remote sensing

MDPI

Article

WCMNet: A Wavelet-Guided and CNN–Mamba Hybrid Network Approach for Unsupervised Domain Adaptation in Building Extraction

Dongjie Yang [1,2], Kuikui Han [1], Yuanwei Yang [1,3,*], Xianjun Gao [1], Kangliang Guo [1], Xinlong Gao [1] and Ruijing Huang [1]

[1] School of Geosciences, Yangtze University, Wuhan 430100, China; 2023730054@yangtzeu.edu.cn (D.Y.); 2022720541@yangtzeu.edu.cn (K.H.); junxgao@yangtzeu.edu.cn (X.G.); guokangliang@yangtzeu.edu.cn (K.G.); 2024720532@yangtzeu.edu.cn (X.G.); 2025720542@yangtzeu.edu.cn (R.H.)
[2] School of Information and Design, Zhejiang Industry Polytechnic College, Shaoxing 312000, China
[3] State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan 430063, China
* Correspondence: 516042@yangtzeu.edu.cn; Tel.: +86-156-2354-2326

Highlights

What are the main findings?

- A wavelet-guided and CNN–Mamba hybrid framework, WCMNet, is proposed for unsupervised domain adaptation in remote sensing building extraction.
- WCMNet achieves superior cross-domain building extraction performance, improving building completeness, boundary delineation, and robustness across the WHU, Massachusetts, and Potsdam datasets.

What are the implications of the main findings?

- Selective low-frequency style alignment, preserving high-frequency details, is effective for reducing structural degradation during cross-domain image translation.
- Joint modeling of local textures and global context provides a practical solution for cross-resolution, cross-scene, and cross-sensor building extraction in remote sensing.

Abstract

With the increasing diversity of remote sensing image acquisition conditions and imaging scenarios, building extraction models often experience significant performance degradation in cross-dataset applications due to variations in sensors and scene characteristics. Improving their cross-domain generalization ability has therefore become a critical research problem. To address the challenges of appearance style discrepancy and feature distribution shift in cross-domain building extraction, this paper proposes WCMNet, a wavelet-guided and CNN–Mamba hybrid network for unsupervised domain adaptation in building extraction. Specifically, a Mamba Wavelet Alignment (MWA) module is designed to align low-frequency style information in the wavelet domain while preserving directional high-frequency edge structures, thereby mitigating cross-domain appearance discrepancies and reducing structural degradation during domain translation. In addition, a Global–Local Mamba Block (GLMB) is developed to jointly model local textures and global semantic dependencies. In GLMB, the CNN branch captures fine-grained local details and boundary cues, while the Mamba branch models long-range contextual information; an adaptive gated fusion mechanism further integrates the two types of features. Experimental results on six cross-domain transfer tasks across the WHU, Massachusetts,

Check for updates

Academic Editor: Massimiliano Pepe

Received: 1 April 2026
Revised: 30 April 2026
Accepted: 6 May 2026
Published: 7 July 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.

and Potsdam datasets demonstrate that WCMNet consistently outperforms existing state-of-the-art domain adaptation methods. In particular, WCMNet achieves an average IoU of 65.13% and an average BIoU of 74.80% across all transfer settings, with improvements of up to 27.35 percentage points in IoU and 38.32 percentage points in BIoU compared with the strongest competing methods. These results demonstrate that the proposed MWA and GLMB effectively improve building completeness, boundary delineation accuracy, and cross-domain robustness.

Keywords: building extraction; domain adaptation; wavelet transform; CNN; Mamba

1. Introduction

With the widespread application of multisource sensors, remote sensing data have become increasingly diverse in terms of visual style [1,2]. While this diversity provides rich information for building extraction and subsequent semantic analysis, it also introduces greater challenges [3,4]. Traditional methods rely on handcrafted features such as shape and texture, which exhibit limited generalization in complex scenes [5]. Although supervised approaches based on convolutional neural networks (CNNs) and Vision Transformers (ViTs) have achieved remarkable progress [6,7], their performance remains constrained by domain shift. Due to significant differences in spectral characteristics and illumination conditions across images acquired by different sensors, models trained on labeled source domains often fail to generalize to unlabeled target domains effectively, thereby limiting their performance in large-scale real-world applications [8,9].

To mitigate discrepancies in feature distributions between the source and target domains and enable effective knowledge transfer in the absence of target-domain annotations, unsupervised domain adaptation (UDA) has emerged as a promising solution with broad applicability. Existing UDA methods for semantic segmentation can be broadly categorized into two types: adversarial learning-based distribution alignment methods and self-training-based pseudo-label learning methods [10,11]. Among them, image-level style transfer strategies based on Generative Adversarial Networks (GANs) have been widely adopted to reduce appearance discrepancies between the source and target domains [12]. By transforming source-domain images into representations closer to the target-domain style, these methods can improve cross-domain generalization to some extent. However, despite their widespread use in domain adaptation, GAN-based methods still exhibit notable limitations in preserving structural consistency and spatial detail integrity. Specifically, during style transfer, these methods often introduce artifacts or degrade high-frequency structural information, leading to blurred building boundaries, distorted contours, and loss of local texture details [13].

In addition, most existing UDA frameworks primarily focus on reducing cross-domain appearance discrepancies while paying insufficient attention to another critical issue: the lack of effective global context modeling. From a feature-representation perspective, mainstream convolutional neural networks (CNNs) are inherently constrained by their local receptive fields, making it difficult to capture the long-range dependencies required in complex scenes and thereby limiting their ability to suppress false-positive predictions. Although Transformer-based methods introduce global receptive fields via self-attention and partially alleviate this limitation, their computational complexity typically scales quadratically with input size, greatly restricting their applicability in high-resolution remote sensing scenarios [14]. Recently, representative state space model (SSM)-based methods, such as Mamba, have demonstrated strong potential for efficient global modeling with linear com-

plexity, providing a promising alternative for high-resolution vision tasks [15–17]. However, this potential has not yet been fully exploited in the remote sensing domain adaptation. A key reason is that cross-domain remote sensing building extraction is fundamentally more challenging than general in-domain vision tasks: in addition to requiring global semantic perception, it must also cope with cross-sensor and cross-resolution appearance variations, complex scene layouts, and severe feature distribution shifts between the labeled source domain and the unlabeled target domain. Moreover, buildings in remote sensing imagery often exhibit dense distributions, slender boundaries, complex geometric structures, and large intra-class appearance variations, which require the model to preserve fine local details while maintaining global structural consistency. Therefore, directly introducing a generic Mamba-based architecture is insufficient for the remote sensing domain adaptation. Instead, a task-oriented design is needed to couple efficient global modeling with local structure preservation and cross-domain alignment.

To address the above challenges, we propose an end-to-end framework, WCMNet (Wavelet-Guided and CNN–Mamba Hybrid Network), for unsupervised domain adaptation in building extraction. Rather than directly applying Mamba as a generic backbone, WCMNet is specifically designed for cross-domain remote sensing building extraction by jointly considering appearance alignment and structural representation. Specifically, we introduce a wavelet-based image-level style-transfer strategy that decouples style from structure. By aligning only the low-frequency components in the wavelet domain while explicitly preserving high-frequency subbands, the proposed method effectively mitigates cross-domain appearance discrepancies while reducing structural degradation during domain translation. Furthermore, to address feature distribution shift and improve structural representation under complex scene conditions, we design a hybrid CNN–Mamba architecture equipped with an adaptive gated fusion mechanism. In this architecture, CNN is used to extract local texture details and boundary cues, while Mamba is employed to model long-range semantic dependencies and global contextual information. By dynamically fusing the two branches, the proposed framework achieves a better balance between local detail preservation and global structural consistency, thereby improving cross-domain building extraction performance.

Compared with existing frequency-domain alignment methods such as FDANet, which mainly rely on Fourier-domain low-frequency alignment to mitigate appearance discrepancy, the proposed method performs style transfer in the wavelet domain with explicit subband differentiation. Specifically, only the low-frequency component is aligned for global style adaptation. At the same time, the directional high-frequency subbands are explicitly preserved and enhanced, which is more suitable for maintaining boundary continuity and structural integrity in cross-domain remote sensing building extraction. Moreover, unlike generic CNN/SSM hybrid architectures that mainly focus on backbone-level feature fusion, our framework is designed as a task-oriented coupling of image-level frequency-domain alignment and global–local structural representation. The wavelet alignment module first generates direction-aware, structure-preserved representations, and the subsequent CNN–Cross-Scan Mamba hybrid modeling further integrates local boundary details with long-range semantic dependencies. In this way, WCMNet forms a unified framework that jointly addresses appearance discrepancy, feature distribution shift, and structural preservation, rather than simply combining CNN and Mamba in parallel. The main contributions of this paper are summarized as follows:

(1) We propose WCMNet, a novel UDA framework for cross-domain building extraction. It is designed to jointly address two key problems: image-level appearance discrepancies caused by cross-sensor and cross-resolution variations, and feature-level distribution shifts that weaken target-domain generalization. By integrating

wavelet-based alignment with a CNN–Mamba hybrid model, WCMNet provides a unified solution to improve building completeness, boundary delineation, and cross-domain robustness.

(2) We design a Mamba Wavelet Alignment (MWA) module to address structural degradation during domain translation. Unlike conventional image-level alignment methods that may blur building contours or distort local textures, MWA selectively aligns low-frequency style statistics while explicitly preserving directional high-frequency edge information. This design reduces cross-domain appearance discrepancy while maintaining boundary-related structural cues, which is particularly important for dense buildings, slender boundaries, and complex roof structures.

(3) We introduce a Global–Local Mamba Block (GLMB) to address the insufficient global context modeling and local detail preservation of existing UDA backbones. In GLMB, Cross-Scan Mamba captures long-range semantic dependencies, while the CNN branch extracts fine-grained textures and boundary details. An adaptive gated fusion mechanism is further used to dynamically balance global semantics and local structural cues, thereby improving structural consistency and reducing omission and false-positive errors in complex cross-domain scenarios.

2. Related Work

2.1. Unsupervised Domain Adaptation in Remote Sensing

Unsupervised domain adaptation (UDA) aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain. In remote sensing building extraction, existing UDA methods can generally be categorized into alignment-based and self-training-based approaches [18]. Alignment-based methods typically rely on adversarial learning to reduce domain discrepancies at the image level or feature level. For instance, CyCADA [12] and AdaptSegNet [10] employ discriminators to align source and target domain distributions. To better accommodate the unique characteristics of remote sensing imagery, some studies further incorporate scale-aware constraints to address variations in spatial resolution [19]. Nevertheless, image-level alignment via CycleGAN often introduces semantic distortions and hallucination artifacts, resulting in blurred building boundaries and degraded structural fidelity. By contrast, self-training-based methods [20] iteratively refine the model by generating pseudo-labels for target-domain samples. Although such methods have shown promising performance, they remain vulnerable to the influence of noisy pseudo-labels. This issue becomes particularly severe when the target-domain distribution differs substantially from that of the source domain, as inaccurate pseudo-labels can provide misleading supervision signals, thereby impairing adaptation.

2.2. State Space Models and Mamba

Although Transformers [7] have become the dominant paradigm for global context modeling, the computational cost of self-attention scales quadratically with image size, limiting their efficiency in high-resolution visual tasks. Recently, state space models (SSMs), especially the Structured State Space Sequence Model (S4) and its selective variant Mamba [15], have emerged as promising alternatives due to their ability to model long-range dependencies with linear complexity. In the vision domain, representative methods such as Vision Mamba [21] and VMamba [22] introduce two-dimensional scanning strategies to adapt one-dimensional SSMs to image understanding tasks, demonstrating strong performance in capturing long-range dependencies and, in some cases, surpassing CNNs and Transformers. Despite their impressive success in natural image classification and segmentation, the potential of Mamba for cross-domain remote sensing tasks has not yet been systematically explored. This is particularly important for building extraction, where

robust global context modeling is crucial for distinguishing buildings from complex and confusing backgrounds. In this work, WCMNet effectively incorporates Mamba into a UDA framework, bridging this research gap.

2.3. Wavelet Transform for Style Transfer

The wavelet transform (WT) provides an effective means of decomposing an image into frequency subbands, thereby separating low-frequency content from high-frequency details. In style transfer, prior studies have shown that style-related information is primarily concentrated in low-frequency components, whereas structural information is typically preserved in high-frequency components [23,24]. For example, PhotoWCT [25] employs wavelet pooling to mitigate the loss of spatial information during stylization. In the context of domain adaptation, recent studies have also explored frequency-domain alignment strategies to enforce style consistency across domains [26]. Unlike Fourier-domain alignment methods such as FDANet, which mainly mitigate appearance discrepancy by aligning low-frequency spectral information, our method performs style transfer in the wavelet domain with explicit subband differentiation. Specifically, only the low-frequency wavelet component is aligned for global style adaptation, while the directional high-frequency subbands are preserved to maintain boundary continuity and fine-grained structural details. This design is better suited to cross-domain remote sensing building extraction, where slender boundaries and dense building distributions require greater structural fidelity during style transfer.

