \times \frac{D}{2}}$. Since the operation of each head is identical, we denote $F_s$ and $F_t$ as the input to one head of the MBA module by omitting the subscripts $i$ for the sake of notational brevity. Thus, two groups of matrices $\{Q_s, K_s, V_s\}$ and $\{Q_t, K_t, V_t\}$ for each head can be computed as follows:

$$Q_s = F_s W_Q, K_s = F_s W_K, V_s = F_s W_V,$$  \hspace{1cm} (2)

$$Q_t = F_t W_Q, K_t = F_t W_K, V_t = F_t W_V.$$  \hspace{1cm} (3)

After the computation of the matrices above, two self-similarity matrices denoted as $S_s$ and $S_t$ and two cross-similarity matrices denoted as $S_{s2t}$ and $S_{t2s}$ are further derived. $S_s$ and $S_t$ access the information of each element on the respective domain feature map. Furthermore, $S_{s2t}$ exploits every element $K_s$ in the source domain using the target query $Q_t$, whereas $S_{t2s}$ explores every element $K_t$ in the target domain using the source query $Q_s$. By exploiting these self- and cross-similarity matrices, the self- and cross-attention matrices can be computed as:

$$A_{s2t} = S_{s2t} V_s^T = \varphi \left( \frac{Q_s^T K_s}{\sqrt{d}} \right) V_s^T,$$  \hspace{1cm} (4)

$$A_s = S_s V_s^T = \varphi \left( \frac{Q_s^T K_s}{\sqrt{d}} \right) V_s^T,$$  \hspace{1cm} (5)

$$A_t = S_t V_t^T = \varphi \left( \frac{Q_t^T K_t}{\sqrt{d}} \right) V_t^T,$$  \hspace{1cm} (6)

$$A_{t2s} = S_{t2s} V_t^T = \varphi \left( \frac{Q_t^T K_t}{\sqrt{d}} \right) V_t^T,$$  \hspace{1cm} (7)

where $d = D/4$ is the channel dimension for each head. Furthermore, $\varphi(\cdot)$ and $(\cdot)^T$ are the softmax function and the matrix transpose operator, respectively. Note that $A_{s2t}$ is designed to transfer information from the source domain to the target domain. By exploiting $\{A_{s2t}, F_{s2t}\}$ and $\{A_{t2s}, F_{t2s}\}$, the proposed MBATrans can fuse information from different domains while learning more transferable features from one domain to another. The fused feature maps are fed into an MLP that is a two-layer fully connected neural network designed to restore the perception of full channel information. In the MBA module, features from two different domains are integrated through fusing information of each point on both feature maps over all channel dimensions. This fusion process is repeated for $N$ stages via stacking MBA modules in which $F_{2s}, F_s, F_t$ and $F_{2t}$ derived by the previous MBA module are further enhanced in the next stage. Furthermore, normalization layers are employed throughout MBATrans while a dropout layer is used after the attention module, which is consistent with the design of other existing methods [31, 34, 37]. Finally, the outputs from MLP are synthesized after the reconstruction.
layer, generating \( \mathbf{F}_s \) and \( \mathbf{F}_t \) of the same size as the input, i.e. \( \mathbb{R}^{D \times \frac{H}{4} \times \frac{W}{4}} \). These augmented representations characterize cross-domain aligned features empowered with improved similarity and transferable features derived during the training process. It is worth mentioning that the proposed MBATrans is such flexibly designed that the structure can be easily adopted by most existing generators with only minor changes.

### C. Multi-Discriminator Module (MDM)

In the proposed network model, an MDM is employed to enhance the adaptation performance. The three discriminators are denoted as \( D^s \), \( D^h \) and \( D^f \), respectively. This multi-discriminator module enables the proposed model to extract highly domain-specific features.

The structures of the proposed \( D^s \) and \( D^h \) are identical. As illustrated in Fig. 5 (a), the discriminators are endowed with \( N \) self-attention modules followed by a reconstruction network and a TailConv. Furthermore, each discriminator receives only one input from either the source domain or the target domain at a time as indicated in Fig. 5 (a). As shown in Fig. 5 (b), the proposed self-attention module utilizes its multi-head design to align the features in both the source or the target domains. Denoted by \( A \), the value of self-attention on the label space can be computed as follows:

\[
A = SV^T = \varphi \left( \frac{Q^T K}{\sqrt{d}} \right) V. \tag{8}
\]

In addition to \( D^s \) and \( D^h \), a third discriminator \( D^f \) is designed to further exploit the aligned high-level feature maps \( \mathbf{F}_s \) and \( \mathbf{F}_t \) generated by MBATrans. As shown in Fig. 6, \( D^f \) consists of multiple convolutional layers and a TailConv.

In summary, the proposed MDM can better utilize diverse information from the generator by exploiting multiple discriminators for adversarial learning. It should be emphasized that all discriminators proposed only process input from either the source domain or the target domain at a time. The Leaky ReLU activation connected after each convolutional layer is omitted in this figure.

### D. Training of Generator and Discriminators

To fully exploit the proposed MBATA-GAN model, four types of loss functions are designed to guide the training of the generator and the discriminators shown in Fig. 2, namely the discrepancy loss \( L_{\text{global}} \), the segmentation loss \( L_{\text{seg}} \), the cross-entropy loss \( \{ L_{\text{ce}}, L_{\text{ce}}^h \} \) and the classification adversarial loss \( \{ L_{\text{adv}}, L_{\text{adv}}^h \} \). Table I summarizes all the loss functions used with information about whether the functions are used for the training of the generator or the discriminators as well as the domain from which the training data comes.

**Generator:** The objective functions, namely \( L_{\text{global}}, L_{\text{seg}}, L_{\text{ce}} \) and \( L_{\text{ce}}^h \) are used to optimize the generator’s parameters. In this work, the following Maximum Mean Discrepancy (MMD) loss [41] is utilized to globally align the enhanced high-level features derived by MBATrans from the source and the target domains due to its high computational efficiency.

\[
L_{\text{global}} \left( \mathbf{F}_s, \mathbf{F}_t \right) = \lambda^g D \sum_{d=1}^{D} \left[ \mathbb{P}(\mathbf{F}_s^d) - \mathbb{P}(\mathbf{F}_t^d) \right], \tag{9}
\]

where \( \lambda^g \) is a weighting coefficient, \( D \) is the number of channels while \( \mathbb{P} \) denotes the average pool to gather spatial

---

**Table I**

The details about loss functions used to train the generator and the discriminators.

