remote sensing

MDPI

Article

Depth-Aware Adversarial Domain Adaptation for Cross-Domain Remote Sensing Segmentation

Lulu Niu [1,2,3], Xiaoxuan Liu [1,2], Enze Zhu [1,2,3], Yidan Zhang [1,2,3], Hanru Shi [1,2,3], Xiaohe Li [1,2], Hong Wang [1,2,*], Jie Jia [1,2] and Lei Wang [1,2]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; niululu17@mails.ucas.ac.cn (L.N.); liuxiaoxuan@aircas.ac.cn (X.L.); zhuenze23@mails.ucas.ac.cn (E.Z.); zhangyidan19@mails.ucas.ac.cn (Y.Z.); shihanru23@mails.ucas.ac.cn (H.S.); lixiaohe@aircas.ac.cn (X.L.); jiajie@aircas.ac.cn (J.J.); wanglei002931@aircas.ac.cn (L.W.)
[2] Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing 100190, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wanghong@aircas.ac.cn

Highlights

What are the main findings?

- Integrating depth information as an auxiliary modality significantly improves the exploitation of geometric cues for cross-domain remote sensing segmentation.
- Mining informative signals from unlabeled target-domain data can promote adaptation performance and mitigate domain gaps.

What are the implications of the main findings?

- The study provides a key solution to enhance segmentation generalizability in unsupervised cross-domain multimodal learning.
- The study offers a valuable reference for applying depth-assisted segmentation to real-world remote sensing scenarios.

Abstract

As a key task in remote sensing analysis, semantic segmentation of remote sensing images (RSI) underpins many practical applications. Despite its importance, obtaining dense pixel-wise annotations remains labor-intensive and time-consuming. Unsupervised domain adaptation (UDA) offers a promising solution by utilizing knowledge from labeled source domains for unlabeled target domains, yet its effectiveness is often compromised by significant distribution shifts arising from variations in imaging conditions. To address this challenge, we propose a depth-aware adaptation network (DAAN), a novel two-branch network that explicitly leverages complementary depth information from a digital surface model (DSM) to enhance cross-domain remote sensing segmentation. Unlike conventional UDA methods that primarily focus on semantic features, DAAN incorporates depth data to build a more generalized feature space. This network introduces three key components: an adaptive feature aggregator (AFA) for progressive semantic-depth feature fusion, a gated prediction selection unit (GPSU) that selectively integrates predictions to mitigate the impact of noisy depth measurements, and misalignment-focused residual refinement (MFRR) module that emphasizes poorly aligned target regions during training. Experiments on the ISPRS and GAMUS datasets demonstrate the effectiveness of the proposed method. In particular, DAAN achieves an mIoU of 50.53% and an F1 score of 65.75% for cross-domain segmentation on ISPRS to GAMUS, outperforming models without depth information by 9.17% and 8.99%, respectively. These results demonstrate the advantage of integrating auxiliary geometric information to improve model generalization

Check for updates

Academic Editor: Shuying Li

Received: 4 December 2025
Revised: 31 March 2026
Accepted: 1 April 2026
Published: 7 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

on unlabeled remote sensing datasets, contributing to higher mapping accuracy, more reliable automated analysis, and enhanced decision-making support.

Keywords: remote sensing; unsupervised domain adaptation; semantic segmentation; adversarial learning

1. Introduction

Semantic segmentation attends to the assignment of a meaningful label to each pixel based on its category to partition an image into regions. It plays a crucial role in applications such as building extraction and scene understanding [1]. However, the efficacy of a semantic segmentation model hinges heavily on the supply of extensive annotated data for its training. Considering the intricacy of remote sensing images, detailed pixel-level annotation can be laborious and time-consuming [2]. To mitigate the demand for intensive manual annotation, some researchers have explored the transfer of models trained on annotated datasets to unlabeled target datasets [3–5]. Nonetheless, the expected gains are often limited by the distributional discrepancies that naturally arise across datasets.

In machine learning, a domain refers to a collection of input data characterized by a shared feature space and a marginal probability distribution. Cross-domain scenarios arise when the source domain used for model training and the target domain where the model is deployed exhibit distinct data distributions. Standard segmentation models trained on a labeled source domain typically suffer severe performance degradation when directly applied to an unlabeled target domain due to such domain shifts [6]. To mitigate this issue, a wide range of domain adaptation methods have been developed to enhance performance in unlabeled target domains [3,6]. Among these, approaches for semantic segmentation commonly focus on narrowing the gap between source and target domains, with representative techniques including adversarial learning methods motivating the development of cross-domain segmentation methods to mitigate this distribution gap, with representative techniques including adversarial learning [7–11] and self-training [12–16]. Adversarial training (AT) minimizes domain discrepancies by confounding domain discriminators. Self-training, on the other hand, iteratively generates pseudo-labels for unlabeled target data and incorporates these selected pseudo-labels into the training process.

Recent efforts have demonstrated the promising achievement of adversarial learning in UDA semantic segmentation tasks [17–19]. Typically, these methods involve a segmentation network and a domain discriminator. During training, these networks engage in a continuous adversarial process in which features, predictions, or both produced by the segmentation network are fed into the discriminator. The segmentation network attempts to fool the discriminator by making outputs from both domains appear indistinguishable, while the discriminator seeks to determine the domain of its inputs. Through this adversarial interplay, the segmentation network is encouraged to learn domain-invariant representations.

For remote sensing images, domain gaps originate from inherent differences across domains, including variations in sensors, spatial resolution, seasonal conditions, and land-cover characteristics. Such variationscan significantly alter the visual appearance of the same land-cover categories. Meanwhile, inconsistencies in scene composition, such as land-use patterns, object density, and landscape structure, further exacerbate cross-domain mismatches. These discrepancies often lead to severe semantic ambiguities among similar classes, which substantially degrade segmentation performance on unseen target datasets. Conventional spectrally based UDA methods struggle to mitigate these issues, as they only model 2D spectral and textural features while lacking awareness of 3D topo-

graphic structure. Spectral information alone cannot disentangle the typical ambiguities in remote sensing. Moreover, RSIsare highly sensitive to variations in the spectrum, atmospheric conditions, and sensors, making it difficult to achieve stable and consistent cross-domain representation.

One way to overcome these limitations is to incorporate supplementary information. Earlier studies [20,21] have shown that incorporating depth information has a positive impact on performance on semantic segmentation tasks. However, there is no assurance that a positive outcome will occur under unsupervised conditions, as there has been inadequate investigation into the effectiveness of introducing another modality of data for domain adaptation.

This work aims to study how to make use of depth information (i.e., a digital surface model, DSM) to improve the UDA semantic segmentation of remote sensing images. DSM data can be easily obtained with photogrammetric methods when creating aerial imagery datasets and are commonly available in both source and target domains. Such data can provide valuable auxiliary information for the semantic segmentation of remote sensing images, and their incorporation typically does not necessitate extensive preprocessing. This advantage arises from the consistent relative relationships among land-cover classes reflected in DSM data; for example, the elevation values associated with roads are generally lower than those corresponding to buildings, irrespective of varying environmental conditions. Exploiting such intrinsic semantic–geometric correlations facilitates the enhancement of segmentation performance in the target domain, particularly in the presence of domain shifts.

Nevertheless, real-world DSM data inevitably contain inherent imperfections, such as incomplete coverage and spatial misalignment with remote sensing images. These disturbances can introduce inconsistencies between DSM and spectral information, thereby impairing segmentation performance, particularly for fine-grained objects such as buildings and roads, where elevation plays a critical role. Compounding this issue, most existing depth-assisted methods [22,23] treat depth either as a regression task label or rely on explicit post-processing steps after segmentation inference. Furthermore, in the absence of supervision for the target domain, the contribution of target -domain images to the updating of network parameters is limited, leading to insufficient utilization of the diverse information inherent in the target domain.

In this work, we propose a method to bridge domain disparities in remote sensing segmentation via integration of extra information. First, a two-branch architecture is designed to separately process RSI and DSM inputs, enabling modality-specific feature extraction and mitigating modality-induced discrepancies. Its end-to-end paradigm enables more efficient and seamless exploitation of depth's geometric cues for cross-domain segmentation. An adaptive feature aggregator (AFA) based on cross-modality attention is developed to adaptively fuse features from different modalities, thereby reducing the potential adverse effects of DSM data. Additionally, a gated prediction selection unit (GPSU) adaptively weights and merges predictions from both branches, enhancing the accuracy of segmentation results. To address potential domain misalignment in the target domain, we use a misalignment-focused residual refinement (MFRR) module, which identifies poorly aligned regions in images and re-weights the target features to make the best of the target-domain information. Beyond reporting improvements over commonly used benchmarks, we additionally evaluate our method on the new GAMUS dataset [24], providing further evidence of its generalization capability across different data distributions. This comprehensive design enhances the model's ability to generalize across domains. The proposed method demonstrates potential for application in scenarios that demand rapid response under data-insufficient conditions, offering an effective solution.

The contributions of this paper are threefold:

- We propose an end-to-end, cross-domain semantic segmentation framework that integrates depth as an auxiliary input modality, enabling the network to exploit geometric cues to improve robustness under domain shifts.
- The network learns the correlation between modalities through a feature aggregator, which it shares across domains through adversarial learning. Moreover, we employ a module to identify poorly aligned features, exploiting informative signals from the unlabeled target domain.
- Extensive experiments on the ISPRS and GAMUSdatasets demonstrate the effectiveness and strong generalization ability of the proposed network.