3. Methodology

3.1. Overall Overview of the Method

We propose a WCMNet approach for unsupervised domain adaptation in building extraction, an end-to-end unsupervised domain adaptation network for building extraction in remote sensing imagery. The proposed method is designed to alleviate both appearance discrepancy and feature distribution shift caused by variations in imaging conditions and sensor characteristics across different remote sensing datasets. By reducing domain shift, WCMNet enables a model trained only on labeled source-domain data to generalize robustly to unlabeled target domains. As shown in Figure 1, the overall training and adaptation process consists of two stages. The first stage performs unsupervised domain-adaptation pretraining using WCMNet. The second stage further improves cross-domain generalization by generating pseudo-labels for the target domain and continuing training using the model from the first stage.

In the first stage, three complementary alignment mechanisms are introduced during domain adaptation: (1) resolution alignment, (2) wavelet-transform-based image-level style transfer, and (3) CNN–Mamba-based global–local dynamic feature extraction with pixel-level alignment.

(1) Resolution Alignment: Due to differences in spatial resolution across datasets, image patches of the same size may correspond to different physical coverage areas. To ensure that source-domain and target-domain input patches are comparable in terms of physical scale, we adopt an adaptive cropping strategy based on the resolution ratio. Specifically, for high-resolution data with a spatial resolution of 0.3 m, a 384 × 384 region is first cropped and then resized to 128 × 128, while for lower-resolution data with a spatial resolution of 1.0 m, a 128 × 128 patch is directly cropped. This strategy normalizes the effective spatial scale of the two domains: the resized high-resolution patch corresponds to an effective ground sampling distance of approximately 0.9 m, which is close to that of the 1.0 m data. Although the resizing operation inevitably removes some very fine spatial details, these details are not consistently

observable at lower resolution and may introduce a source-domain-specific texture bias. Therefore, this resolution alignment strategy is adopted to reduce scale-induced domain discrepancy and encourage the model to learn more transferable building representations across domains.

(2) Image-Level Alignment: To reduce cross-domain appearance discrepancy while preserving building structural boundaries as much as possible, we employ a wavelet-transform-based image-level style alignment strategy. The core idea is to decompose an image into frequency subbands and process the different frequency components separately. Unlike conventional image-level style alignment methods, which may damage edge information or cause excessive smoothing, the proposed method preserves the directional high-frequency components (LH/HL/HH) while aligning style information. In this way, critical edge structures can be stably retained during domain translation. More importantly, this coupling is not a simple serial combination between style transfer and sequence modeling. Instead, the wavelet alignment module is deliberately designed to produce direction-aware, structure-preserved representations that are better aligned with the subsequent Cross-Scan Mamba. Specifically, by suppressing low-frequency appearance discrepancy while retaining directional high-frequency boundary cues, the transformed features provide more reliable structural signals for Mamba to model long-range spatial dependencies along different scanning paths. As a result, the proposed frequency-domain alignment strategy is intrinsically coupled with the CNN–Mamba hybrid architecture, jointly suppressing appearance discrepancy while enhancing structural consistency.

(3) Feature-Level Alignment: At the feature level, we adopt a classical encoder–decoder segmentation framework and introduce a Cross-Scan Mamba Block during the encoding stage for global–local feature extraction. Specifically, the two-dimensional feature maps are unfolded into one-dimensional sequences along the horizontal and vertical directions. Then, the selective state-space modeling capability of the Mamba is leveraged to capture global context and long-range spatial dependencies, thereby enhancing scene-level semantic representation. Meanwhile, the CNN branch is responsible for extracting local texture patterns and edge details. Subsequently, a gated fusion mechanism is employed to adaptively integrate these two types of features, allowing the network to dynamically balance global semantics and local details according to spatial locations. During the decoding stage, progressive upsampling and skip connections are used to fuse multi-scale features, and the final pixel-wise building segmentation map is generated. Compared with pure CNN- or Transformer-based architectures, the proposed strategy achieves better synergy between fine-grained boundary delineation and global spatial layout understanding while maintaining near-linear computational complexity.

In the second stage, a progressive semi-supervised training strategy based on pseudo-label generation is adopted to further improve the generalization ability of the model on the target domain. Specifically, the pretrained model obtained in Stage I is first used to infer target-domain samples. High-confidence regions are then selected as pseudo-labels according to pixel-wise prediction confidence, while low-confidence pixels are ignored to reduce the influence of noisy supervision. The model is subsequently trained jointly on the labeled source-domain data and the pseudo-labeled target-domain data to complete one round of adaptation. In subsequent iterations, the confidence threshold is gradually reduced, allowing the model to progressively transition from high-confidence "easy samples" to lower-confidence "hard samples" and ambiguous boundary regions. This strategy effectively suppresses the accumulation and amplification of pseudo-label noise. At the same time, it progressively enhances the ability of the model to adapt to complex

distributions and uncertain boundaries, thereby improving generalization performance on the target domain.

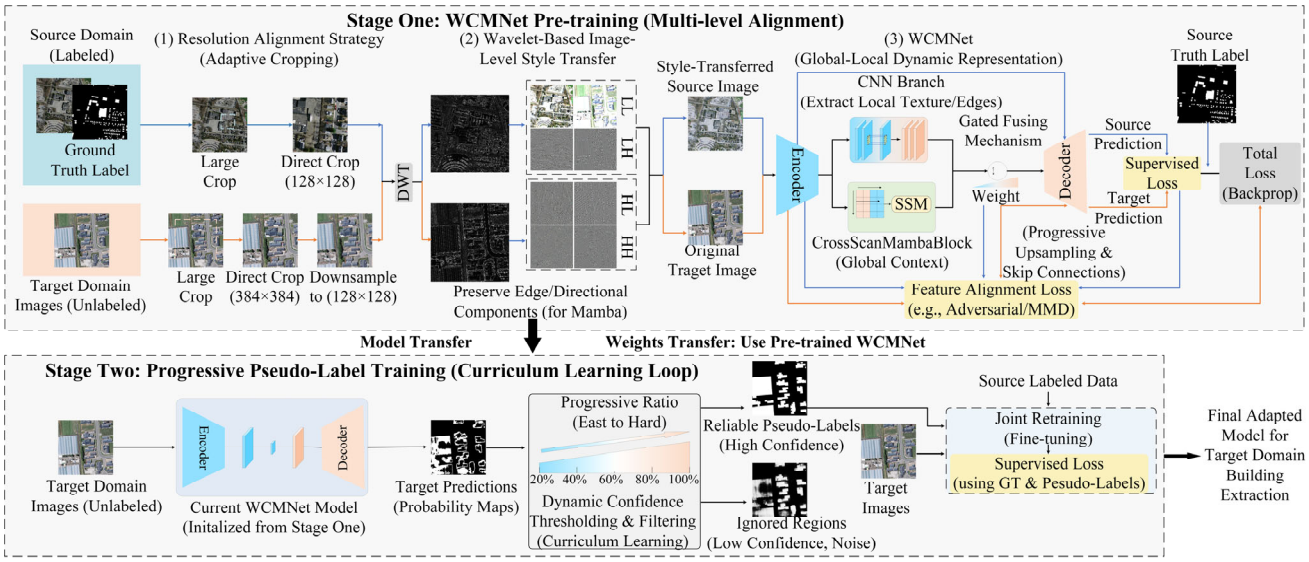


Figure 1. Overall pipeline of WCMNet, an unsupervised domain adaptation framework for building extraction based on wavelet guidance and CNN–Mamba hybrid modeling.

3.2. Mamba Wavelet Alignment-Based Style Transfer Method

As shown in Figure 2, Mamba Wavelet Alignment (MWA) is a domain-level style transfer method specifically tailored for the Cross-Scan Mamba architecture. By integrating frequency-domain decomposition with spatial adaptivity, it is designed to address the challenges posed by significant appearance style discrepancy and the tendency for structural information to degrade in cross-domain remote sensing tasks. Specifically, the proposed method first applies the Haar discrete wavelet transform (DWT) to decompose the source-domain image into a low-frequency approximation subband LL, which captures the overall style tone, and high-frequency detail subbands $\{LH,\ HL,\ HH\}$, which encode directional texture information. Considering that the Mamba architecture relies on horizontal and vertical scanning mechanisms to capture long-range dependencies, the proposed method does not simply replace low-frequency information as in conventional wavelet-based approaches. Instead, it adopts a differentiated frequency-band processing strategy. In particular, Soft Adaptive Instance Normalization (Soft AdaIN) is applied to the component LL based on the target-domain statistical feature $(\mu_t,\ \sigma_t)$. The formulation is given as follows:

$$LL' = \sigma_t\left(\frac{LL - \mu_s}{\sigma_s}\right) + \mu_t \tag{1}$$

where LL and LL' denote the original and aligned low-frequency subbands, respectively; μ_s and σ_s represent the mean and standard deviation of the source-domain low-frequency component; and $(\mu_t,\ \sigma_t)$ denotes the statistical feature used for style alignment. In practice, instead of employing a shared adaptive blending coefficient across all wavelet subbands, we adopt a subband-specific fixed processing policy. Specifically, a single-level Haar discrete wavelet transform is used in all experiments. Full style alignment is applied only to the low-frequency subband LL, i.e., $\alpha_{LL} = 1.0$, so as to eliminate global domain-level discrepancies in illumination and tone. In contrast, Soft AdaIN is not applied to the directional high-frequency subbands LH and HL; these components are directly preserved to retain horizontal and vertical boundary structures that are highly compatible with the Cross-Scan Mamba mechanism. For the diagonal high-frequency subband HH, the original

component is kept unchanged, which can be equivalently interpreted as $\alpha_{HH} = 0.0$, thereby preventing the amplification of unstable diagonal details or the introduction of artifacts during domain translation. Moreover, the statistical feature (μ_t, σ_t) used in Equation (1) is estimated from the target-domain training images and kept fixed throughout training. When edge enhancement is enabled, the preserved LH and HL components are further multiplied by an enhancement factor γ to strengthen orientation-sensitive edge responses along the Mamba scanning paths. In this way, the proposed subband-wise design explicitly coordinates low-frequency style adaptation, high-frequency structural preservation, and subsequent Mamba-based long-range dependency modeling, as formally summarized in Algorithm 1.

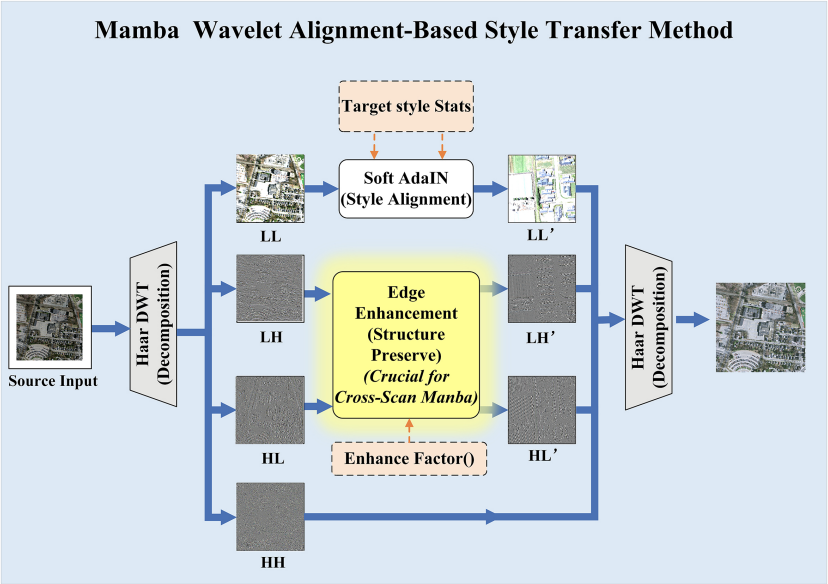


Figure 2. Workflow of the proposed MWA style transfer method with a subband-specific fixed policy: full alignment on LL, preservation on LH, HL, and unchanged HH.

Algorithm 1 Mamba Wavelet Alignment

Input: source image x_s, source-domain statistics (μ_s, σ_s), wavelet level $L = 1$, fixed subband policy $\{\alpha_{LL} = 1.0, \alpha_{HH} = 0.0\}$, optional edge enhancement factor γ, post-reconstruction global correction, adaptive texture-guided fusion
Output: aligned image x_{out}
$\{LL, LH, HL, HH\} \leftarrow DWT(x_s)$ ▷ Single-level Haar wavelet decomposition
$LL' \leftarrow \text{AdaIN}(LL; \mu_t, \sigma_t)$ ▷ Full style alignment on low-frequency subband
$LH' \leftarrow LH, HL' \leftarrow HL$ ▷ Preserve directional high-frequency structure
if edge enhancement is enabled then
$LH' \leftarrow \gamma.LH', HL' \leftarrow \gamma.HL'$ ▷ Optional enhancement for directional edge cues
end if
$HH' \leftarrow HH$ ▷ Keep diagonal high-frequency component unchanged
$Iwave \leftarrow IDWT(LL', LH', HL', HH')$ ▷ Inverse wavelet reconstruction
Iwave ← GlobalStatAlign(Iwave; μ_t, σ_t) ▷ Post-reconstruction global statistical correction
Mtex ← TextureMap(x_s; variance window = 15 × 15, Gaussian smoothing = 21 × 21)
Iglobal ← GlobalStatAlign(x_s; μ_t, σ_t)
$x_{out} \leftarrow$ Mtex ⊙ Iwave + (1 − Mtex) ⊙ Iglobal
return x_{out}

Therefore, the proposed MWA differs fundamentally from Fourier-based alignment methods. Rather than aligning frequency components globally, it adopts a structure-aware subband policy that restricts style adaptation to the low-frequency component. At the same time, directional high-frequency cues are explicitly preserved and enhanced. This enables the transferred representation to better maintain building boundaries and orientation-sensitive structures, which are critical for subsequent Cross-Scan Mamba modeling.