<table>
<thead>
<tr>
<th>Loss Func.</th>
<th>Generator</th>
<th>Discriminators</th>
<th>Data from the Source Domain</th>
<th>Data from the Target Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L_{\text{global}} )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( L_{\text{seg}} )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( L_{\text{ce}} )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( L_{\text{ce}}^h )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( L_{\text{adv}} )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>( L_{\text{adv}}^h )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

---

![Fig. 5. (a) The discriminator \( D^s \) where \( s \in \{ \ell, h \} \). (b) The self-attention module. \( D^s \) and \( D^h \) have identical structure with non-sharing weights. Each discriminator receives only one input from the source domain or the target domain at a time. The Leaky ReLU activation connected after each convolutional layer is omitted in this figure.](image)

![Fig. 6. The details of the discriminator \( D^f \). It receives only one input from either the source domain or the target domain at a time. The Leaky ReLU activation connected after each convolutional layer is omitted in this figure.](image)
information for each channel $d$. The minimization of this global loss function leads to efficient alignment of the feature maps of the source and the target domains. Notably, MMD is used to assist the training of MBATrans and the low-level features can be effectively learned by the multi-level discriminators.

Next, data samples and their corresponding labels from the source domain are employed to compute the following cross entropy-based segmentation loss for the generator to learn the basic features:

$$L_{\text{seg}}(P_s^t, P_s^h) = -\sum_{H, W, C} Y_s^{(H, W, C)} \left[ \lambda^s \log (P_s^{(H, W, C)}) + \lambda^h \log (P_s^{(H, W, C)}) \right],$$

(10)

where $Y_s$ stands for the labels of the source images while $\lambda^s$ and $\lambda^h$ are two weighting coefficients to balance the contributions from the classifications derived from the low-level and high-level features.

After that, feature maps derived from the target domain $\tilde{F}_t$ are used to calculate the feature cross-entropy loss as follows:

$$L_{\text{ce}}^{f} (\tilde{F}_t) = -\sum_{H, W} \lambda^f \log \left( D^{f}(\tilde{F}_t)^{(H, W, 0)} \right),$$

(11)

where $\lambda^f$ is a weighting coefficient. We intentionally set the label to 0 in the superscript $(H, W, 0)$ to mislead the discriminator. As a result, the discriminator will perceive the target classification as the source classification for each element $(H, W)$. Similarly, the classification cross-entropy loss can be computed as:

$$L_{\text{ce}}^{th} (P_t^t, P_t^h) = -\sum_{H, W} \left[ \lambda^t \log (D^{t}(P_t^t)^{(H, W, 0)}) + \lambda^h \log (D^{t}(P_t^h)^{(H, W, 0)}) \right],$$

(12)

where $\lambda^t$ and $\lambda^h$ are two weighting coefficients. $P_t^t$ and $P_t^h$ are the low-level and high-level classifications derived by the generator, respectively.

Summarizing Eqs. (9) - (12), we have the overall objective function used in training the generator as:

$$L_{G} = L_{\text{seg}} + L_{\text{global}} + L_{\text{ce}}^{f} + L_{\text{ce}}^{th},$$

(13)

**Discriminators:** The discriminators are trained to distinguish from which domain the information comes. In particular, the feature discriminator $D^{f}$ is optimized with the feature adversarial loss $L_{\text{adv}}^{f}$. The high-level feature maps from the source domain and the target domain are used to calculate the feature adversarial loss as follows:

$$L_{\text{adv}}^{f}(F_s, \tilde{F}_t) = -\sum_{H, W} \left[ \log(D^{f}(\tilde{F}_t)^{(H, W, 0)}) + \log(D^{f}(\tilde{F}_t)^{(H, W, 1)}) \right],$$

(14)

where we assign labels of 0 and 1 to each element $(H, W)$ of images from the source domain and the target domain, respectively. This design aims to improve the discriminator’s capability of correctly identifying the domain that each input image belongs to.

Finally, the low-level discriminator $D^{l}$ and the high-level discriminator $D^{h}$ are optimized with the following classification adversarial loss functions $L_{\text{adv}}^{l}$ and $L_{\text{adv}}^{h}$, respectively.

$$L_{\text{adv}}^{l} = -\sum_{H, W} \left[ \log(D^{l}(P_s^t)^{(H, W, 0)}) + \log(D^{l}(P_s^h)^{(H, W, 1)}) \right],$$

(15)

$$L_{\text{adv}}^{h} = -\sum_{H, W} \left[ \log(D^{h}(P_s^t)^{(H, W, 0)}) + \log(D^{h}(P_s^h)^{(H, W, 1)}) \right],$$

(16)

Summarizing Eqs. (14) - (16), we have the overall objective function used in training the generator as:

$$L_{D} = L_{\text{adv}}^{l} + L_{\text{adv}}^{f} + L_{\text{adv}}^{h},$$

(17)

It is worth noting that the training of the generator and the training of the discriminators are conducted in an alternating fashion as reported in the GAN literature, i.e. the parameters of all discriminators are first held constant while the generator is trained with Eq. (13). After that, the parameters of the generator are held constant while the discriminators are trained with Eq. (17).

**III. Experiments And Discussions**

**A. Datasets Description**

In this section, adaptation experiments are conducted with two VHR datasets belonging to the ISPRS 2D semantic segmentation benchmark datasets, namely Potsdam and Vaihingen [42]. Images in both datasets were obtained from aerial photography. However, the Ground Sampling Distance (GSD) of Potsdam is 5 cm and the GSD of Vaihingen is 9 cm. In addition, these two datasets possess different channel information. More specifically, the Potsdam dataset contains four bands of information, namely InfraRed, Red, Green and Blue channels (IRRGB). In contrast, the Vaihingen dataset only contains three bands, namely InfraRed, Red and Green channels (IRR). In addition, the difference in geographical locations results in great discrepancies in terms of ground object categories, such as building types and vegetation forms. These discrepancies discussed above represent a great challenge for the proposed UDA method.