2. Related Works

2.1. Overview of Semantic Segmentation

Semantic segmentation is an actively researched topic in computer vision. Since the Fully Convolutional Neural Network [25] model was introduced in 2014, CNN models have been applied to semantic segmentation. The DeepLab series [26–29] gradually improved the performance of such models with the help of atrous convolution and other techniques. DeepLabV3+ [29] utilized dilated convolution and pyramidal pooling modules to effectively gather information across multiple scales, exemplifying the beneficial impact of leveraging multi-scale information to enhance the efficacy of semantic segmentation.

In addition, there have been comprehensive research and applications of the attention mechanism [30] and its successors [31,32]. TransUnet [33] adopted a hybrid variant of ViT [34], integrating the feature map extracted by a CNN into a Transformer for further representation learning. UNetMamba [35] improved the efficiency of high-resolution semantic segmentation by employing the newly developed linear Mamba mechanism [32].

Inspired by the recent exploration of foundational large language models, the computer vision community is developing large-scale vision foundation models [36–38]. CLIP [37] is a multimodality model based on contrastive learning, aiming to enable the model to discern and learn the associative relationships between pairs of text and images. Recently, the SAM prompt-based segmentation model [38] displayed zero-shot generalization capabilities on unseen images without additional training. Its remarkable success has triggered a shift in the paradigm of how semantic segmentation models are trained. AdaptVFMs-RSCD [39] explores the use of vision foundation models for remote sensing change detection by integrating SAM with CLIP to enhance semantic recognition of land-cover categories under remote sensing-specific challenges. The method further designs a semantically aware change detection module to jointly exploit semantic and change cues, achieving strong performance. However, achieving optimal performance on a particular dataset often requires a fine-tuning process. In domain adaptation tasks, constrained data annotation availability increases the impracticalityof training small models. Small models have a crucial role in tackling these research problems and are still worth studying.

2.2. Unsupervised Domain Adaptation in Semantic Segmentation

In remote sensing, the high cost of pixel-wise labeled annotations for large-scale datasets can be addressed by unsupervised domain adaptation methods. These methods enable models trained on a labeled source domain to effectively leverage unlabeled data from a target domain, reducing the dependency on extensive manual annotations and enhancing the utilization of available data resources.

Generative approaches aim to translate source images into the style of the target domain, thereby allowing the model to learn from source labels while enhancing adaptation to the target distribution. ColorMapGAN [40] generated synthetic images that preserve

source semantics while adopting the target domain's spectral style, enabling effective fine-tuning of a pre-trained classifier. On the other hand, adversarial training methods aim to minimize domain shift at the feature level or output level. AdaptSegNet [9] employs two output-space discriminators to determine whether predictions originate from the source domain. Chung [41] proposed a self-attention module to learn domain-invariant inter-pixel correlations. Self-training methods use pseudo-labels assigned to target samples to guide model training for adaptation [13]. Hybrid training methods [42] are not restricted to employing just one of the three aforementioned methods. In real-world applications, striking a balance between accuracy and speed is essential in the implementation of these multi-stage methods.

In particular, adversarially based methods are notable for their ability to align domains at the pixel, feature, and output levels, without the need for explicit object modeling. Remote sensing researchers have developed a range of adversarially based UDA strategies to minimize domain gaps. MDANet [43] proposed a memory-contrastive unsupervised domain adaptation method designed to help the model retain domain-invariant features. CLA-UDA [5] formulated category-level alignment between source and target domains using integer programming. MSDAN [44] proposed a high-performance network utilizing the multi-scale convolutional structure and an attention mechanism, effectively capturing rich multi-scale features of remote sensing images. These approaches aim to mitigate misclassification and reduce confusion among ground-object categories and have contributed to the advancement of unsupervised domain adaptation in semantic segmentation tasks.

2.3. Unsupervised Domain Adaptation with Depth Information

Different modalities contain diverse features. MSFF-Net fuses multimodal remote sensing features in both spatial and frequency domains via wavelet transform and multi-scale contextual attention, effectively preserving edge details and improving segmentation boundary clarity [45]. Integrating heterogeneous data compensates for the limitations of single-modality information and leverages complementary properties to create a more comprehensive feature space, including geometric and semantic aspects. Evidence suggests that incorporating depth information can enhance performance on semantic segmentation tasks [46]; thus, researchers have started to leverage depth as privileged information in UDA for semantic segmentation. Multi-task frameworks leverage depth as a supervisory signal to regularize shared feature representations. DADA [21] proposed a new domain-adaptive framework for depth perception, which changed the segmentation backbone and embedded depth information into a dedicated depth architecture through assisted depth regression tasks to effectively utilize depth as privileged information in an unsupervised domain-adaptive environment. CTRL [22] proposed a network architecture with a multi-task-aware feature alignment mechanism for UDA, introducing a differentiable, non-parametric cross-task relation layer to model interactions between semantic and depth predictions. In remote sensing, Chen et al. proposed MIDANet [47], a mutual information-based domain adaptation network that combines semantic segmentation and nDSM estimation to guide the learning of domain-invariant features through multi-task and entropy-based learning.

Apart from multi-task learning, a model can better adapt to domain shifts and improve segmentation accuracy by leveraging complementary information from different modalities—namely, depth data. Zhou proposed the improvement of segmentation performance by introducing DSM data [23], which can be used to provide the height information of objects in an image and extract domain-invariant features from different images.

Utilizing depth as a supervised task in a multi-task framework and employing depth as an auxiliary modality in a multimodal fusion pipeline represent two distinct technical paradigms. Multi-task approaches can introduce complementary geometric constraints but

may suffer from task competition between segmentation and depth regression, as the optimization objectives of the two tasks may conflict and mutually restrict each other. In contrast, multimodal methods can integrate information from different data sources to enhance the robustness and comprehensiveness of feature representation, yet they rely heavily on the quality and consistency of multimodal data. Consequently, Zhou's work [23] adopted additional post-processing operations to refine the probability map. Moreover, despite their advantages, these methods may struggle to capture domain-specific nuances without detailed supervision from the target domain as models, leading to suboptimal feature alignment and reduced segmentation accuracy. In contrast with these methods, our method relaxes such task competition and further exploits information from the target domain.

3. Methods

3.1. Preliminaries

In this paper, we focus on the task of unsupervised domain adaptation for semantic segmentation in remote sensing scenarios, where the source and target domains differ in data distribution but share an identical label space.

Formally, let S denote the fully labeled source domain and T denote the unlabeled target domain. The source domain consists of a set of remote sensing images ($X_S \subset \mathbb{R}^{H\times W\times 3}$), DSM data ($D_S \subset \mathbb{R}^{H\times W}$), and the corresponding pixel-wise annotations ($Y_S \subset \{1,\ldots,C\}^{H\times W}$). The target domain includes only the RSI ($X_T \subset \mathbb{R}^{H\times W\times 3}$) and DSM data ($D_T \subset \mathbb{R}^{H\times W}$), without available annotations. Here, H and W denote the height and width of the images, and C denotes the number of semantic classes.

The goal is to train a segmentation model using S that can generalize effectively to T, despite the inherent domain gap. To this end, our method fuses complementary information from RSI and DSM modalities and incorporates strategies to mitigate both cross-modal and cross-domain discrepancies. As part of the adversarial learning framework, domain discriminators are trained to distinguish between features from the two domains. For this purpose, we assigned a domain label of 0 to source samples and 1 to target samples. The following sections detail the proposed architecture and adaptation modules.

3.2. Model Architecture

3.2.1. Overall Architecture

As illustrated in Figure 1, the proposed framework consists of two main components—a segmentation network and discriminators—collectively fulfilling three core functions, i.e., feature extraction, semantic classification, and adversarial domain alignment. To facilitate understanding of the data flow, Figure 1 uses color-coded arrows to distinguish domain-specific data pathways: orange for the source domain and blue for the target domain.

Furthermore, inspired by CTRL [22], which models the relationships between semantic and depth predictions, we introduce a gated prediction selection unit to weight the segmentation outputs from both branches, yielding a consolidated segmentation result.

The segmentation network follows an encoder–decoder architecture and comprises two parallel branches dedicated to extracting complementary features from RSI and DSM data, respectively. Specifically, the semantic branch adopts a ResNet101-based encoder (E_I), while the depth branch employs a ResNet34-based encoder (E_D). Both encoders are structured into four stages. At each stage, features extracted from the two modalities are progressively integrated through the AFA, which is strategically placed between stages as shown in Figure 1. The AFA is designed to facilitate interaction between heterogeneous modalities by learnably merging semantic and geometric information, thereby enhancing the network's ability to exploit their complementary characteristics. The fused features are subsequently passed to the next stage of the semantic branch. After the encoder, the

depth and fused features are fed into the classifiers, which both consist of an atrous spatial pyramid pooling (ASPP) module to produce dense predictions with multi-scale context.