In addition, unlike the conventional Wallis filter, which tends to disrupt global color consistency or cause the loss of local texture details, MWA introduces a post-reconstruction global statistical correction and an adaptive spatial fusion mechanism based on a texture map. Specifically, a texture mask M_{tex} is generated by computing the local variance, and the wavelet-reconstructed result I_{wave} and the globally aligned result I_{global} are dynamically fused as follows:

$$I_{out} = M_{tex} \odot I_{wave} + (1 - M_{tex}) \odot I_{global} \quad (2)$$

This design preserves global style consistency in homogeneous regions by leveraging global statistics, while prioritizing the retention of wavelet-preserved structural details in texture-rich regions. As a result, it achieves an effective balance between style transfer and structure preservation, thereby improving the generalization ability of the model in cross-domain building extraction. For reproducibility, the post-reconstruction global statistical correction is always enabled in our implementation. Specifically, after inverse wavelet reconstruction, the reconstructed image I_{wave} is further aligned to the source-domain global statistics through GlobalStatAlign (I_{wave}; μ_t, σ_t). Meanwhile, a texture-guided adaptive fusion strategy is employed by computing the texture map M_{tex} from the local variance of the original image. In our implementation, the local variance is calculated using a 15×15 window, and a 21×21 Gaussian filter further smoothes the resulting texture map. This design allows texture-rich regions to retain more wavelet-preserved structural details, while homogeneous regions rely more on global statistical correction, thereby improving the stability and transparency of the style transfer process.

3.3. Global–Local Dynamic Feature Learning via CNN–Mamba Hybrid Modeling

To overcome the limitations of traditional CNNs, which are constrained by their local receptive field and therefore struggle to capture long-range dependencies, as well as the limitation of pure Transformer architectures, whose computational complexity increases quadratically with image resolution, we introduce a Global–Local Mamba Block (GLMB) in the encoder stage. As shown in Figure 3, this module adopts a dual-branch parallel design, with the CNN branch extracting high-frequency local texture features and the Mamba branch capturing low-frequency global semantic features. Notably, the input to this hybrid modeling stage has already been conditioned by the preceding MWA module, which suppresses cross-domain appearance bias while preserving direction-aware structural cues, thereby making the subsequent Cross-Scan Mamba more effective for long-range building-structure modeling. Therefore, the proposed GLMB is not a conventional CNN/SSM parallel backbone. Instead, it is a task-oriented hybrid representation module in which the Mamba branch operates on direction-aware, structure-conditioned features generated by the preceding wavelet alignment module. In contrast, the CNN branch complements it with local texture and boundary cues. The complete procedure of this dual-branch global–local feature extraction and fusion is formally summarized in Algorithm 2.

Algorithm 2 Dual-Branch Global–Local Feature Extraction and Fusion

Input: Input feature tensor $x \in \mathrm{R}^{BCHW}$
Parameters: θ_{cnn}(Local), θ_{ssm}(Mamba), θ_{gate}(Fusion)
Output: Fused feature representation Y
//Captures high-frequency details (e.g., edges, corners)
//Captures high-frequency details (e.g., edges, corners)
$\mathrm{F}_{local} \leftarrow Conv2d(x;\ \theta_{cnn})$
$\mathrm{F}_{local} \leftarrow \mathrm{ReLU}(BatchNorm(\mathrm{F}_{local}))$
//Branch2: Global Context Modeling (Cross-Scan Mamba)
//Captures long-rang semantic dependencies
$x_{seg} \leftarrow$ Flatten (x) ▷ Reshape to $B \times L \times C$, where $L = H \times W$
$D \leftarrow$ {Top-Left, Top-Right, Bottom-left, Bottom-Right}
$\mathrm{H}_{list} \leftarrow \phi$
for each direction $d \in D$ do
    $x_d \leftarrow$ Scan $(x_{seg},\ dir = d)$ ▷ Reorder spatial sequence
    $H_d \leftarrow$ SSM $(x_d;\ \theta_{ssm})$ ▷ Selective State Space Model
    $H_d^{rev} \leftarrow$ ReverseScan $(\mathrm{H}_d,\ \mathrm{dir} = d)$ ▷ Restore spatial alignment
    H_{list}. append (H_d^{rev})
end for
$\mathrm{F}_{global} \leftarrow \sum H_{list}$ ▷ Aggregate multi-directional features
$\mathrm{F}_{global} \leftarrow$ Reshape $(\mathrm{F}_{global}) \rightarrow \mathrm{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$
//Stage 3: Adaptive Gated Fusing Mechanism
//Dynamically balance local details and global semantics
$\mathrm{F}_{cat} \leftarrow$ Concat $(\left[\mathrm{F}_{local},\ \mathrm{F}_{global}\right],\ \mathrm{dim} = C)$
$\mathrm{M}_{gate} \leftarrow \sigma(\mathrm{Conv}_{1\times1}(\mathrm{F}_{cat};\ \theta_{gate}))$ ▷ Spatial Gating Map $\mathrm{G} \in (0,\ 1)$
$\mathrm{Y} \leftarrow \mathrm{M}_{\mathrm{gate}} \odot \mathrm{F}_{\mathrm{global}} + (1 - \mathrm{M}_{gate}) \odot \mathrm{F}_{local}$ ▷ Complementary Fusion
return Y

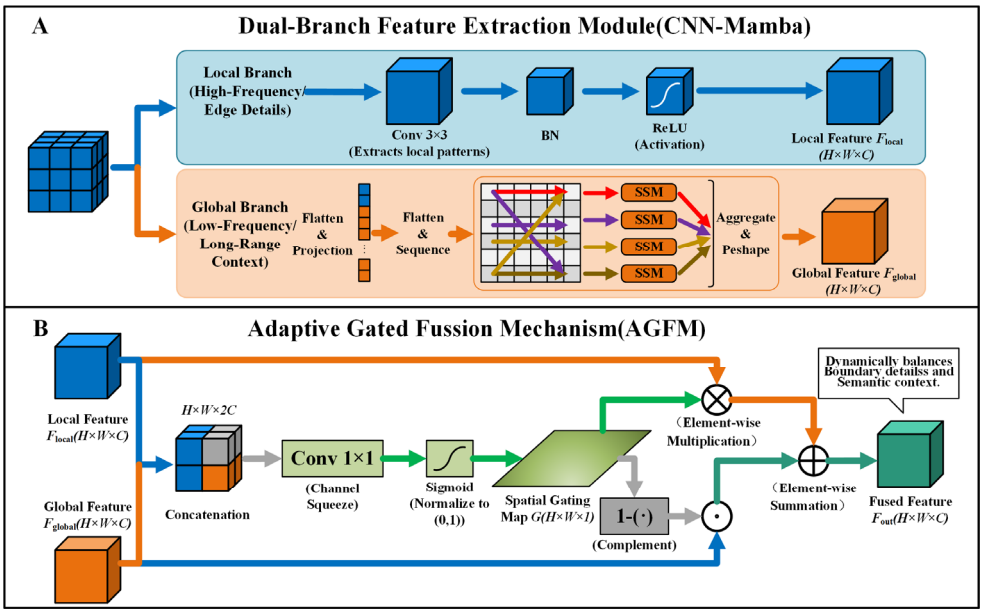


Figure 3. Overall pipeline of the proposed CNN–Mamba-based global–local dynamic feature learning method.

Given an input feature $x \in R^{H\times W\times C}$, we feed it into the local convolutional branch and the global state-space branch, respectively. Local features are particularly important

for building extraction, since the fine-grained boundaries and corner structures of buildings are often embedded in high-frequency information. Therefore, we adopt a residual convolutional block to extract local representations, which can be expressed as F_{local}:

$$F_{local} = F_{CNN}(x) = \mathrm{RELU}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(x))) \tag{3}$$

where $\mathrm{Conv}_{3\times3}$ effectively captures spatial correlations within local neighborhoods, thereby preserving the geometric boundary structures of buildings.

To model global contextual dependencies while retaining linear computational complexity, we further introduce a Cross-Scan Mamba Block based on State Space Models (SSMs). Conventional SSMs are inherently designed for one-dimensional sequential modeling, which makes them not directly applicable to two-dimensional image representations. To address this limitation, the feature map x is first unfolded into directional sequences $x \in R^{L\times C}$ along four scanning directions, namely upper-left, lower-right, upper-right, and lower-left, where $L = H \times W$. For each directional sequence, we adopt the Selective State Space Model (Selective SSM), a variant of the Structured State Space Sequence Model (S4), to perform sequence modeling. The continuous-time state equation is formulated as follows:

$$h'(t) = \mathrm{A}h(t) + \mathrm{B}x(t), \qquad y(t) = \mathrm{C}h(t) \tag{4}$$

where $h'(t) \in R^N$ denotes the latent state, $x(t)$ is the input signal, and $y(t)$ is the output response. In Mamba, the state-space matrices A, B, and C are dynamically predicted from the input, enabling the model to adaptively capture content-dependent dependencies. To facilitate computation in a deep learning framework, the continuous-time state-space equations are discretized using the time-scale parameter Δ with the Zero-Order Hold (ZOH) scheme. The resulting discrete recurrence is implemented through the selective scan operation in Mamba, enabling efficient sequence modeling along each scanning direction while maintaining linear computational complexity with respect to the sequence length.

Since the CNN branch primarily captures fine-grained texture details, while the Mamba branch models global semantic context, directly combining the two branches via simple summation may introduce feature conflicts or lead to mutual suppression. To overcome this limitation, we propose an Adaptive Gated Fusion Module (AGFM) that dynamically balances the contributions of the two branches by learning spatially adaptive attention weights.

Specifically, the local feature representation F_{local} and the global feature representation F_{global} are first concatenated along the channel dimension and then passed through a lightweight convolutional layer followed by a Sigmoid activation function to generate the spatial gating map $G \in \mathrm{R}^{H\times W\times C}$.

$$G = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{Concat}(F_{local},\ F_{global}))) \tag{5}$$

where $\sigma(\cdot)$ represents the Sigmoid activation function, which constrains the learned weights to the (0, 1) interval. Accordingly, the final fused feature F_{out} is computed through a gated weighting operation as follows:

$$F_{out} = G \odot F_{global} + (1 - G) \odot F_{local} \tag{6}$$

where $\odot$ denotes the Hadamard Product. The gating map G essentially acts as a dynamic information router. In regions with strong semantic consistency, such as the interior of buildings, the model tends to assign larger G values, thereby relying more on F_{global} to preserve the overall integrity of the prediction. In contrast, in detail-rich regions such as building boundaries, the model tends to assign smaller G values, thus prioritizing F_{local} to

refine boundary segmentation. This mechanism effectively alleviates the common issues of intra-class inconsistency and inter-class ambiguity in remote sensing imagery, enabling the network to maintain global awareness while preserving fine-grained details.

Based on the above network design, the overall loss functions used in Stage I and Stage II are further formulated as follows.

3.4. Overall Loss

To explicitly present the optimization process of the proposed framework, the overall loss functions used in the two training stages are summarized in this subsection. Let (x_s', y_s) denote a labeled source-domain image-label pair, and let x_t denote an unlabeled target-domain image. It should be noted that the image-level alignment in WCMNet is achieved by the proposed Mamba Wavelet Alignment (MWA) module through deterministic wavelet-domain style transformation before feature extraction, rather than by introducing an additional learnable loss term. Therefore, the loss function in Stage I mainly consists of source-domain supervised segmentation, feature-level alignment, pixel-level alignment, and adversarial adaptation. The overall optimization objective in Stage I is defined as follows.

$$L_{stage1} = L_{seg}^{s} + \lambda_{fea} L_{fea} + \lambda_{pix} L_{pix} + \lambda_{adv} L_{adv} \tag{7}$$

where L_{fea} L_{pix}, and L_{adv} denote the feature-level alignment loss, pixel-level alignment loss, and adversarial loss, respectively. λ_{fea}, λ_{pix} and λ_{adv} are their balancing coefficients. for the feature-level alignment loss, the pixel-level alignment loss, and the adversarial loss, respectively. In our implementation, these coefficients are set to $\lambda_{fea} = 1.0$, $\lambda_{pix} = 0.01$, and $\lambda_{adv} = 0.001$.