The Potsdam dataset contains 24 VRH True Orthophotos of size 6000 x 6000. These 24 orthophotos are divided into a training set of 18 patches and a test set of 6 patches whose IDs are as follows:

- The training set (18 patches): \{6_10, 7_10, 2_12, 3_11, 2_10, 7_8, 5_10, 3_12, 5_12, 7_11, 7_9, 6_9, 7_7, 4_12, 6_8, 6_12, 6_7, 4_11\};
- The test set (6 patches): \{2_11, 3_10, 4_10, 5_11, 6_11, 7_12\}.

In contrast, the Vaihingen dataset includes 16 VHR True Orthophotos of size 2500 x 2000 pixels. These 16 orthophotos are divided into a training set of 12 patches and a test set of 4 patches. The training set contains the orthophotos indexed...
by \{1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34, 37\} and the test set \{5, 21, 15, 30\}.

The data loader developed from vFuseNet is adopted in our work. All methods including our model were trained for 50 epochs each of which contains 1000 batches. During the training process, the models are evaluated every 100 batches before their performance was recorded. Finally, the optimal model corresponding to the best performance was recorded for each method. It is worth mentioning that the optimal model usually appeared before the final epoch.

Fig. 7 illustrates some sample images from Potsdam and Vaihingen. Visual inspection of Fig. 7 suggests clear differences between the two datasets. Indeed, the images of Potsdam are of much larger sizes and more channels as compared to those of Vaihingen. Thus, we divided the Potsdam dataset into two subsets, namely “Potsdam IRRG” and “Potsdam RGB” while forming one dataset with Vaihingen denoted as “Vaihingen IRRG”. Unsupervised domain adaptation tasks are designed by treating these three subsets of data as the source and the target domain. More details about the UDA tasks are provided in Sec. III-E.

B. Evaluation Metrics

In the sequel, the Overall Accuracy (OA), the mean Intersection over Union (mIoU), and the mean F1 score (mF1) are used to quantify the segmentation performance of the cross-domain VHR remote sensing images. These standard statistical indices provide a fair comparison on the performance of the proposed MBATA-GAN and other existing works. More specifically, we compute mF1 and mIoU of the following five ground classes, namely Building, Tree, Low Vegetation (Low Veg), Car, Impervious Surface (Imp Suf.), following the common practice reported in the literature [23, 43]. In addition, we also include Clutter or Background in the evaluation of OA whose definition is given by:

\[
OA = \frac{TP}{TP + FP},
\]

where \(TP\) and \(FP\) are the sum of values on and outside the diagonal of the confusion matrix, respectively. Furthermore, F1 and IoU are calculated for each category \(c\) according to the following formula:

\[
F1 = 2 \times \frac{Q_c \times R_c}{Q_c + R_c},
\]

and

\[
\text{IoU} = \frac{TP_c}{TP_c + FP_c + FN_c},
\]

where \(TP_c\), \(FP_c\), \(TN_c\) and \(FN_c\) are true positives, false positives, true negatives, and false negatives for the \(c\)-th class, respectively. Finally, \(Q_c\) and \(R_c\) are given by:

\[
Q_c = \frac{TP_c}{TP_c + FN_c},
\]

\[
R_c = \frac{TP_c}{TP_c + FP_c}.
\]

Upon obtaining F1 and IoU for the five categories according to the definitions above, we will then derive their means, namely mF1 and mIoU.

Finally, the computational complexity of the proposed MBATA-GAN is analyzed using the following evaluation metrics. First, the floating point operation count (FLOPs) is used to evaluate the complexity. Furthermore, the number of model parameters (M) and the memory footprint (MB) are used to evaluate the memory requirement. Finally, the frames per second (FPS) is used to evaluate the execution speed. Ideally, a computationally efficient model should have smaller values in the first three metrics but a larger FPS value.

C. Experimental settings

In the following experiments, the DeepLab-v2 framework [38] with ResNet-101 [39] is adopted as our multi-level feature extractor. It was pre-trained on ImageNet [40]. Other details about the feature extractor can be found in AdasegNet [13]. The classifiers consist of four branches using \(3 \times 3\) kernels but different atrous rates \{6, 12, 18, 24\} in order to capture objects of different sizes. The channel number of all branches is set to 6, i.e. the total categories of semantic segmentation labels. MBATrans collects high-level features of both domains before aligning them by cross-attention. From the perspective of model development, it is a concise single branch transformer with four stacked MBA modules and four heads per MBA module. An average pooling is used to compute the discrepancy between the source domain and the target domain. The design of \(D^d\) and \(D^h\) follows AdasegNet, and a self-attention based transformer of four stacks is employed with four heads for each attention module. The TailConv in both \(D^d\) and \(D^h\) is the last convolution layer with a kernel size of \(4 \times 4\), a stride of 2 and a channel number of 1, which is designed to determine the domain which each input belongs to. \(D^f\) consists of four convolution layers with a kernel size of \(1 \times 1\) and stride 1. The channel number varies over \{2048, 1024, 512, 1\}. Each convolution layer except the last one follows a Leaky-ReLU [44] activation parameterized by 0.2.

The experiments are implemented with PyTorch on a single NVIDIA GeForce RTX 3090 GPU with 24-GB RAM. To
efficiently read the large images, a sliding window is used to dynamically collect the training batches. The size of the sliding window was set to 256 × 256 with no overlapping during the training stage, i.e. the stride in the training stage was set to 256. The upper limit of the batch size allowed by our GPU was 10. The generator including the feature extractor and the MBATrans were trained with the Stochastic Gradient Descent (SGD) optimizer and the Nesterov acceleration. For the MDM training, the Adam optimizer was utilized. Some key training parameters are listed in the following Table II. Finally, while \( \lambda^{\ell_s}, \lambda^{h_s}, \lambda^{f} \) and \( \lambda^{b} \) used the default values of AdasegNet, the values of \( \lambda^{g} \) and \( \lambda^{f} \) were chosen based on our sensitivity analysis as shown in Sec. III-E.