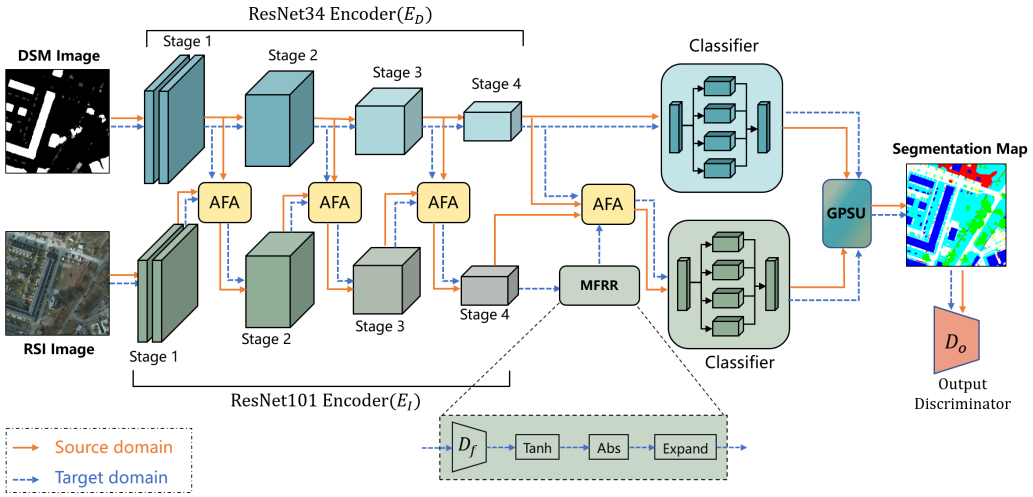


Figure 1. Structure of the proposed DAAN network. The blue dashed arrows and orange solid arrows denote the data flow from the target domain and source domain, respectively.

For unsupervised domain adaptation, we employ two discriminators: a feature-level one (D_f) embedded in the MFRR and an output-level one (D_o) applied after the final prediction.

These components described above jointly facilitate effective domain adaptation, allowing the model to better handle significant domain shifts between the source and target domains.

3.2.2. Adaptive Feature Aggregator

The structure of the AFA is illustrated in Figure 2. The AFA contains three 1 × 1 convolutions and a learnable parameter (α) to aggregate cross-modal features. Specifically, one 1 × 1 convolution projects the RSI branch feature (F_{rsi}) to the query (Q). For the DSM modality feature (F_{dsm}), the two other 1 × 1 convolutions separately generate the key (K) and the value (V):

$$Q = F_{rsi}W_Q, \quad K = F_{dsm}W_K, \quad V = F_{dsm}W_V \tag{1}$$

The attention is then calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2}$$

The RSI features act as queries to retrieve relevant information from the DSM modality, allowing the model to focus on spatial structures semantically consistent with RSI visual patterns. By using the rich semantics of RSI to guide the attention mechanism, this design exploits the complementary structural and depth information in DSM to strengthen the semantic representation of RSI and reduce ambiguity in regions where RSI features alone are inadequate. To achieve adaptive regulation of the information fusion depth, this study introduces a learnable parameter (α) to flexibly control the weight of the attention output in the final fused features. To encourage the model to gradually learn to fuse attention features rather than introducing them abruptly, α is initialized to 0. The attention output weighted by α is added to the original RSI feature (F_{rsi}), generating the final feature (F'_{rsi}) with integrated depth information:

$$F'_{rsi} = \alpha\text{Attention} + F_{rsi} \tag{3}$$

The fused feature (F'_{rsi}) is then passed to the next stage of the RSI branch to perform subsequent feature extraction and network computation. For the DSM branch, the F_{dsm} feature is directly sent to the residual block in the next stage.

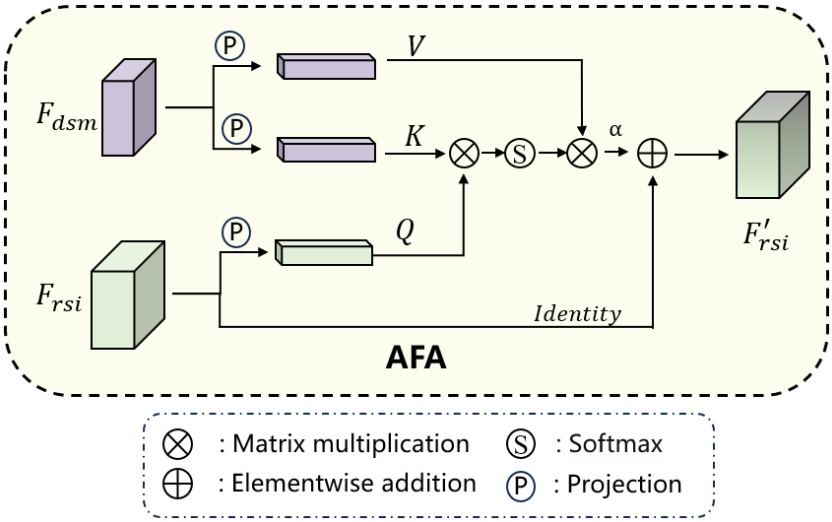


Figure 2. Illustration of the proposed AFA. Features from different modalities are integrated through cross-attention.

3.2.3. Gated Prediction Selection Unit

The design of GPSU draws upon the concept of a gate structure, as illustrated in Figure 3. Specifically, the GPSU comprises a 1 × 1 convolutional layer and a Sigmoid activation function. Let P_{dsm} and P_{rsi} denote the semantic predictions generated by the DSM and RSI branches, respectively, and P_g denote the selected prediction produced by the GPSU. These predictions are first concatenated along the channel dimension to form a joint representation, i.e.,

$$z = [P_{dsm}; P_{rsi}] \tag{4}$$

where $[\cdot\,;\cdot]$ denotes channel-wise concatenation. A 1 × 1 convolution is then applied to generate gating signals, which are transformed into probabilistic weights through the Sigmoid function, i.e.,

$$g = \sigma(\mathrm{Conv}_{1\times1}(z)) \tag{5}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation. The generated weights are applied to the DSM prediction via element-wise multiplication to selectively filter its response:

$$\tilde{P}_{dsm} = g \odot P_{dsm} \tag{6}$$

where $\odot$ denotes element-wise multiplication. Finally, the filtered DSM prediction is combined with the RSI prediction through element-wise addition to produce the fused prediction:

$$P_g = \tilde{P}_{dsm} + P_{rsi} \tag{7}$$

Through this gating mechanism, the GPSU serves as a filtering gate to weight the contributions of both branches. By integrating their complementary information, it enhances the final segmentation result and reduces modality inconsistencies.

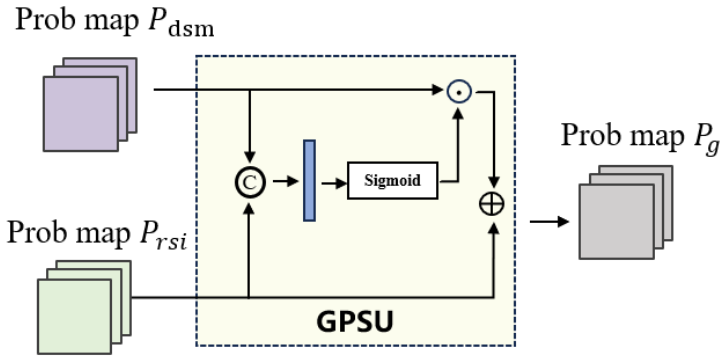


Figure 3. Illustration of the proposed GPSU. Predictions from different branches are fused through the gating mechanism.

3.2.4. Misalignment-Focused Residual Refinement

Within the adversarial learning framework, the probability score predicted by the domain discriminator reflects the domain attribution of the input. Scores near 0 indicate source-domain similarity, whereas scores close to 1 suggest that the input is readily identified as target-domain data. During segmentation network training, high discriminator scores on target-domain input indicate poor feature alignment, as they indicate that the domain discriminator can easily distinguish between source and target features, which breaks the domain-invariance assumption essential for robust cross-domain generalization. Therefore, these outputs not only distinguish domain attribution but also serve as soft indicators of alignment quality in the feature space, which can be used to help the model identify regions in the target domain where adaptation remains insufficient. MFRR is designed to leverage these cues to guide feature-extractor optimization.

As illustrated in Figure 4, MFRR is integrated into the segmentation network before obtaining segmentation predictions. Blue dashed arrows represent the data flow of the target domain, and orange solid arrows represent the source domain. During training, intermediate features of the target domain (F_T) in stage 4 proceed to generate the map (μ) to identify regions with low alignment levels. This map is used to re-weight the target features as $\mu F_T + F_T$, and the result is passed forward to generate pixel-level prediction. In contrast, source-domain features do not participate in this process and are directly propagated forward. During inference, MFRR is not activated, as this stage does not involve the domain discriminator; thus, no alignment confidence is required.

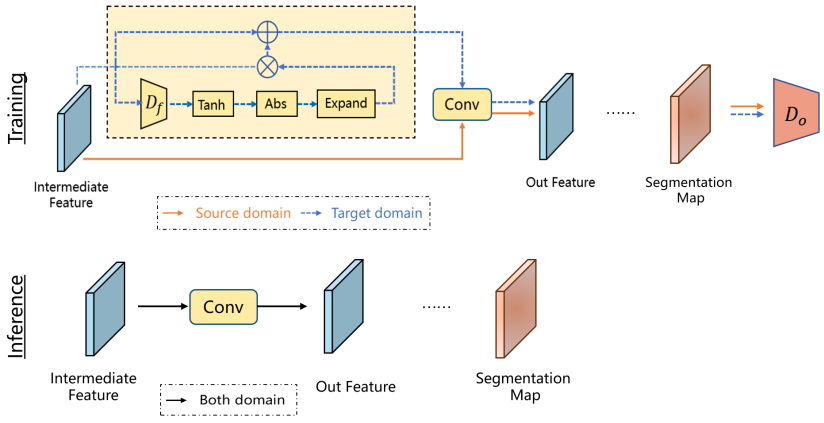


Figure 4. Illustration of data flow and forward computation through the MFRR in the training phase (**top**) and inference phase (**bottom**).