After Stage I pretraining, a progressive pseudo-label learning strategy is adopted in Stage II to further improve target-domain generalization. Specifically, the pretrained model is first used to generate predictions for target-domain samples, and only pixels with sufficiently high confidence are retained as pseudo-labels. The confidence mask at the k-th self-training round is defined as

$$M_i^{(k)} = 1(\max_c p_{i,c}^{(k)} \geq \tau_k) \tag{8}$$

where $p_{i,c}^{(k)}$ denotes the predicted probability of class c for the i-th target-domain pixel at the k-th round, $M_i^{(k)}$ is the corresponding binary confidence mask, and τ_k is the confidence threshold. Based on this mask, the pseudo-label loss is written as follows.

$$L_{pl}^{(k)} = \frac{\Sigma_i M_i^{(k)} \ell(\overset{\wedge}{p_i^t}, \widetilde{y_i^t})}{\Sigma_i M_i^{(k)}} \tag{9}$$

where $\overset{\wedge}{p_i^t}$ denotes the current prediction for the i-th target-domain pixel, $\overset{\wedge}{y_i^t}$ denotes the generated pseudo-label, and $\ell(\cdot, \cdot)$ denotes the same segmentation criterion used in supervised training. This strategy allows the model to learn first from high-confidence easy samples and then gradually adapt to harder samples and ambiguous boundary regions by progressively relaxing the confidence threshold. Accordingly, the overall loss function in Stage II is given as follows.

$$L_{stage2} = L_{\text{seg}}^{S} + \lambda_{pl} L_{pl} \tag{10}$$

where λ_{pl} is the weight of the pseudo-label loss. In our implementation, $\lambda_{pl} = 1.0$, and the initial confidence threshold is set to $\tau = 0.9$. When the mean-teacher strategy is enabled,

the teacher update rate is set to 0.999, and the consistency weight is set to 1.0. In this way, the proposed framework first learns stable cross-domain transferable representations in Stage I and then further improves adaptation performance in Stage II through conservative and progressively expanded pseudo-label supervision.

4. Experiment

4.1. Datasets and Experimental Settings

4.1.1. Dataset Description

To comprehensively evaluate the cross-domain transfer performance of the proposed WCMNet for building extraction, three widely used benchmark datasets, namely Potsdam [27], WHU [28], and Massachusetts [29], were selected for the experiments. These datasets differ substantially in spatial resolution, imaging conditions, scene appearance, building density, and structural complexity. Therefore, they provide a representative benchmark for evaluating the robustness of building extraction models under cross-resolution, cross-scene, and cross-appearance domain shifts. The Potsdam dataset is a very-high-resolution urban remote sensing dataset with an original spatial resolution of 0.05 m. It contains dense urban scenes with complex building layouts, diverse roof appearances, and rich structural details. To reduce the extreme resolution discrepancy with the other datasets and ensure a more reasonable cross-domain comparison, the Potsdam images were first downsampled to 0.3 m and then cropped into 512 $\times$ 512 image patches for training and testing. The WHU dataset has a spatial resolution of 0.3 m and contains large-scale aerial images with various building sizes, roof structures, and surrounding backgrounds. Compared with Potsdam, WHU exhibits differences in scene distribution and building morphology, while maintaining a similar spatial resolution after the downsampling of Potsdam. In this study, the WHU dataset was cropped into 512 $\times$ 512 image patches, resulting in a total of 8189 image samples, consistent with the commonly reported usage in the literature. The Massachusetts dataset has a spatial resolution of 1.0 m, which is lower than that of Potsdam and WHU. As a result, building boundaries and fine roof structures are less distinguishable, and the appearance differences between buildings and surrounding backgrounds become more ambiguous. This dataset is therefore suitable for evaluating the ability of domain adaptation methods to handle cross-resolution transfer and boundary degradation. The Massachusetts images were also cropped into 512 $\times$ 512 patches for all experiments. Based on these three datasets, six bidirectional transfer settings were constructed, including Potsdam $\rightarrow$ WHU (P-W), Potsdam $\rightarrow$ Massachusetts (P-M), WHU $\rightarrow$ Potsdam (W-P), WHU $\rightarrow$ Massachusetts (W-M), Massachusetts $\rightarrow$ WHU (M-W), and Massachusetts $\rightarrow$ Potsdam (M-P). These settings cover transfers between datasets with similar spatial resolution but different scene appearances, as well as transfers between high-resolution and low-resolution datasets. Therefore, they can more comprehensively evaluate whether the proposed method can improve building completeness, boundary preservation, and cross-domain generalization under different domain-shift conditions.

4.1.2. Experimental Setup

All experiments were conducted using the PyTorch 2.1.1 deep learning framework. The training and inference processes were accelerated on an NVIDIA RTX 3090 Ti GPU with 24 GB of memory. To ensure a fair comparison, the proposed method and all competing methods were trained under the same experimental environment. Specifically, the segmentation network was optimized using the Adam optimizer, with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The learning rate was decayed by a factor of 0.5 every 5 epochs. To comprehensively assess the cross-domain generalization capability of the proposed method under different scene conditions, experiments were performed on the

three datasets under six domain transfer settings, namely Potsdam → WHU (abbreviated as P-W), Potsdam → Massachusetts (abbreviated as P-M), WHU → Potsdam (abbreviated as W-P), WHU → Massachusetts (abbreviated as W-M), Massachusetts → WHU (abbreviated as M-W), and Massachusetts → Potsdam (abbreviated as M-P). In all experiments, the proposed MWA module uses a single-level Haar wavelet decomposition. The statistical feature $(\mu_s,\ \sigma_s)$ used for style alignment is estimated from the source-domain training images and fixed during training. Full alignment is applied to the low-frequency subband LL with $\alpha_{LL} = 1.0$, while the directional high-frequency subbands LH and HL are excluded from Soft AdaIN so as to preserve boundary-aware structural information. The diagonal high-frequency subband HH is kept unchanged, equivalently $\alpha_{HH} = 0.0$. In addition, post-reconstruction global statistical correction is enabled, and the final output is obtained through adaptive texture-guided fusion. Specifically, the texture map is computed from local variance using a 15 × 15 window and further smoothed by a 21 × 21 Gaussian filter. When edge enhancement is enabled, the preserved LH and HL components are further scaled by an enhancement factor γ. Edge enhancement was disabled in the reported experiments. This study adopts overall accuracy (OA), precision (P), recall (R), intersection over union (IoU), F1-score (F1), and boundary intersection over union (BIoU) as the evaluation metrics.

4.1.3. Compared Methods

To thoroughly assess the effectiveness of the proposed method in cross-domain building extraction, several representative unsupervised domain adaptation (UDA) approaches were selected as comparison methods. These methods cover a diverse range of adaptation paradigms and therefore provide a comprehensive and meaningful benchmark. Specifically, Source is adopted as the baseline, where a model trained only on the labeled source domain is directly applied to the target domain for inference, thereby reflecting the initial cross-domain generalization ability of the model. AdaptSegNet performs adversarial learning in the output space to reduce the discrepancy between source- and target-domain predictions, thereby representing a typical output-level alignment method. FDANet leverages the Fourier transform to align low-frequency information across the two domains in the frequency domain, thereby mitigating appearance-style discrepancy, and can be regarded as a representative image-level adaptation approach. BDL achieves domain adaptation through bidirectional interaction between image translation and segmentation, highlighting collaborative optimization at the image and task levels. FADA introduces a fine-grained category-level alignment mechanism to enhance the consistency of cross-domain semantic representations. FL-Domain, specifically designed for building extraction, jointly performs adaptation at the image, feature, and output levels, demonstrating strong task relevance. Overall, these methods address domain shift from different perspectives, including style transfer, feature alignment, output-space regularization, and self-training, thereby providing a comprehensive and representative basis for validating the effectiveness of the proposed method. In addition, GLGAN and MBMTA-GAN are included as recent GAN-based UDA methods for remote sensing semantic segmentation, providing further comparisons with advanced image-level adaptation strategies.

4.2. Results with Massachusetts as the Source Domain

4.2.1. Quantitative Analysis

As reported in Table 1, when Massachusetts is used as the source domain, the proposed WCMNet achieves the best overall performance in both the M-P and M-W transfer tasks. In the M-P task, most comparison methods suffer from severe performance degradation because the model is transferred from the lower-resolution Massachusetts dataset to the

higher-resolution and structurally more complex Potsdam dataset. Although FDANet and FL-Domain obtain relatively high Precision, their Recall and IoU remain low, indicating that these methods tend to generate conservative predictions and only identify the most salient building regions while missing many true building pixels. By contrast, WCMNet achieves 87.60% OA, 94.60% Precision, 56.48% Recall, 54.70% IoU, 70.71% F1-score, and 74.52% BIoU, substantially outperforming all comparison methods. In particular, compared with the strongest competing method in terms of IoU and F1-score, WCMNet improves IoU and F1-score by 27.35 and 27.78 percentage points, respectively. The remarkable improvement in BIoU further demonstrates that WCMNet can effectively preserve boundary structures under large domain discrepancies.

Table 1. Quantitative evaluation results of different methods when using Massachusetts as the source domain.

Network Model	Transfer Method	OA/%	P/%	R/%	IoU/%	F1/%	BIoU%
Source	M-P	73.42 ± 0.59	49.92 ± 2.57	21.53 ± 2.40	17.65 ± 1.30	30.00 ± 1.87	36.20 ± 1.82
AdaptSegNet [10]	M-P	79.23 ± 0.46	85.77 ± 0.23	25.95 ± 2.16	24.87 ± 1.97	39.82 ± 2.52	29.58 ± 16.89
FDANet [30]	M-P	77.57 ± 0.20	90.95 ± 1.51	17.11 ± 1.19	16.82 ± 1.09	28.79 ± 1.60	35.40 ± 0.91
BDL	M-P	72.74 ± 0.33	46.79 ± 1.44	20.65 ± 1.08	16.73 ± 0.90	28.65 ± 1.32	14.66 ± 1.75
FADA	M-P	79.52 ± 0.70	83.99 ± 13.90	29.18 ± 4.03	27.35 ± 2.07	42.93 ± 2.55	25.22 ± 1.61
FL-Domain	M-P	78.45 ± 0.21	90.71 ± 1.40	20.85 ± 1.29	20.40 ± 1.17	33.89 ± 1.62	30.50 ± 13.69
GLGAN [31]	M-P	76.08 ± 0.27	60.80 ± 1.65	27.58 ± 0.62	23.41 ± 0.20	37.94 ± 0.26	22.57 ± 0.18
MBMTA-GAN [32]	M-P	77.50 ± 0.24	89.58 ± 0.12	17.11 ± 1.07	16.78 ± 1.03	28.72 ± 1.50	16.41 ± 1.14
WCMNet	M-P	87.60 ± 0.33	94.60 ± 1.24	56.48 ± 2.15	54.70 ± 1.60	70.71 ± 1.34	74.52 ± 0.52
Source	M-W	84.70 ± 1.07	39.39 ± 1.94	68.99 ± 1.67	33.44 ± 1.01	50.12 ± 1.14	33.56 ± 1.00
AdaptSegNet	M-W	92.56 ± 0.04	76.62 ± 0.92	47.68 ± 1.61	41.61 ± 0.95	58.76 ± 0.95	30.53 ± 12.80
FDANet	M-W	94.87 ± 0.17	79.73 ± 1.04	72.31 ± 3.65	61.04 ± 1.99	75.80 ± 1.53	48.16 ± 9.94
BDL	M-W	86.93 ± 4.34	47.05 ± 10.84	71.25 ± 17.34	38.09 ± 2.19	55.15 ± 2.30	22.44 ± 3.97
FADA	M-W	92.18 ± 1.72	73.24 ± 14.70	48.21 ± 3.15	41.03 ± 6.90	58.02 ± 6.95	34.97 ± 9.14
FL-Domain	M-W	94.86 ± 0.18	80.24 ± 4.53	71.94 ± 8.84	60.76 ± 3.75	75.56 ± 2.90	40.96 ± 1.26
GLGAN	M-W	90.25 ± 0.68	55.32 ± 2.98	65.81 ± 1.79	42.92 ± 1.04	60.06 ± 1.02	41.98 ± 0.57
MBMTA-GAN	M-W	89.85 ± 1.16	56.70 ± 8.41	43.52 ± 4.82	32.31 ± 0.09	48.84 ± 0.11	19.78 ± 0.04
WCMNet	M-W	96.62 ± 0.07	88.16 ± 1.41	80.54 ± 2.38	72.65 ± 0.98	84.16 ± 0.66	60.12 ± 0.32

In the M-W task, the domain gap between Massachusetts and WHU is relatively smaller than that in M-P, and several comparison methods show improved adaptation performance. Nevertheless, WCMNet still achieves the best results across all evaluation metrics, with 96.62% OA, 88.16% Precision, 80.54% Recall, 72.65% IoU, 84.16% F1-score, and 60.12% BIoU. Compared with FDANet, WCMNet improves IoU, F1-score, and BIoU by 11.61, 8.36, and 11.96 percentage points, respectively. These improvements indicate that the proposed method not only enhances the overlap between predicted and reference building regions but also improves the geometric accuracy of building boundaries.

Overall, the quantitative results show that WCMNet achieves a more favorable balance between Precision and Recall than existing UDA methods. The proposed method avoids both overly conservative predictions and excessive false positives, thereby improving building completeness, regional consistency, and boundary quality simultaneously. This confirms the effectiveness of combining wavelet-domain style alignment, high-frequency structural preservation, and CNN–Mamba-based global–local dynamic feature modeling for cross-domain building extraction.

4.2.2. Visual Analysis

The building extraction visualization results are shown in Figure 4. In the M-W task, most methods are able to recover the main building regions to some extent because the

domain discrepancy between Massachusetts and WHU is relatively moderate. However, obvious differences can still be observed in the highlighted and enlarged regions. Source and AdaptSegNet tend to miss small buildings and incomplete roof regions, resulting in fragmented building masks. FDANet and FL-Domain improve the recovery of some building shapes, but they still produce discontinuous predictions and blurred local contours. GLGAN and MBMTA-GAN show unstable adaptation performance, with either missing building components or introducing background noise. By contrast, WCMNet produces more complete building masks in the boxed regions, and its predictions are more consistent with the ground-truth labels in terms of both building shape and spatial extent.