### Table II

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD momentum</td>
<td>0.9</td>
</tr>
<tr>
<td>SGD weight decay</td>
<td>( 5 \times 10^{-4} )</td>
</tr>
<tr>
<td>SGD learning rate</td>
<td>( 2.5 \times 10^{-4} )</td>
</tr>
<tr>
<td>Adam momentum</td>
<td>0.9</td>
</tr>
<tr>
<td>Adam learning rate</td>
<td>( 10^{-4} )</td>
</tr>
<tr>
<td>( \lambda^{g} )</td>
<td>2.0</td>
</tr>
<tr>
<td>( \lambda^{\ell_s}, \lambda^{h_s} )</td>
<td>[0.1, 1]</td>
</tr>
<tr>
<td>( \lambda^{f} )</td>
<td>0.005</td>
</tr>
<tr>
<td>( \lambda^{f}, \lambda^{b} )</td>
<td>[0.0002, 0.001]</td>
</tr>
</tbody>
</table>

#### D. Benchmarking

The proposed MBATA-GAN is benchmarked against six existing UDA methods, namely DANN [12], AdasegNet [13], Advent [45], GANAI [22], TriADA [23] and CCAGAN [18]. As DANN was originally designed to classify images, its classifier is used to predict pixel-level categories. Furthermore, AdasegNet is commonly used as the baseline in the literature while Advent exploits an entropy loss and an adversarial loss to minimize the entropy of the probability map. Finally, GANAI, TriADA and CCAGAN are well-established methods for semantic segmentation of VHR remote sensing images.

For comparison purposes, we also consider two methods without domain adaption. In one non-DA method, data from the source domain is utilized for training while the test is performed in the target domain. For instance, the Potsdam IRRG is used for training while Vaihingen IRRG for testing. Clearly, the mismatch between the source and the target domains will incur substantial performance degradation. In the sequel, this method is referred as the non-DA method (Source-Only). In contrast, the other non-DA method under consideration utilizes data and labels from the target domain for both training and testing, which is regular supervised learning within a single domain. In the sequel, this method is referred to as the non-DA (Supervised). The classical ResUNet++ [46] is chosen for these two non-DA methods for its good balance between performance and computational efficiency. Clearly, the non-DA (Supervised) is expected to exhibit the best performance.

#### E. Experimental results and discussion

Extensive experiments were performed on the following three domain adaption tasks listed in Table III.
In Task 1, Potsdam IRRG and Vaihingen IRRG possess the same band but different ground object styles and GSD, which is a typical UDA problem between cities. Furthermore, Task 2 adapts Potsdam RGB to Vaihingen IRRG while Task 3 Vaihingen IRRG to Potsdam RGB. The latter two tasks are more challenging as the two domains in these two tasks show larger discrepancies than the first task. These three experiments were carefully designed to investigate the performance of the proposed MBATA-GAN on cross-band and in-band adaption capabilities.

**Task 1: Potsdam IRRG to Vaihingen IRRG**

Table IV presents the results for Task 1. For the non-DA (Source-Only) method, its corresponding OA, mF1 and mIoU were 63.03%, 50.33% and 37.13%, respectively. First, it is observed that all conventional DA methods outperformed the non-DA (Source-Only) method. For instance, AdasegNet improved the OA, mF1 and mIoU to 76.29%, 71.55% and 57.14%, respectively. This confirmed the necessity of the DA processing. Furthermore, the proposed MBATA-GAN achieved the highest scores among all methods examined. More specifically, the OA, mF1 and mIoU attained by MBATA-GAN were 80.75%, 76.77% and 63.50%, respectively, which amounts to an improvement of 4.46%, 5.22% and 6.36% as compared to AdasegNet. The performance improvement was achieved by aligning the high-level features of the source and the target domains while exploiting the adversarial strategy of improving the discriminative capability of MDM. Fig. 8 depicts the segmentation results achieved by all methods under consideration.

Inspection of Table IV suggests that the non-DA (Source-Only) method had particularly poor classification performance of Car, which was evidenced by the recognition rate of 7.06%. This poor classification performance of car can be explained by the fact that the small size of a car made it difficult for the segmentation model to accurately learn the features of Car. In contrast, the classification rate of Car was improved to 66.38% with the proposed MBATA-GAN, which was 8.79% higher than AdasegNet. Visual inspection of cars in Fig. 8 reveals that the car boundaries were sharpened by MBATA-GAN. Furthermore, MBATA-GAN has improved IoU for Building, Tree, Low Veg and Imp Sur to 90.98%, 79.51%, 63.41% and 82.80%, respectively, as compared to 87.87%, 79.44%, 56.73 and 76.10% obtained by AdasegNet. The proposed MBATA-GAN provided improved performance in all categories in Task 1. Compared with other methods, the proposed MBATA-GAN has fewer cluttered pixels, indicating that it can better recognize complete ground objects. In summary, Table IV showed that the proposed MBATA-GAN achieved substantial performance improvement across all categories and performance metrics in Task 1.

Next, the sensitivity of the two hyper-parameters $\lambda^g$ and $\lambda^f$ is analyzed. $\lambda^g$ is designed to adjust the contribution derived from aligning cross-domain high-level feature maps while $\lambda^f$ adjusts the contribution from the feature discriminator $D^f$. These two parameters play an important role in balancing the transferability and discriminability of the proposed GAN-based model. Fig. 9 (a) shows the performance in terms of mIoU using various $\lambda^g$ values. Clearly, a smaller value of $\lambda^g$ will reduce the contribution from aligning cross-domain features while a larger value may exaggerate the significance of the alignment. Therefore, setting $\lambda^g \geq 4.0$ incurred non-negligible performance degradation. In contrast, the performance is less sensitive to the value of $\lambda^f$ for $\lambda^f \in [0.5, 2.0]$. Therefore, unless otherwise specified, $\lambda^f = 2.0$ was employed in our experiments.

In addition, Fig. 9 (b) shows the model performance when varying $\lambda^f \in [0.0001, 0.5]$ with $\lambda^g = 2.0$. A smaller $\lambda^f$ value made the contribution from the feature discriminator $D^f$ insignificant in learning the model. As $\lambda^f$ grew to 0.005, the MBATA-GAN achieved the best mIoU up to 63.50% as shown in Fig. 9 (b). However, further increase in $\lambda^f$ entailed performance degradation.

Extensive ablation experiments will be conducted in Sec. III-F, including complete removal of the proposed MBATrans and other proposed modules as compared to the baseline.