3.2.5. Discriminator Network

For domain adaptation, two discriminators are incorporated to implement adversarial learning at the feature and output levels. The output-level discriminator (D_o) operates on the fused predictions from the GPSU and performs output-level alignment by adversarially distinguishing between source predictions and target predictions. On the other hand, the other discriminator (D_f) is embedded within the MFRR. D_f receives the final-stage features of target data and performs feature-level adversarial learning to align source-domain and target-domain distributions. During adversarial training, D_f and D_o are optimized to distinguish the domain of input features or predictions, while the feature extractor is trained to confuse them. As a result, the learned representations from both domains become increasingly indistinguishable, promoting domain-invariant and transferable features. This adversarial interplay helps reduce domain discrepancy and enhances the segmentation network's generalization to the target domain.

3.3. Training Procedure

The adversarial training process is formulated as a min–max adversarial game between the segmentation network and the domain discriminator. The segmentation network is trained to learn domain-invariant representations that can fool the discriminator, while the discriminator attempts to distinguish features extracted from the source and target domains. Training is performed in an alternating fashion, as detailed below.

3.3.1. Segmentation Network Training

The segmentation network is trained using both the supervised segmentation loss in the source domain and the adversarial loss in the target domain. The supervised segmentation loss in the labeled source domain serves to learn representative semantic features and provides a reliable initialization for subsequent domain adaptation. The adversarial loss guides the segmentation network to generate target-domain features that can deceive the discriminator into classifying them as source-domain features.

To supervise the segmentation learning in the source domain, cross-entropy (CE) loss is applied to the semantic predictions. The segmentation loss ($\mathcal{L}_{seg}$) is defined as follows:

$$\mathcal{L}_{seg}(P_S, Y_S) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} Y_S^{(h,w,c)} \log P_S^{(h,w,c)} \tag{8}$$

where P_S refers to the predictions for the source data and Y_S refers to the true label. H, W, and C denote the height, width, and number of semantic classes, respectively.

This loss is applied to the predictions (P_{dsm}, P_{rsi}, and P_g) of source data. P_{dsm} and P_{rsi} denote the semantic predictions generated by the DSM and RSI branches, respectively, and P_g denotes the selected prediction produced by the GPSU. The overall segmentation objective is therefore defined as follows:

$$\min_{\theta_E} \mathcal{L}_{seg}(P_{Sdsm}, Y_S) + \mathcal{L}_{seg}(P_{Srsi}, Y_S) + \mathcal{L}_{seg}(P_{Sg}, Y_S) \tag{9}$$

where θ_E represents the parameters of the segmentation network.

During the adversarial training stage, the parameters of the discriminators are fixed while the segmentation network is updated. The binary cross-entropy (BCE) loss is employed to measure the discrepancy between the discriminator predictions and the desired

domain labels. The adversarial loss ($\mathcal{L}_{adv}^{E}$) used to update the segmentation network is defined as follows:

$$\min_{\theta_E} \mathcal{L}_{adv}^{E} = \sum \log\left(D_o\left(\text{softmax}(P_{Tg})\right)\right) + \sum \log\left(D_f\left(\text{softmax}(f_T)\right)\right) \tag{10}$$

where P_{Tg} refers to the final prediction of the target input and f_T refers to the intermediate feature of the target input.

3.3.2. Discriminator Network Training

In the next step, the discriminators are trained while the parameters of the segmentation network are fixed. The objective of the discriminators is to distinguish the domain origin of the input features or predictions. In this study, the discriminators are trained to output 0 for source-domain inputs and 1 for target-domain inputs.

The training objectives ($\mathcal{L}_{adv}^{D}$) of the output-level discriminator (D_o) and the feature-level discriminator (D_f) are defined as follows:

$$\min_{D_o} \mathcal{L}_{adv}^{D_o} = L_{BCE}\left(D_o(P_{Sg}), 0\right) + L_{BCE}\left(D_o(P_{Tg}), 1\right)$$
$$\min_{D_f} \mathcal{L}_{adv}^{D_f} = L_{BCE}\left(D_f(f_S), 0\right) + L_{BCE}\left(D_f(f_T), 1\right)$$

where P_{Sg} refers to the final prediction of the source input and f_S refers to the intermediate feature of the source input.

4. Experimental Results and Discussion

4.1. Datasets

In this study, we evaluated the proposed method on three publicly available remote sensing datasets: the Potsdam and Vaihingen datasets from the ISPRS Semantic Labeling benchmark [48] and the GAMUS dataset [24]. Each dataset provides reliable annotated remote sensing imagery paired with corresponding digital surface models or normalized DSMs, enabling effective evaluation of semantic segmentation performance. The datasets were collected from regions in different countries, each exhibiting distinct landscape patterns, urban planning styles, and land-cover distributions. A detailed comparison of their specific attributes is provided in Table 1. Together, these datasets offer a diverse set of challenges, further highlighting the need for models that can adapt to varying semantic and visual characteristics, making the evaluation more comprehensive and representative of real-world applications.

Table 1. Comparison of attributes of the datasets.

Property	Potsdam	Vaihingen	GAMUS
Landscape	Urban	Village	Multiple Cities
Pixel	6000 × 6000	2000 × 2500 or so	1024 × 1024
GSD	5 cm/pixel	9 cm/pixel	33 cm/pixel
Channels	IR-R-G/R-G-B/R-G-B-IR	IR-R-G	R-G-B
Imp. surf. ratio	29.55%	27.72%	42.14%
Building ratio	25.70%	26.40%	29.67%
Tree ratio	15.56%	22.91%	12.12%
Low veg. ratio	22.51%	20.94%	16.08%
Car ratio	1.79%	1.27%	-
Clutter ratio	4.89%	0.77%	-

The Potsdam and Vaihingen datasets are widely used for evaluating semantic segmentation models due to their high-quality annotations and diverse land-cover types. The annotations in these datasets were originally divided into six categories: "clutter/background", "impervious surfaces", "car", "tree", "low vegetation", and "building". For the purpose of our experiments, we focused on the latter five categories when it came to evaluation. GAMUS is a recently introduced multimodal remote sensing dataset of the Technical University of Munich. It is annotated with six slightly different land-cover categories: ground, low vegetation, building, water, road, and tree. In this study, 500 images were selected from the Philadelphia (PHL) subset due to its relatively balanced class distribution. For alignment with the label space of the ISPRS datasets, the ground and road classes were merged, and the water class was treated as background.

For data pre-processing, we first normalized both the RSI and DSM data to ensure consistent value ranges across modalities. The original high-resolution images were then divided into smaller patches to facilitate model training. Specifically, for all datasets, RGB or IRRG images, along with their corresponding DSM data and ground-truth labels, were cropped into patches of 512 × 512 pixels without overlap. Following this pre-processing pipeline, we obtained a total of 5472 image patches from the Potsdam dataset, 742 patches from the Vaihingen dataset, and 2000 patches from the GAMUS dataset. We also applied random vertical or horizontal rotation during training as data augmentation before feeding the data into the model.

4.2. Experimental Settings

To enhance scalability for large-scale and high-resolution remote sensing data, we adopted depthwise separable convolutions [49] in the the ResNet-34 network [50] of the DSM branch, effectively reducing the parameter count and computational load while maintaining performance. Meanwhile, the DeeplabV2 framework [27] with a ResNet-101 backbone [50] we used to encode the RSI features was pre-trained on the ImageNet dataset. The ASPP [27] was used as a classifier for both branches with sampling rates of [6, 12, 18, 24], as was the case for AdaptSegNet [9]. The stride of convolution layers in the last two residual blocks of the DSM branch was adjusted to 2 and 4, respectively, to ensure that the output feature maps matched the size of the other branch.

The output-level discriminatorwas set up the same as AdaptSegNet [9]. Specifically, the network consisted of five 4 × 4-kernel convolutional layers, and the stride was 2. These layers included [64, 128, 256, 512, 1] channels, respectively. In contrast, the feature-level discriminator consisted of four 3 × 3-kernel convolutional layers, and the stride was 2; the number of channels for these convolutional layers was set to [1024, 512, 256, 1]. The output from each convolutional layer underwent processing through a Leaky ReLU layer, utilizing a parameter value of 0.2.

We selected PyTorch 1.11.0 as the deep learning framework, which was run on Python 3.8 and CUDA 11.3 under a Ubuntu 20.04 operating system with a single NVIDIA RTX 4090 24 GB GPU. During the training of the segmentation model, we used a stochastic gradient descent optimizer with Nesterov acceleration, with the momentum set to 0.9, the weight decay set as 5×10^{-4}, and the initial learning rate set as 2.5×10^{-4}. For the training of the discriminator, We adopted the settings from AdaptSegNet [9]. The batch size and number of epochs were set to 4 and 50, with early stopping adopted to avoid overfitting.

4.3. Evaluation Metrics

For quantitative evaluation, we adopted two widely used metrics in semantic segmentation: the F1 score and the Intersection over Union (IoU), which jointly assess the model's classification accuracy and spatial precision. The F1 score is defined as the harmonic mean

of precision and recall. It provides a balanced evaluation of model performance by considering both false positives and false negatives, making it particularly suitable for imbalanced datasets. Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. Precision and recall are defined as

$$precision = \frac{TP}{TP + FP} \tag{11}$$

$$recall = \frac{TP}{TP + FN}. \tag{12}$$

The F1 score is then computed as

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}. \tag{13}$$

The IoU evaluates the overlap between the predicted segmentation and the ground-truth region, directly reflecting the spatial accuracy of the model. It is defined as

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \tag{14}$$

A higher IoU value indicates better alignment between the predicted and ground-truth regions.