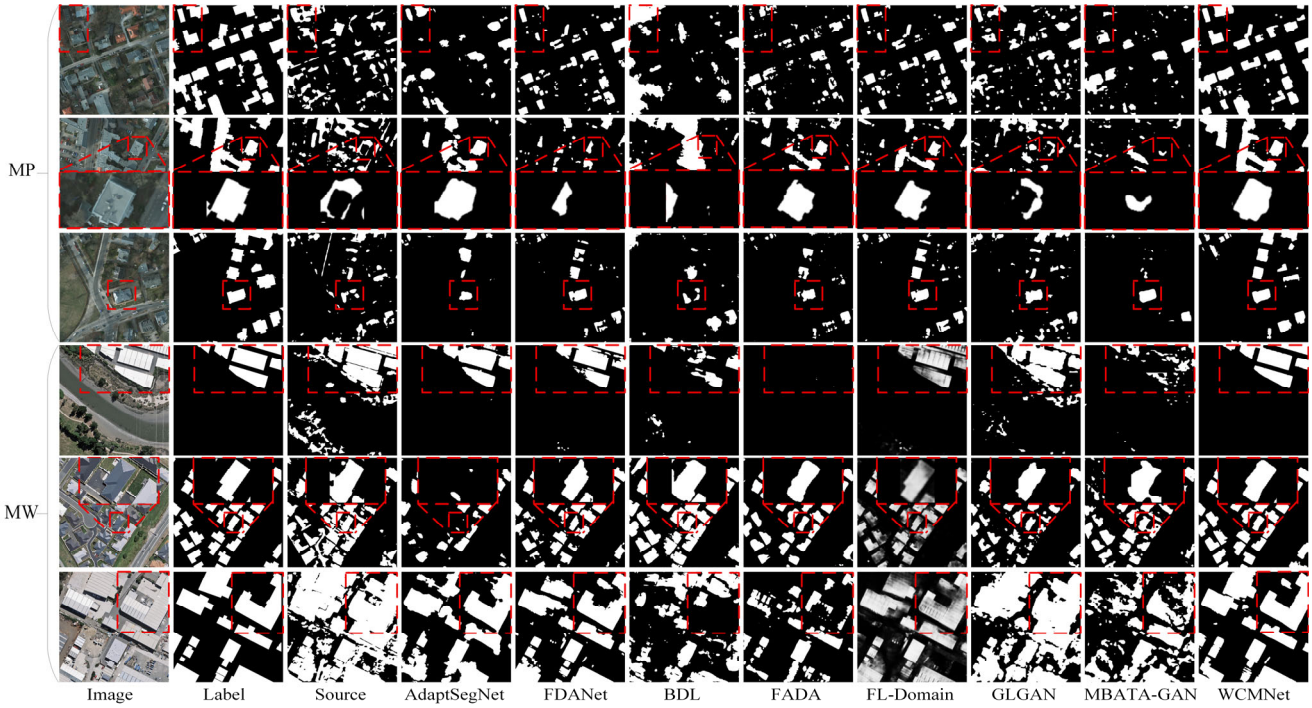


Figure 4. Visualization results of different methods when transferring from the domain of Massachusetts.

In the more challenging M-P task, the advantage of WCMNet becomes more evident. Due to the large resolution and scene gap between Massachusetts and Potsdam, many comparison methods fail to recover complete building structures. As shown in the enlarged regions of Figure 4, some methods only extract partial building bodies, while others confuse roads, bare ground, or roof-like background regions with buildings. These errors lead to either severe omission or false-positive predictions. WCMNet, however, can better preserve the overall building layout and recover more complete building regions. In the highlighted areas, the proposed method maintains more accurate building shapes, reduces holes inside building masks, and suppresses background interference more effectively than the comparison methods.

Figure 5 further provides boundary-level visual comparisons. The boundary results show that most comparison methods suffer from contour discontinuity, boundary shift, or incomplete edge responses. In particular, in the boxed and enlarged regions, Source, BDL, GLGAN, and MBMTA-GAN often generate broken or inaccurate boundaries, while FDANet and FL-Domain still show local contour deviations despite their improved regional extraction performance. In contrast, the boundaries produced by WCMNet are more continuous and more consistent with the reference contours. The proposed method better preserves building corners, elongated edges, and closed boundary structures, especially in complex roof regions and densely distributed building areas. This visual evidence

is consistent with the BIoU results in Table 1 and further confirms that WCMNet can effectively maintain high-frequency structural information during cross-domain transfer.

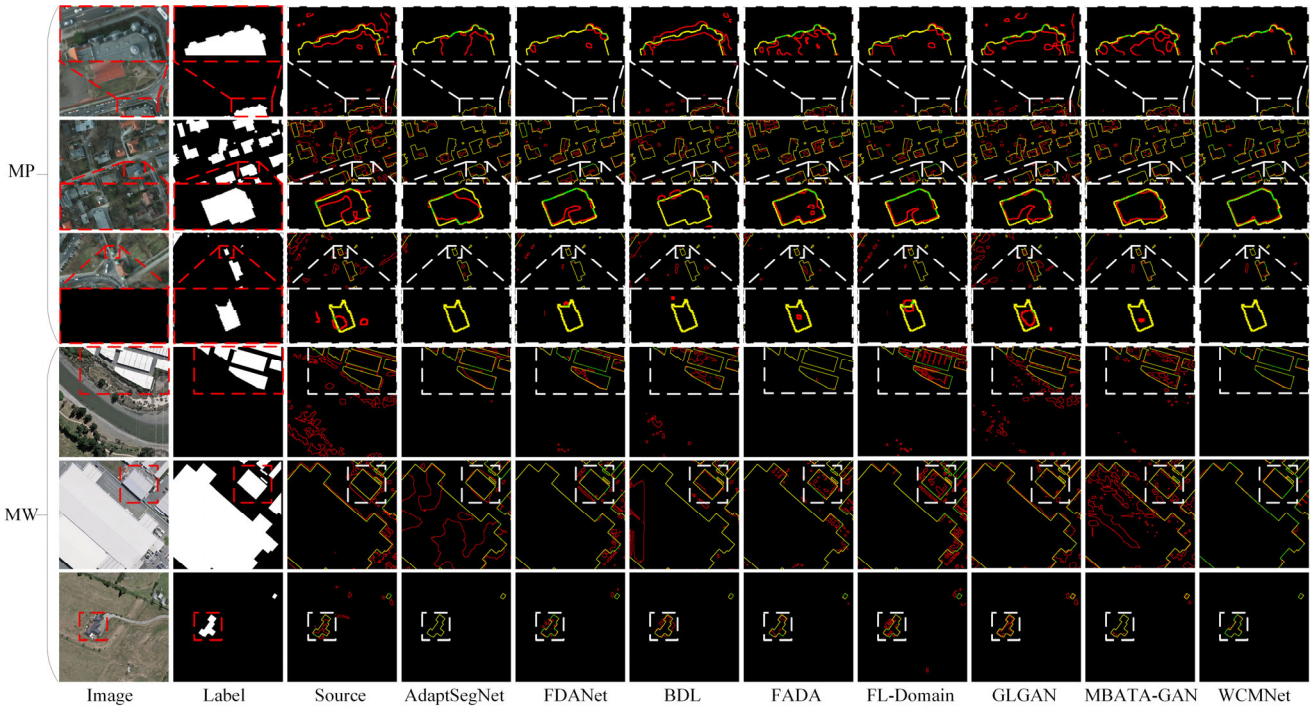


Figure 5. Visualization results of boundary extraction using different methods during migration from the Massachusetts domain.

4.3. Results with WHU as the Source Domain

4.3.1. Quantitative Analysis

As reported in Table 2, when WHU is used as the source domain, WCMNet consistently achieves the best overall performance in both the W-M and W-P transfer tasks. In the W-M task, the model is transferred from WHU to the lower-resolution Massachusetts dataset, where differences in spatial resolution, image texture, and scene appearance introduce substantial domain discrepancies. In this challenging setting, WCMNet achieves 91.09% OA, 78.03% Precision, 72.95% Recall, 60.37% IoU, 75.29% F1-score, and 87.12% BIoU. Compared with FDANet, WCMNet improves OA, Recall, IoU, F1-score, and BIoU by 4.03, 10.51, 13.05, 11.05, and 34.40 percentage points, respectively. The large improvement in BIoU indicates that WCMNet provides a clear advantage in boundary preservation, rather than merely improving coarse regional classification.

In the W-P task, FDANet and FL-Domain show relatively strong competitiveness, suggesting that frequency-domain alignment and multi-level adaptation can alleviate domain discrepancy to some extent. Nevertheless, WCMNet still obtains the best results in most metrics, including 93.64% OA, 86.51% Recall, 78.25% IoU, 87.80% F1-score, and 87.84% BIoU. Compared with FDANet, WCMNet improves Recall, IoU, F1-score, and BIoU by 12.44, 6.69, 4.38, and 2.97 percentage points, respectively. Although its Precision is slightly lower than that of FDANet and FL-Domain, the much higher Recall and IoU indicate that WCMNet can recover building regions more completely while maintaining strong segmentation quality.

These results demonstrate that WCMNet is particularly effective when both appearance discrepancy and structural discrepancy exist between domains. Existing methods may obtain high Precision in some cases, but they often sacrifice Recall and regional completeness. In contrast, WCMNet maintains stable improvements in IoU, F1-score, and BIoU

across both transfer tasks, indicating that the proposed framework can simultaneously improve building completeness, reduce omission errors, and enhance boundary delineation.

Table 2. Quantitative evaluation results of different methods when using WHU as the source domain.

Network Model	Transfer Method	OA/%	P/%	R/%	IoU/%	F1/%	BIoU%
Source	W-M	80.85 ± 0.83	48.47 ± 2.53	34.42 ± 11.67	24.74 ± 5.59	39.50 ± 7.20	46.23 ± 7.55
AdaptSegNet	W-M	83.91 ± 0.23	66.53 ± 1.48	27.23 ± 0.73	23.95 ± 0.76	38.64 ± 0.98	18.72 ± 1.77
FDANet	W-M	87.06 ± 0.40	66.23 ± 2.18	62.44 ± 1.92	47.32 ± 0.01	64.24 ± 0.01	52.72 ± 18.43
BDL	W-M	74.24 ± 0.61	35.72 ± 0.15	48.05 ± 4.78	25.74 ± 1.45	40.93 ± 1.84	21.96 ± 1.94
FADA	W-M	80.68 ± 0.50	47.63 ± 1.51	37.32 ± 2.31	26.43 ± 0.70	41.81 ± 0.87	19.83 ± 0.86
FL-Domain	W-M	83.21 ± 0.74	55.62 ± 2.98	50.14 ± 4.42	35.69 ± 1.02	52.61 ± 1.11	33.56 ± 0.45
GLGAN	W-M	79.65 ± 0.28	44.28 ± 0.69	36.08 ± 0.87	24.81 ± 0.19	39.75 ± 0.25	22.55 ± 0.25
MBMTA-GAN	W-M	83.47 ± 0.01	63.32 ± 3.56	27.29 ± 5.93	23.40 ± 3.92	37.84 ± 5.15	21.11 ± 2.85
WCMNet	W-M	91.09 ± 0.45	78.03 ± 4.33	72.95 ± 3.87	60.37 ± 0.06	75.29 ± 0.04	87.12 ± 0.06
Source	W-P	77.32 ± 0.52	74.77 ± 5.44	22.41 ± 6.32	20.66 ± 5.02	34.10 ± 6.90	34.69 ± 6.51
AdaptSegNet	W-P	85.86 ± 0.48	91.41 ± 3.35	51.55 ± 0.28	49.15 ± 0.72	65.91 ± 0.64	55.85 ± 16.56
FDANet	W-P	92.39 ± 0.33	95.56 ± 2.82	74.07 ± 2.18	71.56 ± 0.45	83.42 ± 0.30	84.87 ± 0.37
BDL	W-P	80.74 ± 1.39	75.35 ± 9.19	41.52 ± 2.31	36.37 ± 0.38	53.34 ± 0.41	32.16 ± 1.21
FADA	W-P	87.02 ± 0.55	91.66 ± 0.45	56.18 ± 2.60	53.44 ± 2.20	69.64 ± 1.87	46.79 ± 2.16
FL-Domain	W-P	92.06 ± 0.06	94.91 ± 1.81	74.07 ± 1.81	71.21 ± 0.65	83.19 ± 0.45	74.33 ± 14.78
GLGAN	W-P	83.48 ± 0.24	75.92 ± 2.23	55.30 ± 1.78	47.02 ± 0.43	63.96 ± 0.40	39.71 ± 0.27
MBMTA-GAN	W-P	81.74 ± 1.56	92.76 ± 2.75	33.92 ± 7.56	32.94 ± 6.82	49.36 ± 7.72	29.00 ± 4.75
WCMNet	W-P	93.64 ± 0.29	89.27 ± 2.11	86.51 ± 1.50	78.25 ± 0.51	87.80 ± 0.33	87.84 ± 0.47

4.3.2. Visual Analysis

The building extraction results in Figure 6 further validate the quantitative improvements reported in Table 2. In the W-M task, the target Massachusetts images contain relatively low-resolution building structures and complex background textures. As shown in the highlighted and enlarged regions, Source and AdaptSegNet fail to recover many building regions, especially small buildings, narrow roof structures, and fragmented residential areas. Their predictions are incomplete and often disconnected. BDL and FADA recover some building bodies, but they still show obvious structural degradation, including broken masks, rounded corners, and missing roof details. FDANet and FL-Domain produce better regional predictions than these methods, but they still suffer from local over-smoothing and boundary adhesion, particularly in densely distributed buildings.