**Task 2: Potsdam RGB to Vaihingen IRRG**

Table V shows the results for Task 2 in which the Potsdam RGB domain was adapted to the Vaihingen IRRG domain while exploiting the adversarial strategy of improving the discriminative capability of MDM. This poor classification performance of car can be explained by the fact that the small size of a car made it difficult for the segmentation model to accurately learn the features of Car. In contrast, the classification rate of Car was improved to 66.38% with the proposed MBATA-GAN, which was 8.79% higher than AdasegNet. Visual inspection of cars in Fig. 8 reveals that the car boundaries were sharpened by MBATA-GAN. Furthermore, MBATA-GAN has improved IoU for Building, Tree, Low Veg and Imp Sur to 90.98%, 79.51%, 63.41% and 82.80%, respectively, as compared to 87.87%, 79.44%, 56.73 and 76.10% obtained by AdasegNet. The proposed MBATA-GAN provided improved performance in all categories in Task 1. Compared with other methods, the proposed MBATA-GAN has fewer cluttered pixels, indicating that it can better recognize complete ground objects. In summary, Table IV showed that the proposed MBATA-GAN achieved substantial performance improvement across all categories and performance metrics in Task 1.

Next, the sensitivity of the two hyper-parameters $\lambda^g$ and $\lambda^f$ is analyzed. $\lambda^g$ is designed to adjust the contribution derived from aligning cross-domain high-level feature maps while $\lambda^f$ adjusts the contribution from the feature discriminator $D^f$. These two parameters play an important role in balancing the transferability and discriminability of the proposed GAN-based model. Fig. 9 (a) shows the performance in terms of mIoU using various $\lambda^g$ values. Clearly, a smaller value of $\lambda^g$ will reduce the contribution from aligning cross-domain features while a larger value may exaggerate the significance of the alignment. Therefore, setting $\lambda^g \geq 4.0$ incurred non-negligible performance degradation. In contrast, the performance is less sensitive to the value of $\lambda^f$ for $\lambda^f \in [0.5, 2.0]$. Therefore, unless otherwise specified, $\lambda^f = 2.0$ was employed in our experiments.

In addition, Fig. 9 (b) shows the model performance when varying $\lambda^f \in [0.0001, 0.5]$ with $\lambda^g = 2.0$. A smaller $\lambda^f$ value made the contribution from the feature discriminator $D^f$ insignificant in learning the model. As $\lambda^f$ grew to 0.005, the MBATA-GAN achieved the best mIoU up to 63.50% as shown in Fig. 9 (b). However, further increase in $\lambda^f$ entailed performance degradation.

Extensive ablation experiments will be conducted in Sec. III-F, including complete removal of the proposed MBATrans and other proposed modules as compared to the baseline.
TABLE IV

<table>
<thead>
<tr>
<th>Method</th>
<th>Building</th>
<th>Tree</th>
<th>Low Veg</th>
<th>Car</th>
<th>Imp Sur</th>
<th>OA</th>
<th>mF1</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-DA</td>
<td>76.93</td>
<td>70.79</td>
<td>40.19</td>
<td>7.06</td>
<td>56.70</td>
<td>63.03</td>
<td>50.33</td>
<td>37.13</td>
</tr>
<tr>
<td>(Source-Only)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DANN</td>
<td>83.07</td>
<td>75.93</td>
<td>54.15</td>
<td>55.38</td>
<td>69.73</td>
<td>71.82</td>
<td>67.66</td>
<td>52.24</td>
</tr>
<tr>
<td>AdasegNet</td>
<td>87.87</td>
<td>79.44</td>
<td>56.73</td>
<td>57.59</td>
<td>76.10</td>
<td>76.29</td>
<td>71.55</td>
<td>57.14</td>
</tr>
<tr>
<td>Advent</td>
<td>90.85</td>
<td>78.78</td>
<td>58.31</td>
<td>55.07</td>
<td>79.53</td>
<td>77.63</td>
<td>72.51</td>
<td>58.68</td>
</tr>
<tr>
<td>GANAI</td>
<td>87.13</td>
<td>77.20</td>
<td>52.48</td>
<td>58.78</td>
<td>74.38</td>
<td>73.52</td>
<td>69.99</td>
<td>55.29</td>
</tr>
<tr>
<td>TriADA</td>
<td>89.34</td>
<td>79.06</td>
<td>47.50</td>
<td>65.72</td>
<td>81.11</td>
<td>77.56</td>
<td>72.55</td>
<td>58.88</td>
</tr>
<tr>
<td>CCAGAN</td>
<td>89.02</td>
<td>78.10</td>
<td>59.29</td>
<td>65.42</td>
<td>79.10</td>
<td>76.90</td>
<td>74.19</td>
<td>60.09</td>
</tr>
<tr>
<td>Proposed</td>
<td>90.98</td>
<td>79.51</td>
<td>63.41</td>
<td>66.38</td>
<td>82.80</td>
<td>80.75</td>
<td>76.77</td>
<td>63.50</td>
</tr>
</tbody>
</table>