In the following experiments, we report the IoU and F1 scores for each category, as well as their mean values, i.e., mIoU and mF1, to evaluate the model's performance from both category-specific and overall perspectives.

4.4. Comparative Analysis

4.4.1. Experiments Configuration

Three sets of experiments were conducted to evaluate the performance of the proposed method by comparing it with the source-only baseline and several leading adversarial learning approaches—namely, AdaptSegNet [9], Advent [10], CLAN [51], DAST [52], DADA [21], CTRL [22], DAFormer [53], and MBATA-GAN [54]. The configuration details are listed in Table 2.

Table 2. Cross-domain dataset configurations for three sets of experiments.

Setting	Source Domain	Target Domain
(a)	Vaihingen IRRG	Potsdam IRRG
(b)	Potsdam RGB	Vaihingen IRRG
(c)	Potsdam RGB	GAMUS

The labeled data and labels from the source domain, as well as unlabeled data from the target domain, were used in the training process. As for the testing process, the ground-truth images from the target domain were used to verify the performance of the model.

4.4.2. Comparative Results on Vaihingen IRRG to Potsdam IRRG

In this experiment, the Vaihingen dataset was used as the source domain, and the Potsdam dataset with IRRG channels was used as the target domain. Both the source- and target-domain images consist of IRRG channels. However, their GSD are different, resulting in distinct spatial resolutions.

Table 3 presents the quantitative results of the cross-domain semantic segmentation task on the Vaihingen IRRG and Potsdam IRRG datasets. The best results are highlighted in bold.

Table 3. Quantitative results (%) of the Vaihingen IRRG → Potsdam IRRG experiment.

Method	Metric	Imp. Surf.	Building	Tree	Low Veg.	Car	Avg.
Source only	F1	31.07	54.23	50.93	19.38	47.38	40.60
	IoU	18.39	37.21	34.17	10.73	31.04	26.31
AdaptSegNet	F1	65.83	67.19	56.35	54.98	60.89	61.05
	IoU	49.06	50.60	39.23	37.92	43.77	44.12
Advent	F1	66.57	69.01	56.88	53.44	61.39	61.46
	IoU	49.89	52.69	39.75	36.46	44.29	44.62
CLAN	F1	69.64	68.34	61.45	44.15	59.67	60.65
	IoU	53.42	51.91	44.35	28.33	42.52	44.11
DAST	F1	71.56	74.87	59.43	**60.01**	57.38	64.65
	IoU	55.72	59.84	42.28	**42.87**	40.23	48.19
DADA	F1	61.80	69.65	60.06	58.15	65.79	63.09
	IoU	44.72	53.44	42.91	41.00	49.02	46.22
CTRL	F1	67.45	57.87	49.38	42.91	65.49	56.62
	IoU	50.89	40.72	32.78	27.32	48.68	40.08
MBATA-GAN	F1	71.96	80.22	47.71	59.92	**67.53**	65.47
	IoU	56.20	66.97	31.33	42.78	**50.98**	49.65
DAFormer	F1	73.86	75.83	47.49	43.32	36.18	55.34
	IoU	58.56	61.07	31.14	27.65	22.09	40.10
Ours	F1	**78.73**	**84.56**	**66.26**	56.34	57.56	**68.69**
	IoU	**64.92**	**73.25**	**49.55**	39.22	40.41	**53.47**

Bold values indicate the best performance in each column.

Our proposed method achieved the best overall performance, with an mF1 score of 68.69% and an mIoU of 53.47%, outperforming all comparative methods by a clear margin. Specifically, compared with the second-best method, MBATA-GAN, our method yielded an absolute gain of 3.22% in mF1 and 3.82% gain in mIoU, which validates the effectiveness of our network for remote sensing image segmentation. From the category-wise perspective, our method excels particularly in the impervious surface and building classes, achieving F1 scores of 78.73% and 84.56% and IoU scores of 64.92% and 73.25%, respectively, which are significantly higher than those of all other competing methods. This demonstrates that our approach can effectively alleviate the domain shift of large-scale structured objects in urban scenes. For the tree and low vegetation classes, our method obtained F1 scores of 66.26% and 56.34%, respectively, also surpassing many comparative methods, indicating its capability in fine-grained feature adaptation for vegetation categories. Although the F1 score of 57.56% for the car class is not the highest, it still maintains a competitive level.

In addition, several prediction examples are illustrated in Figure 5. The segmentation results of the "source-only" method shown in column (d) are highly noisy. After the adversarial training of domain adaptation, the segmentation results are significantly improved in column (e) and column (f). With the additional support of DSM data, classification errors were significantly reduced, as shown in column (i). It is clear that compared to the source-only model, the segmentation accuracy of our method was significantly improved. Classes such as buildings and roads, which were often misclassified in the source-only model, were better delineated by our method. This suggests that the model benefited from both domain adaptation and DSM-enhanced spatial cues. Compared to the source-only model, our method produced clearer object boundaries, indicating improved spatial precision.

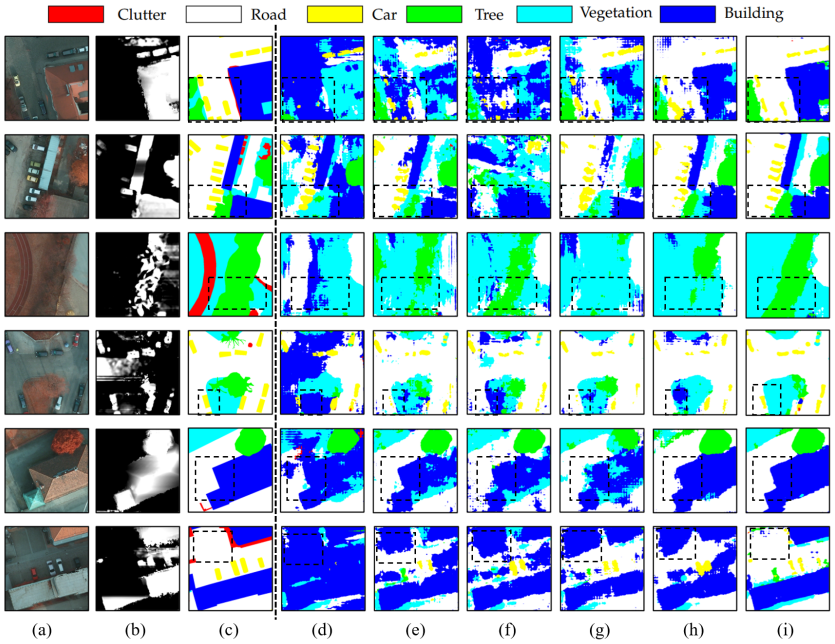


Figure 5. Visualization of segmentation results of the Vaihingen IRRG-to-Potsdam IRRG experiment: Columns representing the (**a**) RSI, (**b**) DSM, (**c**) ground-truth, and segmentation results of the (**d**) source-only, (**e**) AdapSegNet, (**f**) ADVENT, (**g**) CLAN, and (**h**) DAST models in comparison with (**i**) our model. Square regions highlight representative areas for better comparison.

4.4.3. Comparative Results on Potsdam RGB and Vaihingen IRRG Datasets

In this experiment, the Potsdam dataset with RGB channels was used as the source domain, and the Vaihingen dataset was used as the target domain. Differences in the spectral-band configurations mean that the datasets not only differ in terms of spectral information but also in how pixel values are encoded. As a result, the domain shift between the two datasets is relatively large, which poses challenges for models trained on the source domain to generalize effectively to the target domain.

The experimental results are reported in Table 4. The best results are highlighted in bold. Our proposed method attained the best overall performance, with an mF1 score of 65.17% and an mIoU of 50.87%. Specifically, compared with the second-best method, DAST, our method achieved an absolute improvement of 4.08% in mF1 and a 1.29% improvement in mIoU, which further validates the effectiveness and generalization ability of our network across different spectra. From the category-wise perspective, our method exhibits overwhelming advantages in the impervious surface and building classes, obtaining F1 scores of 81.95% and 89.77% and IoU scores of 67.30% and 80.29%, respectively, which are significantly higher than those of all other competing methods. In contrast, performance on the remaining three categories was comparatively lower. For the tree and low vegetation classes, our method achieved F1 scores of 56.15% and 66.38%, respectively, also outperforming most comparative methods. However, the F1 score of 31.61% for the car class is lower than that of some compared methods. We hypothesize that this discrepancy may stem from class imbalance, as the under-representation of certain classes in both datasets limits the model's ability to learn robust and discriminative features for those categories.

To further illustrate the effectiveness of our approach, representative visual results are presented in Figure 6. The segmentation results of the "source-only" method, shown in column (d), are inaccurate and suffer from significant misclassifications due to the large domain shift. Under adversarial training for domain adaptation, the results were improved substantially, as seen in columns (e)–(h). Moreover, with the integration of depth information, the classification accuracy was significantly improved, as demonstrated

in column (i). These qualitative comparisons clearly show that the proposed method produced more accurate and coherent segmentation results than existing single-modality UDA methods. These visual improvements align with the quantitative gains reported in Table 4, confirming that our method not only improved quantitative scores but also produced semantically meaningful and visually coherent outputs.

Table 4. Quantitative results (%) of the Potsdam RGB → Vaihingen IRRG experiment.