In the W-P task, the target Potsdam images contain richer structural details and more complex building layouts. Several comparison methods can detect the main building regions, but their predictions are often inaccurate in the boxed local areas. For example, adjacent buildings are easily merged, small buildings are missed, and roof boundaries are not clearly separated from the surrounding background. In contrast, WCMNet produces more complete and coherent building masks. In the enlarged regions of Figure 6, WCMNet better preserves the shapes of large buildings, the separation of adjacent structures, and the integrity of small-scale targets. These results indicate that the proposed global–local dynamic representation can improve both semantic completeness and local structural accuracy.

Figure 7 presents the corresponding boundary visualization results. The boundary maps reveal that most comparison methods have difficulty maintaining accurate contours after domain transfer. Source and AdaptSegNet exhibit many missing boundary segments, while BDL and FADA show contour deformation and boundary drift. FDANet and FL-Domain improve the overall boundary response, but they still suffer from boundary adhesion and local misalignment in complex regions. In the highlighted and enlarged areas, WCMNet produces boundaries that are more continuous, sharper, and more consistent with the reference labels. It better preserves corners, straight edges, and closed contours

of buildings, indicating that the proposed wavelet-guided structural preservation and CNN–Mamba hybrid modeling are effective in enhancing boundary-level adaptation.

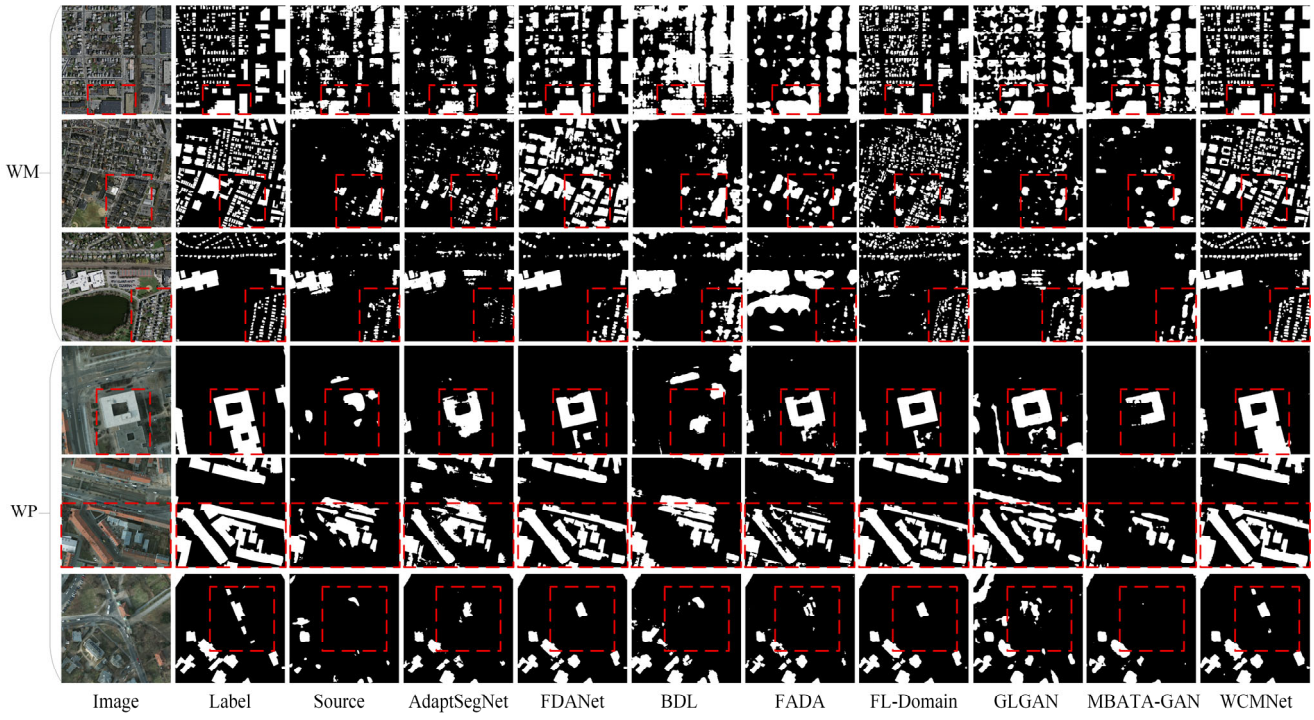


Figure 6. Visualization results of different methods when transferring from the domain WHU.

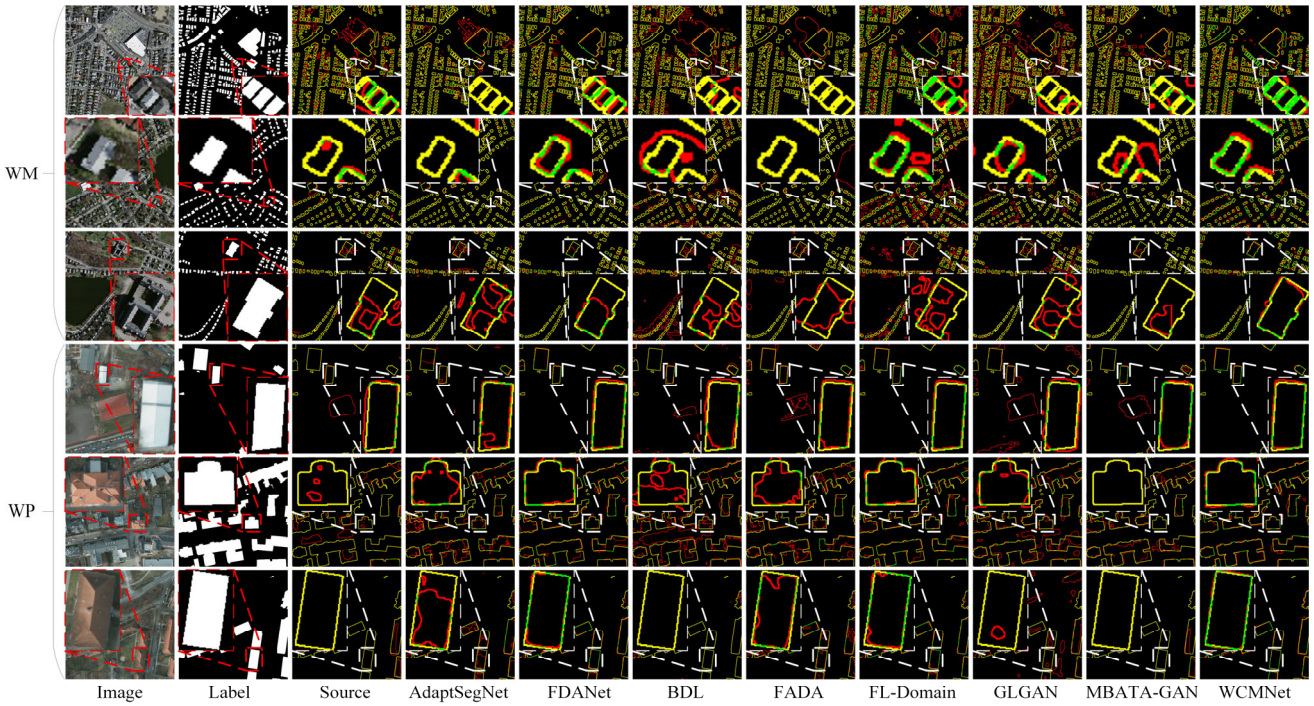


Figure 7. Visualization results of boundary extraction using different methods during migration from the WHU domain.

4.4. Results with Potsdam as the Source Domain

4.4.1. Quantitative Analysis

As reported in Table 3, when Potsdam is used as the source domain, WCMNet achieves the best overall performance in both the P-W and P-M transfer tasks. In the P-W task,

WCMNet obtains 95.83% OA, 77.28% Precision, 88.56% Recall, 70.24% IoU, 82.52% F1-score, and 57.98% BIoU. Compared with FL-Domain, which achieves the highest IoU among the comparison methods, WCMNet improves IoU and F1-score by 10.29 and 7.57 percentage points, respectively. Moreover, compared with FDANet, which obtains the highest BIoU among the competing methods, WCMNet improves BIoU by 7.41 percentage points. Although the Recall of WCMNet is slightly lower than that of FDANet and FL-Domain, its substantially higher Precision, IoU, F1-score, and BIoU indicate that the proposed method better suppresses false positives while maintaining high building detection completeness.

Table 3. Quantitative evaluation results of different methods when using Potsdam as the source domain.

Network Model	Transfer Method	OA/%	P/%	R/%	IoU/%	F1/%	BIoU%
Source	P-W	72.49 ± 2.00	20.70 ± 0.66	51.86 ± 3.51	17.34 ± 0.07	29.56 ± 0.11	21.73 ± 0.08
AdaptSegNet	P-W	90.22 ± 0.59	54.76 ± 2.37	70.56 ± 1.34	44.54 ± 1.03	61.63 ± 0.99	24.92 ± 8.99
FDANet	P-W	92.22 ± 0.47	60.00 ± 1.82	90.48 ± 1.16	56.42 ± 1.16	72.14 ± 0.95	50.57 ± 0.51
BDL	P-W	87.76 ± 2.33	47.00 ± 8.31	51.04 ± 9.05	31.73 ± 0.29	48.17 ± 0.33	17.24 ± 1.16
FADA	P-W	92.37 ± 2.46	62.84 ± 10.63	81.73 ± 0.46	54.98 ± 7.96	70.78 ± 6.65	41.51 ± 5.62
FL-Domain	P-W	93.24 ± 0.45	63.84 ± 2.06	90.81 ± 1.10	59.95 ± 1.34	74.95 ± 1.05	44.96 ± 11.51
WCMNet	P-W	95.83 ± 0.21	77.28 ± 1.95	88.56 ± 1.48	70.24 ± 0.68	82.52 ± 0.47	57.98 ± 0.32
GLGAN	P-W	81.46 ± 1.20	34.83 ± 1.05	76.12 ± 4.79	31.36 ± 0.03	47.75 ± 0.04	19.72 ± 0.55
MBMTA-GAN	P-W	88.46 ± 0.34	48.90 ± 0.90	81.52 ± 0.35	44.01 ± 0.62	61.13 ± 0.60	22.47 ± 0.61
Source	P-M	38.52 ± 17.14	20.74 ± 0.84	80.52 ± 26.98	19.47 ± 0.93	32.59 ± 1.29	34.33 ± 9.64
AdaptSegNet	P-M	77.46 ± 1.32	41.61 ± 1.51	51.41 ± 6.41	29.75 ± 1.39	45.85 ± 1.65	25.59 ± 14.91
FDANet	P-M	82.53 ± 0.73	52.25 ± 1.58	72.99 ± 3.46	43.73 ± 0.14	60.86 ± 0.13	60.94 ± 20.67
BDL	P-M	72.02 ± 0.24	29.49 ± 0.04	36.18 ± 0.82	19.40 ± 0.22	32.49 ± 0.31	13.11 ± 0.04
FADA	P-M	78.66 ± 1.03	44.58 ± 1.77	59.68 ± 0.70	34.24 ± 0.82	51.02 ± 0.91	38.94 ± 17.10
FL-Domain	P-M	83.12 ± 1.66	52.93 ± 2.69	70.60 ± 3.98	43.28 ± 0.30	60.42 ± 0.29	76.14 ± 0.18
WCMNet	P-M	88.08 ± 0.84	65.44 ± 3.48	76.81 ± 3.73	54.54 ± 0.54	70.58 ± 0.45	81.21 ± 0.33
GLGAN	P-M	63.98 ± 2.08	30.52 ± 0.98	73.03 ± 2.67	27.41 ± 0.42	43.03 ± 0.52	23.38 ± 0.33
MBMTA-GAN	P-M	72.23 ± 1.65	36.30 ± 1.16	64.82 ± 4.23	30.29 ± 0.12	46.49 ± 0.14	25.13 ± 0.11

In the P-M task, the model is transferred from the high-resolution Potsdam dataset to the lower-resolution Massachusetts dataset, which causes considerable loss of fine structural details. Most comparison methods show clear performance degradation under this challenging setting. BDL obtains only 19.40% IoU, indicating its limited robustness to large resolution and appearance discrepancies. FDANet and FL-Domain remain competitive, achieving IoU values of 43.73% and 43.28%, respectively, but they still lag behind WCMNet by a large margin. WCMNet achieves 88.08% OA, 65.44% Precision, 76.81% Recall, 54.54% IoU, 70.58% F1-score, and 81.21% BIoU. Compared with FDANet, WCMNet improves IoU and F1-score by 10.81 and 9.72 percentage points, respectively. Compared with FL-Domain, WCMNet further improves BIoU by 5.07 percentage points.