TABLE V

<table>
<thead>
<tr>
<th>Method</th>
<th>Building</th>
<th>Tree</th>
<th>Low Veg</th>
<th>Car</th>
<th>Imp Sur</th>
<th>OA</th>
<th>F1</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-DA</td>
<td>64.48</td>
<td>0.04</td>
<td>5.26</td>
<td>1.57</td>
<td>58.21</td>
<td>35.71</td>
<td>25.91</td>
<td>18.43</td>
</tr>
<tr>
<td>(Source-Only)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DANN</td>
<td>73.20</td>
<td>62.46</td>
<td>41.50</td>
<td>54.94</td>
<td>64.39</td>
<td>61.85</td>
<td>59.30</td>
<td>42.94</td>
</tr>
<tr>
<td>AdasegNet</td>
<td>74.04</td>
<td>75.16</td>
<td>40.85</td>
<td>57.02</td>
<td>62.12</td>
<td>64.75</td>
<td>61.84</td>
<td>45.92</td>
</tr>
<tr>
<td>Advent</td>
<td>83.29</td>
<td>77.85</td>
<td>42.37</td>
<td>49.92</td>
<td>65.73</td>
<td>67.26</td>
<td>63.83</td>
<td>48.84</td>
</tr>
<tr>
<td>GANAI</td>
<td>76.87</td>
<td>56.25</td>
<td>41.50</td>
<td>57.93</td>
<td>66.07</td>
<td>61.81</td>
<td>59.72</td>
<td>43.57</td>
</tr>
<tr>
<td>TriADA</td>
<td>77.86</td>
<td>73.85</td>
<td>36.57</td>
<td>58.70</td>
<td>66.48</td>
<td>66.09</td>
<td>62.69</td>
<td>47.20</td>
</tr>
<tr>
<td>CCAGAN</td>
<td>85.60</td>
<td>75.27</td>
<td>36.31</td>
<td>63.24</td>
<td>66.03</td>
<td>68.38</td>
<td>65.25</td>
<td>50.58</td>
</tr>
<tr>
<td>Proposed</td>
<td>87.67</td>
<td>75.30</td>
<td>39.06</td>
<td>64.95</td>
<td>67.33</td>
<td>69.29</td>
<td>66.86</td>
<td>52.31</td>
</tr>
<tr>
<td>Non-DA</td>
<td>94.84</td>
<td>90.15</td>
<td>78.25</td>
<td>80.16</td>
<td>90.47</td>
<td>89.66</td>
<td>86.77</td>
<td>77.20</td>
</tr>
<tr>
<td>(Supervised)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE VI

<table>
<thead>
<tr>
<th>Method</th>
<th>Building</th>
<th>Tree</th>
<th>Low Veg</th>
<th>Car</th>
<th>Imp Sur</th>
<th>OA</th>
<th>F1</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-DA</td>
<td>42.87</td>
<td>0.10</td>
<td>7.36</td>
<td>3.55</td>
<td>43.27</td>
<td>34.14</td>
<td>19.43</td>
<td>12.11</td>
</tr>
<tr>
<td>(Source-Only)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DANN</td>
<td>63.70</td>
<td>42.50</td>
<td>40.26</td>
<td>58.21</td>
<td>62.10</td>
<td>54.53</td>
<td>53.35</td>
<td>37.00</td>
</tr>
<tr>
<td>AdasegNet</td>
<td>73.23</td>
<td>46.23</td>
<td>33.88</td>
<td>59.48</td>
<td>71.61</td>
<td>58.38</td>
<td>56.89</td>
<td>41.27</td>
</tr>
<tr>
<td>Advent</td>
<td>71.13</td>
<td>46.62</td>
<td>31.70</td>
<td>66.76</td>
<td>70.95</td>
<td>57.30</td>
<td>57.43</td>
<td>41.90</td>
</tr>
<tr>
<td>GANAI</td>
<td>64.97</td>
<td>38.79</td>
<td>45.29</td>
<td>67.45</td>
<td>67.17</td>
<td>57.02</td>
<td>56.73</td>
<td>40.58</td>
</tr>
<tr>
<td>TriADA</td>
<td>73.95</td>
<td>38.06</td>
<td>47.87</td>
<td>70.46</td>
<td>72.55</td>
<td>61.34</td>
<td>60.58</td>
<td>44.99</td>
</tr>
<tr>
<td>CCAGAN</td>
<td>76.45</td>
<td>44.86</td>
<td>47.28</td>
<td>72.82</td>
<td>71.26</td>
<td>61.68</td>
<td>62.53</td>
<td>46.87</td>
</tr>
<tr>
<td>Proposed</td>
<td>81.54</td>
<td>30.75</td>
<td>58.14</td>
<td>69.83</td>
<td>75.37</td>
<td>65.82</td>
<td>63.13</td>
<td>48.42</td>
</tr>
<tr>
<td>Non-DA</td>
<td>92.08</td>
<td>74.91</td>
<td>80.13</td>
<td>90.79</td>
<td>89.35</td>
<td>84.69</td>
<td>85.46</td>
<td>75.19</td>
</tr>
<tr>
<td>(Supervised)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RGB serves as the source domain while the Vaihingen IRRG is the target domain. In contrast to Task 1, the domain shift in Task 2 is caused by variations in both the geographic locations and the imaging bands. Inspection of Table V indicates that the proposed MBATA-GAN achieved OA of 69.29%, mF1 of 66.86%, and mIoU of 52.31%, which stands for an increase of 4.54%, 5.02%, and 6.39% as compared to the corresponding performance of AdasegNet, respectively.

Furthermore, the non-DA (Source-Only) method showed very poor performance in classifying Tree, Low Veg and Car. The corresponding F1 scores were 0.04%, 5.26% and 1.57%. In sharp contrast, the F1 scores achieved by the proposed MBATA-GAN for Building, Car and Imp Sur were 87.67%, 64.95% and 67.33%, respectively, which amounts to an increase of 13.63%, 7.93% and 5.21% as compared to AdasegNet. Meanwhile, in Task 2, the MBATA-GAN improved the accuracy for Tree by 0.14% but with degraded performance for Low Veg by 1.79%. This observation can be explained by the fact that Building, Car and Imp Sur share great similarities in shape and texture between the source domain and the target domain. However, as Tree and Low Veg can change greatly depending on their geological location, it was particularly challenging for the UDA methods to differentiate Tree and Low Veg. From Table V, Advent showed stronger performance in Tree and Low Veg while the proposed MBATA-GAN exhibited a definite advantage in overall performance across all categories.

Fig. 10 visualizes the segmentation results attained with dif-
Fig. 10. Task 2: qualitative results of the adaptation Potsdam RGB to Vaihingen IRRG with the size of 768 × 768. (a) RGB images, (b) Ground Truth, (c) Non-DA (Source-Only), (d) DANN, (e) AdasegNet, (f) Advent, (g) GANAI, (h) TriADA, (i) CCAGAN, (j) The proposed MBATA-GAN and (k) Non-DA (Supervised). Color representation is as follows. Blue: Building, green: Tree, cyan: Low Veg, yellow: Car, white: Imp Suf. Red: clutter/background.