Method	Metric	Imp. Surf.	Building	Tree	Low Veg.	Car	Avg.
Source only	F1	3.86	31.85	20.25	40.85	7.27	20.82
	IoU	1.97	18.94	11.27	25.67	3.77	12.32
AdaptSegNet	F1	65.36	73.59	37.38	68.14	38.43	56.58
	IoU	48.54	58.22	22.99	51.67	23.78	41.04
Advent	F1	66.47	77.75	40.09	65.84	43.26	58.68
	IoU	49.78	63.60	25.07	49.07	27.60	43.02
CLAN	F1	65.55	77.24	42.21	55.00	12.83	50.57
	IoU	48.76	62.92	26.75	37.93	6.86	36.64
DAST	F1	74.21	84.00	**58.90**	60.13	48.18	65.09
	IoU	59.00	72.42	**41.74**	42.99	31.73	49.58
DADA	F1	59.51	72.46	38.51	59.94	42.34	54.55
	IoU	42.36	56.81	23.85	42.79	26.86	38.53
CTRL	F1	65.97	63.75	30.37	53.22	17.28	46.12
	IoU	49.22	46.79	17.90	36.25	9.46	31.92
MBATA-GAN	F1	57.52	72.91	39.87	60.02	**58.29**	57.72
	IoU	40.37	57.37	24.90	42.88	**41.13**	41.33
DAFormer	F1	76.59	77.47	37.97	**69.39**	32.07	58.70
	IoU	62.06	63.23	23.44	**53.12**	19.10	44.19
Ours	F1	**81.95**	**89.77**	56.15	66.38	31.61	**65.17**
	IoU	**67.30**	**80.29**	39.56	49.48	17.73	**50.87**

Bold values indicate the best performance in each column.

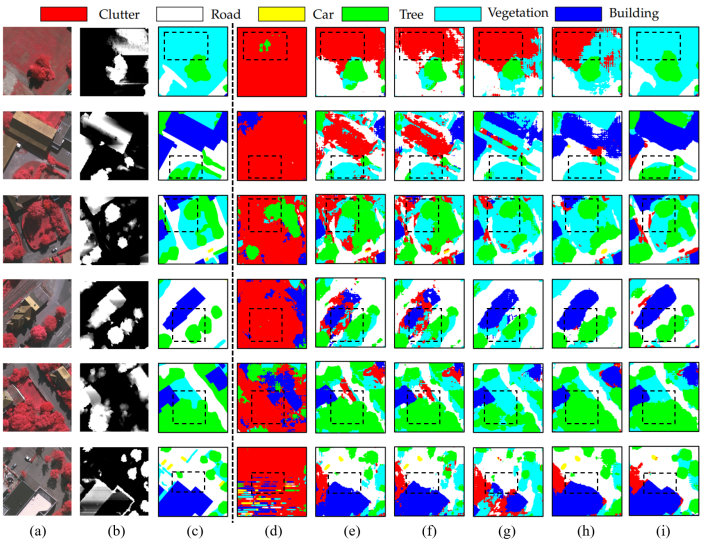


Figure 6. Visualization of segmentation results of the Potsdam RGB to Vaihingen IRRG experiment: Columns representing (**a**) RSI, (**b**) DSM, (**c**) ground-truth, and segmentation results of the (**d**) source-only, (**e**) AdapSegNet, (**f**) ADVENT, (**g**) CLAN, and (**h**) DAST models in comparison with (**i**) ours model. Square regions highlight representative areas for better comparison.

4.4.4. Comparative Results on Potsdam RGB and GAMUS Datasets

In this experiment, the Potsdam RGB dataset served as the source domain, and the GAMUS dataset was used as the target domain. Although both domains contain RGB imagery, the two domains exhibit distinct urban layouts and significant resolution differences due to differing GSD. This discrepancy introduces a domain shift not only in appearance but also in object scale and texture granularity, making it a challenging domain adaptation scenario. Since there is no "car" category in the target domain, we merged the "car" category of the source domain with the "impervious surfaces" category during training. Additionally, due to the presence of a "clutter" category in the source domain and the lack of one-to-one correspondence with target-domain categories, the CLAN method, which involves category-specific discrimination, was not applicable. Experiments confirmed that CLAN does not perform effectively in this scenario, so it was excluded from the comparative analysis.

The experimental results in Table 5 show that the proposed method performed excellently in both mIoU and mF1 score, with scores of 50.53% and 65.75%, respectively. Specifically, compared with the second-best method, AdaptSegNet, our method yielded an absolute gain of 8.99% in average F1 and 9.17% in average IoU. From the category-wise perspective, our method exhibits advantages across all classes. For the impervious surface and building categories, our method obtained F1 scores of 78.59% and 79.40% and IoU scores of 64.74% and 65.84%, respectively, which are significantly higher than those of all other competing methods. This further validates that elevation information can help domain adaptation tasks in situations with significant domain differences, allowing the model to learn more universal and immutable features that do not vary across domains. It is worth noting that the DAST method, which performed well in the former two tasks, did not perform satisfactorily in this task. However, our method still achieved significant improvement, indicating that our method has greater generalization ability than the compared methods.

Table 5. Quantitative results (%) of the Potsdam RGB → GAMUS experiment.

Method	Metric	Imp. Surf.	Building	Tree	Low Veg.	Avg.
Source only	F1	60.57	54.95	48.80	40.53	51.21
	IoU	43.44	37.88	32.28	25.42	34.75
AdaptSegNet	F1	70.39	70.94	55.31	30.41	56.76
	IoU	54.31	54.97	38.23	17.93	41.36
Advent	F1	70.29	71.03	55.56	24.57	55.36
	IoU	54.19	55.07	38.46	14.00	40.43
DAST	F1	70.07	67.66	54.07	20.85	53.16
	IoU	53.92	51.13	37.05	11.64	38.44
MBATA-GAN	F1	61.53	56.35	35.12	66.07	54.77
	IoU	44.44	39.23	21.30	49.34	38.58
DADA	F1	61.89	54.49	48.98	35.06	50.11
	IoU	44.81	37.45	32.43	21.25	33.99
CTRL	F1	55.09	56.67	53.73	9.13	43.66
	IoU	38.02	39.54	36.74	4.78	29.77
Ours	F1	**78.59**	**79.40**	**57.85**	**47.15**	**65.75**
	IoU	**64.74**	**65.84**	**40.69**	**30.85**	**50.53**

Bold values indicate the best performance in each column.

The visualization results of this set of experiments are presented in Figure 7. Compared to other methods, our approach clearly leads to a notable reduction in the red-colored areas, which indicate regions where the model fails to produce valid predictions. This

demonstrates the superior capability of our method in handling challenging or ambiguous regions. In addition, consistent with previous results, our model outperformed other approaches in accurately delineating roads and buildings, showing enhanced precision in structured and fine-grained semantic categories. These improvements can be attributed to the effective integration of multimodal features and the domain alignment mechanisms introduced in our framework.

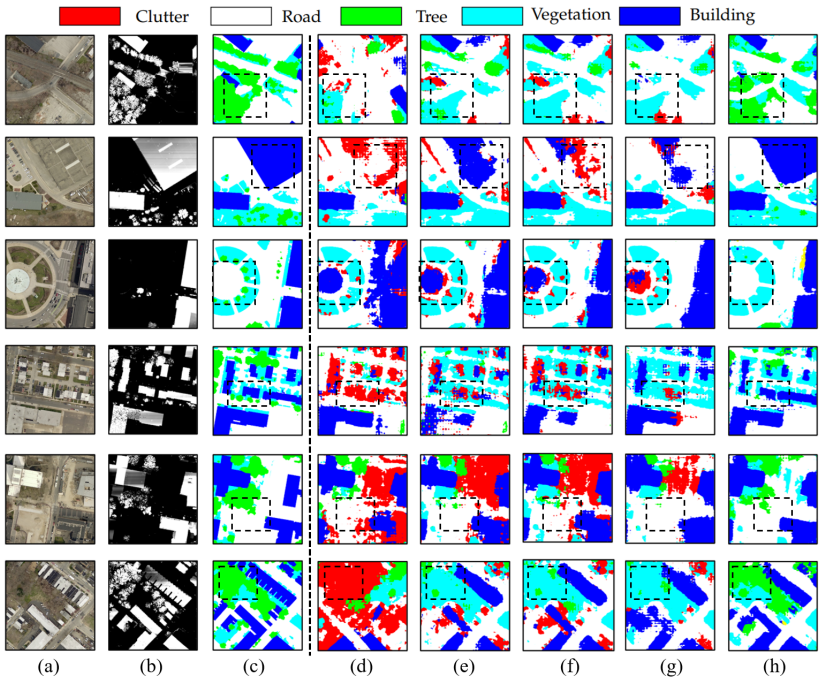


Figure 7. Visualization of segmentation results of the Potsdam RGB to GAMUS experiment: Columns representing (**a**) RSI, (**b**) DSM, and (**c**) ground-truth and segmentation results of the (**d**) source-only, (**e**) AdapSegNet, (**f**) ADVENT, and (**g**) DAST methods in comparison with (**h**) our method. Square regions highlight representative areas for better comparison.

In summary, the experimental results show that our method can improve the segmentation accuracy of the model in the target domain, especially for categories with clear depth properties such as buildings and impervious surfaces. Therefore, our method is promising to be further applied to downstream tasks like building extraction, road extraction, and urban planning.

4.5. Efficiency and Complexity Analysis

To further evaluate the efficiency of different methods, we compared them in terms of the duration of a complete training cycle (Training Time), peak GPU memory usage (Peak GPU Use), the number of parameters (Params), and computational complexity (FLOPs), as shown in Table 6. All methods were tested with an input resolution of 512 × 512, except for MBATA-GAN *, for which we adopted the original 256 × 256 setting, since it encounters CUDA out of memory when the input is resized to 512 × 512. Thus, its efficiency metrics are not directly comparable with those of other methods.