The quantitative results demonstrate that WCMNet maintains strong robustness when transferring from Potsdam to both WHU and Massachusetts. The consistent improvements in IoU and F1-score indicate better regional segmentation quality, while the gains in BIoU confirm their advantage in boundary preservation. These results verify that the proposed combination of low-frequency style alignment, high-frequency structural preservation, and global–local dynamic feature learning is effective for handling cross-resolution and cross-scene domain adaptation in building extraction.

4.4.2. Visual Analysis

The building extraction visualization results in Figure 8 show that the Source baseline suffers from severe domain shift when Potsdam is used as the source domain. In both P-W

and P-M tasks, Source produces fragmented building masks, large omission errors, and obvious background confusion. These problems are particularly visible in the highlighted and enlarged regions, where building regions are either partially missed or incorrectly merged with surrounding non-building areas. This indicates that the features learned from Potsdam cannot be directly transferred to target domains with different resolutions and scene appearances.

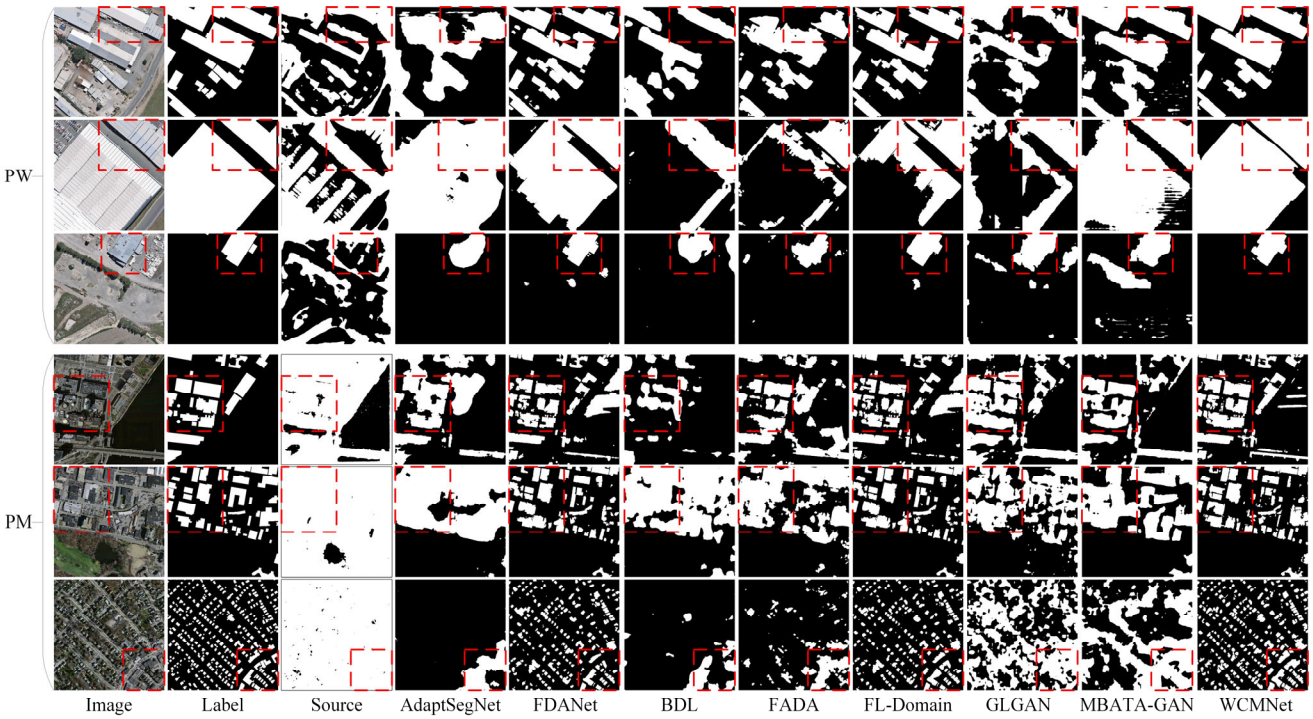


Figure 8. Visualization results of different methods when transferring from the domain of Potsdam.

AdaptSegNet and BDL recover some building regions, but their predictions remain coarse and unstable. They frequently produce boundary expansion, incomplete building interiors, and adhesion between adjacent buildings. FADA improves the recovery of some structural regions, but it still misses small buildings and produces discontinuous masks. FDANet and FL-Domain show more competitive visual results because they can reduce part of the appearance discrepancy through frequency-domain or multi-level adaptation. However, in the boxed regions of Figure 8, these methods still exhibit blurred contours, fragmented edges, and insufficient recovery of fine local structures, especially in dense building areas and low-resolution target scenes.

In contrast, WCMNet achieves the most accurate and stable visual performance in both P-W and P-M tasks. In the highlighted and enlarged regions, WCMNet recovers more complete building bodies, maintains clearer building contours, and better preserves the separation between adjacent buildings. For large buildings, the proposed method reduces internal holes and maintains more coherent shapes. For small or densely distributed buildings, it produces fewer omissions and less background confusion. These advantages indicate that WCMNet can effectively improve both the completeness and structural consistency of building extraction under challenging cross-domain conditions.

The boundary visualization results in Figure 9 further confirm the superiority of WCMNet in boundary delineation. Most comparison methods generate incomplete or inaccurate boundary responses after domain transfer. Source and BDL show severe boundary loss, while AdaptSegNet, FADA, FDANet, and FL-Domain still exhibit boundary drift, contour discontinuity, or adhesion between neighboring buildings. In the enlarged local regions,

WCMNet generates more complete and well-aligned boundary contours, with clearer corners and more continuous edges. This demonstrates that the proposed wavelet-guided alignment can preserve high-frequency structural cues, while the CNN–Mamba hybrid backbone further strengthens global structural consistency and local boundary refinement. Therefore, WCMNet achieves superior visual performance not only in region-level extraction but also in boundary-level reconstruction.

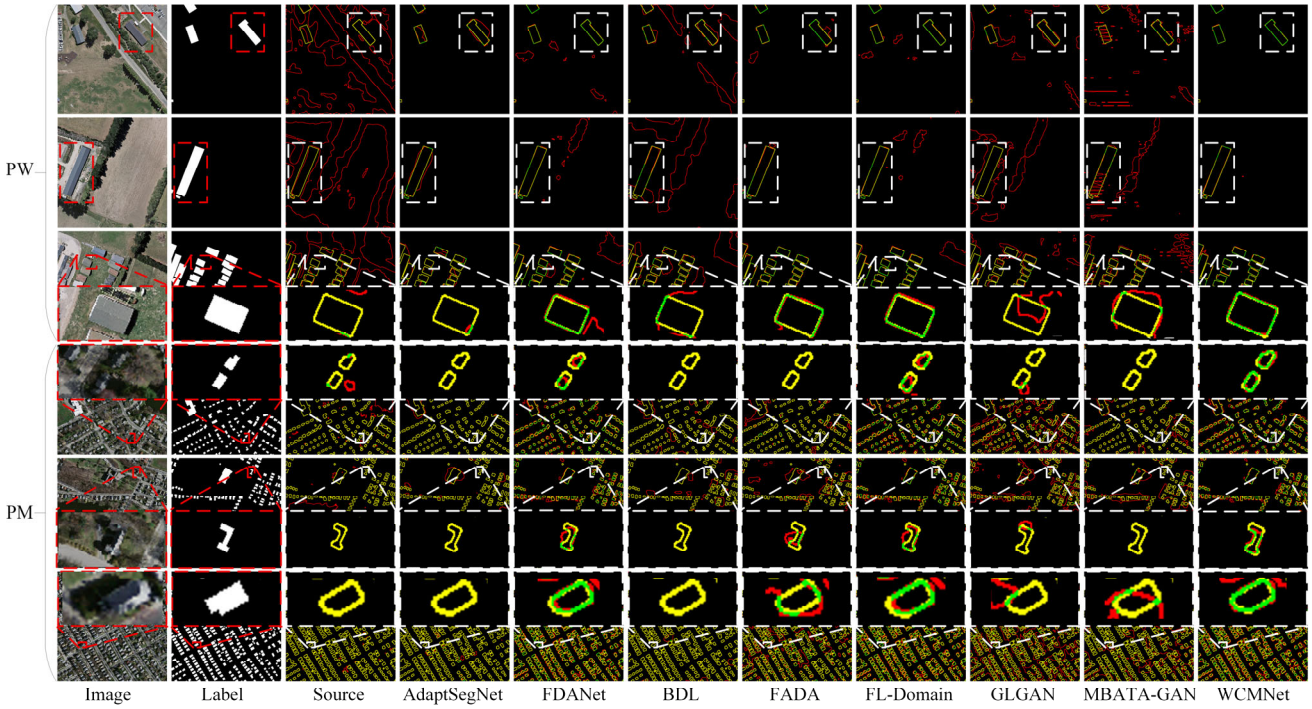


Figure 9. Visualization results of boundary extraction using different methods during migration from the Potsdam domain.

5. Discussion

To further investigate the effectiveness of the proposed WCMNet and clarify the contribution of each component, we conducted both progressive and independent ablation analyses on the WHU → Massachusetts transfer task. This transfer setting is representative and challenging because it involves obvious differences in spatial resolution, texture distribution, and scene appearance between the source and target domains. Therefore, it provides a suitable testbed for evaluating whether each module can effectively improve cross-domain building extraction.

5.1. Progressive Contribution of the Main Modules

As reported in Table 4, the progressive ablation results demonstrate the complementary contributions of the proposed modules. The baseline model achieves 81.44% OA, 50.26% Precision, 26.16% Recall, 20.78% IoU, and 34.41% F1-score. This indicates that directly transferring a model trained on the source domain to the target domain results in severe performance degradation, mainly due to domain discrepancies in spatial resolution, image texture, and building appearance.

After introducing the proposed Mamba Wavelet Alignment (MWA) module, the performance improves substantially. Specifically, Base + MWA achieves 88.79% OA, 78.05% Precision, 55.32% Recall, 47.87% IoU, and 64.75% F1-score. Compared with the baseline, OA, Precision, Recall, IoU, and F1-score improved by 7.35, 27.79, 29.16, 27.09, and 30.34 percentage points, respectively. These improvements demonstrate that the wavelet-guided

image-level alignment effectively reduces cross-domain appearance discrepancy by aligning low-frequency style information while preserving high-frequency structural cues. As a result, the model can better maintain building contours and reduce segmentation distortion caused by domain shift.

Table 4. Quantitative evaluation results of ablation experiments on the WHU-Mass dataset.

Network Model	OA/%	P/%	R/%	IoU/%	F1/%
Base	81.44	50.26	26.16	20.78	34.41
Base + MWA	88.79	78.05	55.32	47.87	64.75
WCMNet	90.77	74.97	75.69	60.41	75.32

When the complete WCMNet is adopted, the performance is further improved to 90.77% OA, 74.97% Precision, 75.69% Recall, 60.41% IoU, and 75.32% F1-score. Compared with Base + MWA, the complete model further improves Recall, IoU, and F1-score by 20.37, 12.54, and 10.57 percentage points, respectively. Although Precision slightly decreases, the substantial increase in Recall and IoU indicates that WCMNet detects building regions more comprehensively and effectively reduces omission errors. This improvement mainly benefits from the CNN–Mamba-based global–local dynamic feature learning and the pseudo-label-based target-domain adaptation strategy, which enhance both structural representation and target-domain generalization. Overall, the progressive ablation results confirm that MWA primarily alleviates image-level style discrepancy, while the complete WCMNet further strengthens feature representation and target-domain adaptation.

5.2. Independent Contribution of Key Components

To more clearly distinguish the contribution of each component, we further conducted independent ablation experiments, as shown in Table 5. Compared with the baseline, Base + Mamba Only improves OA from 81.44% to 87.18%, IoU from 20.78% to 37.07%, and F1-score from 34.41% to 54.09%. These results indicate that introducing Mamba-based global context modeling can significantly enhance the ability of the model to capture long-range dependencies and suppress background interference. However, its Recall remains only 40.57%, suggesting that using Mamba alone is still insufficient to fully recover complete building regions under cross-domain conditions.

Table 5. Quantitative results of independent ablation experiments for key components of WCMNet on the WHU-Massachusetts transfer task.

Network Model	OA/%	P/%	R/%	IoU/%	F1/%
Base	81.44	50.26	26.16	20.78	34.41
Base + Mamba Only	87.18	81.14	40.57	37.07	54.09
Base + Gated Fusion Only	83.88	59.18	43.17	33.26	49.92
Base + Pseudo-Labeling Only	87.49	80.80	43.01	39.02	56.14
WCMNet	90.77	74.97	75.69	60.41	75.32

Base + Gated Fusion Only achieves 83.88% OA, 59.18% Precision, 43.17% Recall, 33.26% IoU, and 49.92% F1-score. Compared with the baseline, the improvement in Recall and IoU indicates that the gated fusion mechanism helps balance local details and global semantic information. Nevertheless, without the explicit global modeling capability of Mamba and the target-domain refinement provided by pseudo-labeling, its overall performance remains limited. This suggests that gated fusion is more effective when it works together with complementary local and global feature branches.