TABLE VII
SENSITIVITY ANALYSIS ON THE STRUCTURE OF FVIT.
(BOLD VALUES ARE THE BEST)

<table>
<thead>
<tr>
<th>Model</th>
<th>MBATrans</th>
<th>$L_{global}$</th>
<th>Enhanced $D^f/D^h$</th>
<th>$D^f$</th>
<th>OA</th>
<th>mF1</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>76.29</td>
<td>71.55</td>
<td>57.14</td>
</tr>
<tr>
<td>ablation1</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>77.73</td>
<td>73.01</td>
<td>59.08</td>
</tr>
<tr>
<td>ablation2</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>76.10</td>
<td>69.84</td>
<td>55.47</td>
</tr>
<tr>
<td>ablation3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>77.21</td>
<td>74.02</td>
<td>59.91</td>
</tr>
<tr>
<td>ablation4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>78.95</td>
<td>74.86</td>
<td>61.16</td>
</tr>
<tr>
<td>ablation5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>77.35</td>
<td>72.06</td>
<td>58.12</td>
</tr>
<tr>
<td>MBATA-GAN*</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.01</td>
<td>75.31</td>
<td>61.57</td>
</tr>
<tr>
<td>MBATA-GAN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.75</td>
<td>76.77</td>
<td>63.50</td>
</tr>
</tbody>
</table>

MBATA-GAN*: only computing cross-attention and adding $\{F_s, A_{s2t}\}$ and $\{F_t, A_{t2s}\}$ directly.

Different methods in Task 2. Clearly, the non-DA (Source-Only) method suffered from the domain shift while the existing DA methods were not able to handle the large domain discrepancy in Task 2. In contrast to the cluttered segmentation results derived from these existing methods, the proposed MBATA-GAN demonstrated clear segmentation of better object integrity and smoother object boundary. For instance, the images in the fourth row show some vehicles parking on the rooftop. For such complex pixels, the proposed MBATA-GAN provided much more reliable classification performance.

Task 3: Vaihingen IRRG to Potsdam RGB
Finally, we performed Task 3 that exhibits the largest domain shift among the three tasks. In addition to the discrepancies in the geographic locations and the imaging modes, the Vaihingen dataset is smaller in size than the Potsdam dataset. The results of Task 3 are presented in Table VI. The proposed MBATA-GAN achieved OA of 65.82%, mF1 of 63.13%, and mIoU of 48.42%, which stands for an increase of 7.44%, 6.24%, and 7.15% as compared to AdasegNet. The results on Tree and Low Veg exhibit greater volatility in Task 3 due to the larger domain shift. Fig. 11 visualizes the segmentation results achieved by different methods in Task 3. The non-DA (Source-Only) showed similar performance in both Task 2 and Task 3 as the larger domain shift made it completely incompetent of classification in the target domain. In sharp contrast, the proposed MBATA-GAN provided classification with impressive accuracy.

In summary, the proposed MBATA-GAN demonstrated improved performance in overcoming domain shifts caused by variations in both the geographic locations and the imaging modes as compared to AdasegNet and other existing UDA methods.
Fig. 11. Task 3: qualitative results of the adaptation Vaihingen IRRG to Potsdam RGB with the size of 1024 × 1024. (a) RGB images, (b) Ground Truth, (c) Source-only, (d) DANN, (e) AdasegNet, (f) Advent, (g) GANAI, (h) TriADA, (i)CCAGAN, (j) The proposed MBATA-GAN and (k) Supervised method. Each color represents, blue: Building, green: Tree, cyan: Low Veg, yellow: Car, white: Imp Suf. Red denotes the clutter/background.

F. Ablation Studies

In this section, ablation experiments were conducted to verify the necessity and effectiveness of each module proposed as compared to the baseline. Table VII compares the performance of the baseline and five ablation experiments investigating various module dependencies. Unless otherwise specified, we set $\lambda^g = 2.0$. It is worth mentioning that the fourth column, “Enhanced $D^\ell/D^h$” indicates whether to add self-attention based transformer to $D^\ell$ or $D^h$.

Inspection of Table VII suggests that all ablation experiments resulted in better performance as compared to the baseline, except “ablation2” that employed MBATrans without the proposed discriminators. More specifically, the ablation experiment labeled as “ablation1” showed that an improvement of 1.44%, 1.46% and 1.94% in terms of OA, $mF1$ and $mIoU$ respectively, has been achieved as compared to baseline. These improvements confirmed that MBATrans can effectively analyze complex remote sensing image content by exploring deep features. In addition, the performance improvement demonstrated by “ablation1” and “ablation3” showed that the proposed generator or discriminators are more effective in feature extraction on remote sensing VHR images as compared to the pure CNN-based baseline model. In contrast, the experiment labeled as “ablation2” showed degraded performance, which hints that the GAN framework is sensitive to the hyper-parameter $\lambda^g$ as indicated in Fig. 9(a). Furthermore, the ablation experiments labeled as “ablation4” and “ablation5” demonstrated noticeable performance gains. Finally, the ablation experiment labeled as “MBATA-GAN*” computed cross-attention layers while naively adding $\{F_s, A_{s2t}\}$ and $\{F_t, A_{t2s}\}$ directly. This is in sharp contrast to the proposed MBATrans in which $F_{s2t}$ and $F_{t2s}$ keep the intermediate inter-domain feature representations while four branches of MBA modules are employed to learn intra- and inter-domain features between different domains. Table VII clearly confirmed that the proposed MBATrans is more effective in exploiting these intra- and inter-domain features.

G. Model Complexity Analysis

Table VIII shows the complexity evaluation of all methods considered in our experiments. Compared with the baseline AdasegNet, the proposed MBATA-GAN achieved improved framework. Both “ablation4” and “ablation5” demonstrated noticeable performance gains. Further, the ablation experiment labeled as “MBATA-GAN**” computed cross-attention layers while naively adding $\{F_s, A_{s2t}\}$ and $\{F_t, A_{t2s}\}$ directly. This is in sharp contrast to the proposed MBATrans in which $F_{s2t}$ and $F_{t2s}$ keep the intermediate inter-domain feature representations while four branches of MBA modules are employed to learn intra- and inter-domain features between different domains. Table VII clearly confirmed that the proposed MBATrans is more effective in exploiting these intra- and inter-domain features.