Previous methods employed a single-branch architecture, processing only one modality, while our method uses a two-branch architecture to process two modalities. Our method takes the most FLOPs (234.14 G) among comparable methods, indicating strong modeling capacity for complex domain shifts, while maintaining a moderate parameter count (55.86 M) only slightly larger than that of DAST (46.92 M) and DADA (46.68 M). DAFormer and MBATA-GAN adopt a Transformer structure, leading to higher parameter counts. Notably, our method consumes only 12.733 G of peak GPU memory, which sig-

nificantly lower than the peak GPU memory of AdaptSegNet (18.10 G), DAST (18.32 G), CTRL (16.99 G), and MBATA-GAN (16.97 G). The superior memory efficiency stems from extensive use of depthwise separable convolutions, which decompose standard convolutions into depthwise and pointwise operations, significantly reducing memory access patterns and intermediate tensor storage requirements while maintaining high computational capacity. Regarding the training time, our method takes 154.77 s per cycle, which is longer than lightweight methods like DADA (71 s) and Advent (89.49 s) but shorter than CTRL (331.5 s). The longer training time of CTRL is likely due to its self-training mechanism. Overall, our method achieves a favorable trade-off between accuracy, computational complexity, and memory efficiency, demonstrating its suitability for deployment on resource-constrained devices.

Table 6. Comparison of efficiency.

Method	FLOPs (G)	Params (M)	Peak GPU Use (GB)	Training Time (s)
AdaptSegNet	186.12	43.16	18.10	84.73
Advent	186.12	43.16	18.10	89.49
CLAN	187.05	43.38	18.08	89.03
DAST	202.01	46.92	18.32	113.43
DADA	200.97	46.68	18.12	71
CTRL	202.86	47.13	16.99	331.5
DAFormer	116.56	85.16	17.72	85.7
MBATA-GAN *	174.20	144.43	16.97	63.51
Ours	234.14	55.86	12.73	154.77

* MBATA-GAN is tested with input resolution 256 × 256 while others adopt 512 × 512, so its efficiency metrics are not directly comparable with others.

4.6. Ablation Experiments

4.6.1. Component Ablation

To verify the effectiveness of each component in our method, we conducted comprehensive ablation studies of the adaptation task on the Vaihingen IRRG and Potsdam IRRG datasets. We progressively added key components of the model to evaluate their individual contributions. Specifically, we considered the following ablation settings: (1) baseline, using only the RSIs of the source domain without adversarial training; (2) +AFA, introducing DSM by AFA without adversarial training; (3) RSI+AT, using adversarial training without DSM input; (4) dual-branch +DSM (concat features) +AT, introducing DSM by concat features with adversarial training; (5) +AFA +AT, introducing DSM by AFA with adversarial training; (6) +AFA +Simple Fusion (Add) +AT, with predictions of the two branches fused via element-wise addition; (7) +AFA +Simple Fusion (Concat) +AT, with predictions of the two branches fused via channel-wise concatenation; and (8) the complete model, corresponding to our complete model.

The experimental results in Table 7 show a clear upward trend in performance as more modules were progressively added. The first two ablation settings evaluate models trained only on the source domain. Even without adversarial training on the target domain, incorporating DSM data significantly boosted the segmentation accuracy, with an mIoU gain of 8.5%. This observation demonstrates that depth information helps the model learn more generalizable and domain-transferable representations. The remaining ablation experiments were conducted under adversarial training, which enabled the model to exploit knowledge from the target domain and, thus, enhanced the segmentation performance in the target domain. Directly concatenating DSM features at each stage yielded an mIoU of 47.65%. In comparison, employing our proposed cross-attention fusion module, i.e., the AFA, further elevated the mIoU to 49.83%, indicating that the AFA can better model the semantic interaction between multi-level semantic features and DSM information. These

ablation results verify the effectiveness of the proposed AFA module and the advantages of integrating geometry-aware information with semantic features. Regarding the fusion of predictions from the two branches, we investigated three strategies: element-wise addition, channel-wise concatenation, and our proposed gated prediction selection unit. The two simple fusion strategies led to a performance drop compared to using only the AFA feature fusion module, whereas our GPSU achieved an improvement. Compared with the baseline, our full model achieved a substantial mIoU gain of 27.16%. The full model also presents stronger cross-domain generalization ability, which further validates the rationality and superiority of the overall framework.

Table 7. IoU (%) results of ablation components (Vaihingen IRRG → Potsdam IRRG).

Ablation Setting	DSM	Fusion Strategy	Imp. Surf.	Building	Tree	Low Veg.	Car	Avg.
Source only	–	–	18.39	37.21	34.17	10.73	31.04	26.31
+AFA (w/o AT)	✓	–	41.41	50.84	39.60	30.57	11.64	34.81
RSI+AT	–	–	50.84	53.18	43.00	39.45	42.89	45.87
+DSM dual-branch (concat) +AT	✓	Concat Feature	63.30	68.24	34.69	34.11	37.9	47.65
+AFA +AT	–	–	66.99	73.60	36.28	29.90	42.39	49.83
+AFA +Simple Fusion +AT	✓	Add Prediction	61.64	71.78	45.67	28.84	24.99	46.58
+AFA +Simple Fusion +AT	✓	Concat Prediction	65.58	65.77	45.64	23.57	37.19	47.55
Complete Model	✓	Ours	64.92	73.25	49.55	39.22	40.41	53.47

AT refers to adversarial training; DSM (Concat) indicates whether dual-branch DSM with concatenated features is enabled; Fusion Strategy lists the fusion method adopted in each setting. Check mark indicates the corresponding settings are used.

4.6.2. Fusion Stage Analysis

To further investigate the impact of multimodal fusion stages, we conducted an ablation study on different feature aggregation positions. Previous studies [55] have shown that different layers of deep neural networks capture features with varying degrees of generality. Leveraging this insight, we experimented with fusion in different residual blocks:

- AFA1234 denotes feature fusion after all four residual blocks;
- AFA234 applies fusion after blocks 2 to 4;
- AFA34 corresponds to fusion after blocks 3 and 4;
- AFA4 performs fusion only after the final block;
- Single-branch refers to using only the RSI data without any DSM input.

The results in Table 8 show that AFA1234 achieved the best overall performance, especially in categories such as impervious surfaces and buildings. This suggests that integrating features at multiple semantic levels enabled the model to leverage both shallow spatial details and deep semantic context. In contrast, fusing only at the final stage (AFA4) yielded the worst results, even exhibiting inferior performance relative to the single-modality baseline in some cases. A potential reason is that high-level features derived from RSI and DSM exhibit strong abstraction but weak alignment, posing challenges for effective cross-modal fusion. These findings underscore that hierarchical fusion not only strengthens cross-modal feature alignment but also facilitates more robust gradient flow throughout the network, enabling DSM features to participate more fully in representation learning.

Table 8. IoU (%) results of different fusion stages (Vaihingen IRRG → Potsdam IRRG).

	Imp. Surf.	Building	Tree	Low Veg.	Car	Avg.
AFA1234	68.42	72.99	43.20	27.31	39.68	50.32
AFA234	68.24	70.63	40.64	24.41	44.40	49.66
AFA34	69.46	67.33	38.15	28.24	47.13	50.06
AFA4	58.86	43.43	39.80	24.52	33.58	40.04
Single branch	50.84	53.18	43.00	39.45	42.89	45.87

4.6.3. Fusion Strategy Analysis

We designed three cross-attention fusion strategies to explore the optimal interaction between RSI and DSM features, as detailed in Table 9:

Experimental results show that Strategy 1 achieved the highest mIoU of 53.47%, surpassing Strategy 2 and Strategy 3 by 11.34% and 13.19% points, respectively. The result suggests that RSI features provide more reliable semantic cues to guide the attention mechanism, whereas DSM features are better suited to provide auxiliary geometric and elevation information. In contrast, employing DSM as the query is less effective, mainly owing to its limited semantic richness and ambiguous structural cues. This demonstrates the importance of modality-aware attention design and verifies that the semantically rich modality (RSI) should act as the guide for feature selection in cross-modal fusion.

Table 9. Comparison of different cross-attention fusion strategies for RSI and DSM modalities.

Strategy	RSI Role	DSM Role	Residual Operation	mIoU (%)
1	Query (Q)	Key & Value (KV)	Add to RSI features	53.47
2	Key & Value (KV)	Query (Q)	Add to DSM features	42.13
3	Bidirectional	Bidirectional	Add to own features	40.28

A plausible explanation lies in the intrinsic roles of Q, K, and V in the attention mechanism. In our cross-attention module, the query provides semantic guidance for retrieving relevant information, while keys and values from the complementary modality supply the indexed content to be aggregator. Specifically, Q defines the alignment criterion and feature-matching focus, K computes the inter-modal relevance map, and V carries the concrete features to be fused. Benefiting from dense semantic cues, RSI generates highly discriminative queries for distinguishing various ground objects. When RSI serves as the query, the attention mechanism appropriately theoretically weights DSM features according to their semantic relevance.

5. Conclusions

In this paper, we have explored the possibility of mining and utilizing more information in unsupervised domain adaptation for the semantic segmentation of remote sensing images. To this end, we integrated depth information—namely, DSM data—with semantic information through the use of an adaptive feature aggregator. Considering the semantic ambiguity of DSM data, we exploited the gated prediction selection unit to obtain a more reliable prediction. In addition, to take advantage of target-domain images under the unsupervised setting, we utilized a misalignment-focused residual refinement module to determine the poorly aligned areas of the target domain and increase the weight of the corresponding area. These strategies enabled the model to integrate complementary information from different modalities, thereby facilitating the learning of more domain-invariant feature spaces. Our proposed approach, which jointly leverages DSM and RSI data, outperformed single-modality UDA methods and achieved improved pixel-level classification accuracy. Across multiple cross-domain experiments in remote sensing semantic segmentation, our method consistently achieved better performance than competing UDA approaches, highlighting its generalization across diverse domain shifts.

However, there remain opportunities for further enhancement. First, the method was designed and validated under the premise that DSM data are accessible and reliable, as the geometric information encoded in DSM serves as a critical cue that is complementary to the optical features. Performance gains were also limited for classes with weak geometric features, such as the low vegetation and car categories, because DSM provides little discriminative height and structural information for these categories. This deficiency is likely

attributable to the fact that our depth-aware module was designed to capture structured spatial layouts, making it more effective for large-scale, compact objects such as buildings but less robust for tiny, scattered objects. In comparison, objects in the building, car, and low vegetation categoriesusually exhibit smaller sizes, sparser and irregular depth signals, and more severe occlusion in remote sensing scenarios. In addition, the multi-branch design leads to lower computational efficiency compared with conventional single-branch networks due to the need for extra feature extraction and fusion operations. While adversarial training enhances feature distribution alignment and domain invariance, it is known to introduce inherent challenges including training instability and mode collapse. However, our proposed method does not employ advanced loss design, dynamic weighting, or dedicated regularization techniques to mitigate adversarial training instability and mode collapse. These limitations highlight the contextual constraints of the approach and point to meaningful directions for future work to enhance robustness, efficiency, and stability.

Author Contributions: Conceptualization, L.N., X.L. (Xiaoxuan Liu) and E.Z.; methodology, L.N. and Y.Z.; software, L.N., E.Z. and H.S.; validation, L.N. and X.L. (Xiaoxuan Liu); resources, X.L. (Xiaohe Li) and H.W.; writing—original draft preparation, L.N.; writing—review and editing, X.L. (Xiaoxuan Liu) and E.Z.; supervision, H.W. and L.W.; funding acquisition, X.L. (Xiaoxuan Liu), J.J. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Program of the Chinese Academy of Sciences under Grants KGFZD-145-25-38 and RCJJ-145-24-13 and the Science and Disruptive Technology Program under Grant AIRCAS2024-AIRCAS-SDTP-03.

Data Availability Statement: The Potsdam and Vaihingen datasets from the ISPRS 2D Semantic Labeling benchmark are available at https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab, accessed on 30 March 2026. The GAMUS dataset is available at https://huggingface.co/datasets/earthflow/GAMUS, accessed on 30 March 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
2. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
3. Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; Krishnan, D. Unsupervised pixel-level domain adaptation with generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3722–3731.
4. Chen, J.; Zhu, J.; Guo, Y.; Sun, G.; Zhang, Y.; Deng, M. Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
5. Ni, H.; Liu, Q.; Guan, H.; Tang, H.; Chanussot, J. Category-level assignment for cross-domain semantic segmentation in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
6. Tasar, O.; Happy, S.; Tarabalka, Y.; Alliez, P. SemI2I: Semantically consistent image-to-image translation for domain adaptation of remote sensing data. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 1837–1840.
7. Hoffman, J.; Wang, D.; Yu, F.; Darrell, T. FCNs in the Wild: Pixel-Level Adversarial and Constraint-Based Adaptation. *arXiv* **2016**, arXiv:1612.02649.
8. Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T. Adversarial discriminative domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7167–7176.
9. Tsai, Y.H.; Hung, W.C.; Schulter, S.; Yang, M.H.; Chandraker, M. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7472–7481.

10. Vu, T.H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2517–2526.
11. Wang, M.; Zhang, Z.; Liu, H. An Adaptive Semantic Segmentation Network for Adversarial Learning Domain Based on Low-Light Enhancement and Decoupled Generation. *Appl. Sci.* **2024**, *14*, 3295. [CrossRef]
12. Zou, Y.; Yu, Z.; Kumar, B.; Wang, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 289–305.
13. Morerio, P.; Volpi, R.; Ragonesi, R.; Murino, V. Generative pseudo-label refinement for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3130–3139.
14. Zhang, L.; Lan, M.; Zhang, J.; Tao, D. Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
15. Herath, S.; Fernando, B.; Abbasnejad, E.; Khan, S.; Shah, M. Energy-based self-training and normalization for unsupervised domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 11653–11662.
16. Zhao, Q.; Lyu, S.; Zhao, H.; Li, Y.; Zhang, L.; Du, Q. Self-training guided disentangled adaptation for cross-domain remote sensing image semantic segmentation. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *127*, 103646. [CrossRef]
17. Wang, Z.; Zhang, Y.; Zhang, Z.; Li, H.; Liu, X.; Chen, J. SDAT-Former++: A foggy scene semantic segmentation method with stronger domain adaption teacher for remote sensing images. *Remote Sens.* **2023**, *15*, 5704. [CrossRef]
18. Dan, J.; Jin, T.; Chi, H.; Wang, Y.; Liu, X.; Zhao, Y. Trust-aware conditional adversarial domain adaptation with feature norm alignment. *Neural Netw.* **2023**, *168*, 518–530. [CrossRef]
19. Yang, M.; Yang, R.; Tao, S.; Liu, Y.; Zhang, H.; Li, X. Unsupervised domain adaptive building semantic segmentation network by edge-enhanced contrastive learning. *Neural Netw.* **2024**, *179*, 106581. [CrossRef] [PubMed]
20. Chen, Z.; Zhang, Y.; Qi, X.; Wang, H.; Li, J.; Zhang, L. HeightFormer: A multilevel interaction and image-adaptive classification–regression network for monocular height estimation with aerial images. *Remote Sens.* **2024**, *16*, 295. [CrossRef]
21. Vu, T.H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. Dada: Depth-aware domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7364–7373.
22. Saha, S.; Obukhov, A.; Paudel, D.P.; Van Gool, L.; Gehler, P.V. Learning to relate depth and semantics for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021; pp. 8197–8207.
23. Zhou, S.; Feng, Y.; Li, S.; Wang, X.; Du, Q. DSM-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
24. Xiong, Z.; Chen, S.; Wang, Y.; Mou, L.; Zhu, X.X. GAMUS: A geometry-aware multi-modal semantic segmentation benchmark for remote sensing data. *arXiv* **2023**, arXiv:2305.14914.
25. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
26. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv* **2014**, arXiv:1412.7062.
27. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
28. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
29. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
31. Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; Re, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16344–16359.
32. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
33. Chen, J.; Lu, Y.; Yu, Q.; Chen, X.; Xing, E. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [CrossRef]

34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
35. Zhu, E.; Chen, Z.; Wang, D.; Li, Y.; Zhang, H.; Sun, X. UNetMamba: An Efficient UNet-Like Mamba for Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2025**, *22*, 1–5. [CrossRef]
36. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
37. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
38. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
39. Jiang, W.; Sun, Y.; Lei, L.; Kuang, G.; Ji, K. AdaptVFMs-RSCD: Advancing Remote Sensing Change Detection from binary to semantic with SAM and CLIP. *ISPRS J. Photogramm. Remote Sens.* **2025**, *230*, 304–317. [CrossRef]
40. Tasar, O.; Happy, S.; Tarabalka, Y.; Alliez, P. ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7178–7193. [CrossRef]
41. Chung, I.; Yoo, J.; Kwak, N. Exploiting inter-pixel correlations in unsupervised domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 12–21.
42. Wang, L.; Xiao, P.; Zhang, X.; Chen, X. A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2023**, *16*, *10156. [CrossRef]*
43. Chen, J.; He, P.; Zhu, J.; Guo, Y.; Sun, G.; Deng, M.; Li, H. Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
44. Chen, F.; Tan, X.; Liu, H. MSDAN: Multi-Scale Domain Adaptive Network for Cross-Domain Remote Sensing Images Adaptive Segmentation. In Proceedings of the 2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Chengdu, China, 18–20 April 2025; IEEE: Piscataway, NJ, USA; pp. 976–981.
45. Zhi, Y.; Wang, Y.; Zhang, F.; Ma, M.; Mei, S. MSFFNet: Multimodal Spatial–Frequency Fusion Network for RGB-DSM Remote Sensing Image Segmentation. *Remote Sens.* **2025**, *17*, 3745. [CrossRef]
46. Cao, Z.; Fu, K.; Lu, X.; Diao, W.; Sun, H.; Yan, M.; Yu, H.; Sun, X. End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1766–1770. [CrossRef]
47. Chen, H.; Zhang, H.; Yang, G.; Li, Y.; Wang, X.; Du, Q. A mutual information domain adaptation network for remotely sensed semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
48. ISPRS. ISPRS Test Project on Urban Classification and 3D Building Reconstruction. 2013. Available online: https://www.cvlibs.net/projects/autonomous_vision_survey/literature/Rottensteiner2013.pdf (accessed on 1 January 2026).
49. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
51. Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; Yang, Y. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2507–2516.
52. Yu, F.; Zhang, M.; Dong, H.; Wang, C.; Feng, J. Dast: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 10754–10762.
53. Hoyer, L.; Dai, D.; Van Gool, L. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 9924–9935.
54. Ma, X.; Zhang, X.; Wang, Z.; Pun, M.O. Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15. [CrossRef]
55. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014; Volume 27.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.