TABLE VIII

<table>
<thead>
<tr>
<th>Model</th>
<th>Complexity (G)</th>
<th>Parameters (M)</th>
<th>Memory (MB)</th>
<th>Speed (FPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResUNet++</td>
<td>241.82</td>
<td>353.46</td>
<td>4365</td>
<td>25.10</td>
</tr>
<tr>
<td>AdasegNet</td>
<td>47.59</td>
<td>42.83</td>
<td>2491</td>
<td>43.31</td>
</tr>
<tr>
<td>Advent</td>
<td>47.95</td>
<td>43.16</td>
<td>2491</td>
<td>30.65</td>
</tr>
<tr>
<td>GANAI</td>
<td>10.27</td>
<td>44.55</td>
<td>1075</td>
<td>6.30</td>
</tr>
<tr>
<td>TriADA</td>
<td>22.15</td>
<td>59.23</td>
<td>1075</td>
<td>4.54</td>
</tr>
<tr>
<td>CCAGAN</td>
<td>145.14</td>
<td>104.02</td>
<td>4503</td>
<td>13.04</td>
</tr>
<tr>
<td>MBATA-GAN</td>
<td>174.20</td>
<td>143.90</td>
<td>5295</td>
<td>8.11</td>
</tr>
</tbody>
</table>
results on the well-known ISPRS Vaihingen and Potsdam datasets have confirmed the superior segmentation performance achieved by the proposed MBATA-GAN as compared to AdasegNet and other existing UDA methods. This work has successfully demonstrated the feasibility of solving the domain shift problem by exploiting cross-attention and self-attention based transformers.

There are several extensions of this study that can be further explored. In particular, it has been observed that the classification rates of Tree and Low Veg achieved by MBATA-GAN in Task 2 were 75.30% and 39.06% whereas the F1 score of Tree achieved by MBATA-GAN in Task 3 was 30.75%. These less satisfactory results on Tree and Low Veg were caused by the large discrepancies of these two classes in the source and the target domains. In other words, Tree and Low Veg can exhibit large changes in their characteristics such as shape and texture, depending on their geographical location. Thus, it is worth further investigation on sophisticated feature extraction techniques or new strategies to exploit auxiliary information such as depth for such hard UDA problems. Finally, it has been shown in Sec. III-G that the performance gain by the proposed MBATA-GAN was achieved at the cost of increased computational complexity. Therefore, it is of great practical importance to investigate low-complexity models for MBATA-GAN.

H. Gradient-based Localization Visualization

In this section, we will investigate the characteristics of the proposed MBATrans and MBATA-GAN by visually inspecting the gradients inside the backbone using Grad-CAM [47] in a manner similar to [35]. More specifically, we compare the heat maps generated by AdasegNet and the proposed MBATA-GAN. The outputs after feature extraction in both methods are shown in Fig. 12 in which a bold black dot is used to indicate the pixel under classification consideration. The first and the fourth columns show the images and the corresponding ground truth, respectively. Furthermore, the second and the third columns show the gradients used by AdasegNet and MBATA-GAN for classification, respectively. Comparison between the second and the third columns revealed that there were more high-score areas with respect to the dot of interest for MBATA-GAN, which resulted in higher classification accuracy. Fig. 12 confirmed that the CNN-based feature extractor endowed with MBATrans and MDM can learn transferable features more effectively, which enables the classifier to make more confident classification decisions based on the information gathered from more pixels of the same category.

IV. CONCLUSION

In this paper, a novel GAN-based architecture called MBATA-GAN has been proposed to perform unsupervised domain adaptation for semantic segmentation of VHR remote sensing imagery by exploiting cross-domain high-level feature alignment in the rich semantic subspace. In particular, an MBATrans module that leverages both cross-attention and self-attention is developed to learn semantic consistency via global alignment. Furthermore, the generator in MBATA-GAN learns the transferable feature whereas the proposed MDM adversarially learns the discriminative features. Extensive simulation

REFERENCES


Xianping Ma received his Bachelor in Geographical Information Science from Wuhan University, China in 2019. He is currently pursuing the PhD degree at the Chinese University of Hong Kong, Shenzhen, China. His research interests include remote sensing and machine learning.

Xiaokang Zhang (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University in 2018. From 2019-2022, he was a postdoctoral research associate at the Hong Kong Polytechnic University and the Chinese University of Hong Kong (Shenzhen). He is currently an Associate Professor in the School of Information Science and Engineering, Wuhan University of Science and Technology. His research interests include remote sensing image analysis, computer vision, and deep learning. He has authored or co-authored more than 20 scientific publications in international journals and conferences. He frequently serves as a reviewer for several renowned international journals, such as IEEE Transactions on Neural Networks and Learning Systems, Information Fusion, and IEEE Transactions on Geoscience and Remote Sensing.

Zhiguo Wang (Member, IEEE) is an associate professor with the School of Information Science and Engineering, Wuhan University of Science and Technology, China. He received the Ph.D. degree from Sichuan University, Chengdu, China, in 2018. From 2018 to 2020, he was Postdoctoral Researcher with The Chinese University of Hong Kong, Shenzhen, China. Since 2020, he has been working as a specially appointed Associate Research Fellow with the College of Mathematics, Sichuan University. His current research interests include information fusion, target tracking, machine learning, nonconvex optimization, and analysis and control of uncertain systems.

Man-On Pun (Senior Member, IEEE) received his BEng. degree in Electronic Engineering from the Chinese University of Hong Kong (CUHK) in 1996, the MEng. degree in Computer Science from University of Tsukuba, Japan in 1999 and the Ph.D. degree in Electrical Engineering from University of Southern California (USC) in Los Angeles, U.S.A in 2006, respectively. He was a postdoctoral research associate at Princeton University from 2006 to 2008. Currently, he is an associate professor with the School of Science and Engineering, the Chinese University of Hong Kong, Shenzhen (CUHKSZ). Prior to joining CUHKSZ in 2015, he held research positions at Huawei (USA) in New Jersey, Mitsubishi Electric Research Labs (MERL) in Boston and Sony in Tokyo, Japan. Prof. Pun’s research interests include AI Internet of Things (AoT) and applications of machine learning in communications and satellite remote sensing. Prof. Pun has received best paper awards from IEEE VTC’06 Fall, IEEE ICC’08 and IEEE Infocom’09. He served as associate editor for the IEEE Transactions on Wireless Communications from 2010 to 2014. He is the founding chair of the IEEE Joint SPS-ComSoc Chapter, Shenzhen.