Base + Pseudo-Labeling Only obtains 87.49% OA, 80.80% Precision, 43.01% Recall, 39.02% IoU, and 56.14% F1-score. This setting achieves better performance than Base + Mamba Only and Base + Gated Fusion Only in terms of IoU and F1-score, indicating that pseudo-label-based target-domain learning can improve the adaptation ability of the model. However, the Recall is still much lower than that of the complete WCMNet, which means that pseudo-labeling alone cannot sufficiently compensate for weak feature representation or structural inconsistency caused by domain shift.

The complete WCMNet achieves the best overall performance, especially in Recall, IoU, and F1-score. Compared with Base + Mamba Only, WCMNet improves Recall, IoU, and F1-score by 35.12, 23.34, and 21.23 percentage points, respectively. Compared with Base + Pseudo-Labeling Only, WCMNet further improves Recall, IoU, and F1-score by 32.68, 21.39, and 19.18 percentage points, respectively. These results clearly show that the performance gain of WCMNet does not come from a single component. Instead, it is achieved through the complementary effects of wavelet-guided style alignment, Mamba-based global context modeling, adaptive gated fusion, and pseudo-label-based target-domain refinement. Therefore, the independent ablation results verify that each component contributes positively to the final performance, while their joint integration produces the most robust and complete building extraction results.

5.3. Visual Interpretation of Module Effects

As shown in Figure 10, the visual ablation results further illustrate the progressive effects of different modules on cross-domain building extraction. Under the direct transfer setting, the baseline model produces highly fragmented predictions. Many small and densely distributed buildings are missed, and the detected building regions often exhibit incomplete shapes, internal holes, and blurred or broken boundaries. These results indicate that the baseline model cannot effectively overcome the domain discrepancy between WHU and Massachusetts, resulting in severe omission errors and structural degradation.

After introducing the MWA module, the visual quality is noticeably improved. Compared with the baseline, Base + MWA recovers more building regions and produces more complete building contours. This demonstrates that wavelet-guided low-frequency style alignment can effectively reduce cross-domain appearance discrepancy, while the preservation of high-frequency components helps retain boundary-related structural cues. Nevertheless, some local discontinuities and missing regions still remain, especially in dense building areas and complex background regions. This suggests that image-level alignment alone is insufficient to fully address feature-level distribution shift and structural inconsistency.

In contrast, the complete WCMNet produces the most accurate and structurally coherent results. It recovers more complete building regions, generates clearer boundaries, and better preserves the separation between adjacent buildings. In dense residential areas and complex background scenes, WCMNet substantially reduces omission errors and suppresses false-positive predictions. These improvements indicate that the CNN–Mamba hybrid modeling strategy enhances the joint representation of global semantic dependencies and local structural details, while the pseudo-label-based refinement further improves target-domain adaptation. Overall, the visual results confirm that the proposed modules work collaboratively to improve building completeness, boundary delineation, and cross-domain robustness.

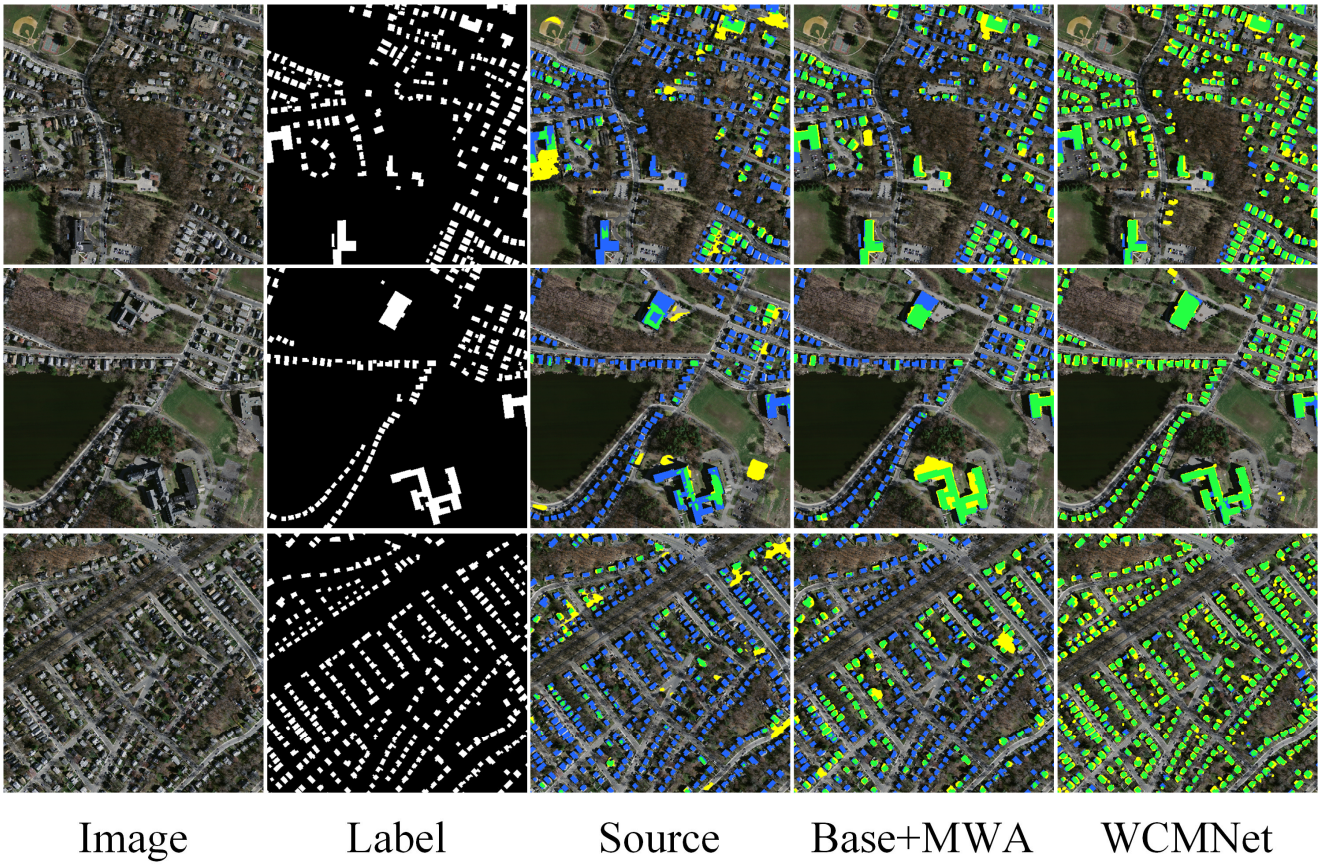


Figure 10. Visualization results of ablation experiments on the WHU-Mass dataset.

6. Conclusions

This paper addresses the problem of cross-domain building extraction in remote sensing imagery, with a particular focus on the challenges faced by existing methods under variations in spatial resolution, sensor characteristics, and scene conditions, including limited generalization ability, boundary degradation, and structural information loss. To overcome these limitations, we propose WCMNet, an unsupervised domain adaptation framework based on wavelet-guided style alignment and CNN–Mamba hybrid modeling. Specifically, the proposed wavelet module performs low-frequency style alignment in the frequency domain while explicitly preserving high-frequency edge structures, thereby mitigating inter-domain appearance discrepancy and reducing contour degradation during domain transfer. Meanwhile, the proposed hybrid backbone, together with an adaptive gated fusion mechanism, dynamically integrates local texture details and global semantic dependencies, which enhances the structural representation capability of buildings under complex scene conditions. Extensive experiments conducted on the Massachusetts, WHU, and Potsdam datasets under multiple cross-resolution, cross-scene, and cross-sensor transfer settings demonstrate that the proposed WCMNet consistently achieves either the best or highly competitive performance. In particular, its superior results in IoU, F1-score, and BIoU verify its effectiveness in preserving building completeness, improving boundary delineation quality, and enhancing cross-domain generalization ability. Despite these promising results, the proposed method still degrades in performance when facing extreme resolution gaps or highly complex background interference, indicating that there is room for further improvement in cross-scale structural modeling and fine-grained feature discrimination. In future work, we will further explore explicit multi-scale alignment strategies, boundary-aware and topology-aware constraints, and extend the proposed framework to

more challenging and practical scenarios involving multimodal remote sensing data, such as SAR, DSM, and LiDAR, as well as downstream applications, including change detection, post-disaster assessment, and temporal remote sensing analysis.

Author Contributions: Conceptualization, D.Y. and K.H. and R.H.; methodology, D.Y. and K.H.; validation, D.Y., Y.Y. and K.H.; investigation, D.Y., K.H., Y.Y., X.G. (Xianjun Gao) and K.G.; data curation, D.Y., Y.Y. and K.H.; writing—original draft preparation, D.Y. and K.H.; writing—review and editing, D.Y. and K.H.; visualization, D.Y., Y.Y., K.H. and X.G. (Xinlong Gao). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tianjin Key Laboratory of Rail Transit Navigation Positioning and Spatio-temporal Big Data Technology, No. TKL2026A06, Open Project of National Key Laboratory for Waterway Traffic Control (26-3-3), Tibet Autonomous Region Science and Technology Major Project XZ202402ZD0001, Key Project of the Scientific Research Plan of Hubei Provincial Department of Education (D20231304), the China National Science and Technology Major Project (Grant 2024ZD1001003), Open Fund of National Engineering Laboratory for Digital Construction and Evaluation Technology of Urban Rail Transit (No. 2023ZH01), Open Fund of Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake, Ministry of Natural Resources (No. MEMI-2021-2022-08), Tianjin Science and Technology Plan Project (No. 23YFYSHZ00190, No. 23YFZCSN00280), Hunan Natural Science Foundation Department Joint Fund (No. 2024JJ8327), Jiangxi Provincial Natural Science Foundation (No. 20232ACB204032), Hunan Provincial Department of Natural Resources Science and Technology Project (20230153CH).

Data Availability Statement: Publicly available datasets were analyzed in this study. The Massachusetts, WHU, and Potsdam datasets can be found here: https://www.cs.toronto.edu/~vmnih/data (accessed on 25 April 2026) and https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/Default.aspx (accessed on 25 April 2026).

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

References

1. Song, W.; Zhao, Y.; Tu, J.; Chen, M.; Xie, Y.; Cui, X. A visual attention-guided approach for concrete crack detection in complex environments. *Eng. Appl. Artif. Intell.* **2026**, *173*, 114439. [CrossRef]
2. Xie, Y.; Liu, S.; Chen, H.; Cao, S.; Zhang, H.; Feng, D.; Wan, Q.; Zhu, J.; Zhu, Q. Localization, balance, and affinity: A stronger multifaceted collaborative salient object detector in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 4700117. [CrossRef]
3. Ran, S.; Gao, X.; Yang, Y.; Li, S.; Zhang, G.; Wang, P. Building multi-feature fusion refined network for building extraction from high-resolution remote sensing images. *Remote Sens.* **2021**, *13*, 2794. [CrossRef]
4. Yang, D.; Gao, X.; Yang, Y.; Guo, K.; Han, K.; Xu, L. Advances and Future Prospects in Building Extraction from High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 6994–7016. [CrossRef]
5. Chang, J.; Gao, X.; Yang, Y.; Wang, N. Object-oriented building contour optimization methodology for image classification results via generalized gradient vector flow snake model. *Remote Sens.* **2021**, *13*, 2406. [CrossRef]
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2015; pp. 3431–3440.
7. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
8. Yang, D.; Gao, X.; Yang, Y.; Jiang, M.; Guo, K.; Liu, B.; Li, S.; Yu, S. Csa-net: Complex scenarios adaptive network for building extraction for remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 938–953. [CrossRef]
9. Cheng, X.; Han, K.; Xu, J.; Li, G.; Xiao, X.; Zhao, W.; Gao, X. SPFDNet: Water extraction method based on spatial partition and feature decoupling. *Remote Sens.* **2024**, *16*, 3959. [CrossRef]
10. Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; Chandraker, M. Learning to adapt structured output space for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 7472–7481.
11. 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 (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 289–305.

12. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2018; pp. 1989–1998.
13. Li, D.; Yang, J.; Kreis, K.; Torralba, A.; Fidler, S. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 8300–8311.
14. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022.
15. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
16. Xie, Y.; Zhan, N.; Zhu, J.; Xu, B.; Chen, H.; Mao, W.; Luo, X. Landslide extraction from aerial imagery considering context association characteristics. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103950. [CrossRef]
17. Song, W.; Cui, X.; Xie, Y.; Wang, G.; Ma, J. Monocular depth estimation via a detail semantic collaborative network for indoor scenes. *Sci. Rep.* **2025**, *15*, 10990. [CrossRef] [PubMed]
18. Tuia, D.; Persello, C.; Bruzzone, L. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
19. Chen, M.; Xue, H.; Cai, D. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 2090–2099.
20. Mei, K.; Zhu, C.; Zou, J.; Zhang, S. Instance adaptive self-training for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*; IEEE: Piscataway, NJ, USA, 2020; pp. 415–430.
21. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. [CrossRef]
22. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J. Vmamba: Visual state space model. In *Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*; Curran Associates Inc.: Red Hook, NY, USA, 2024; pp. 103031–103063.
23. Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; Ha, J.-W. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 9036–9045.
24. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1501–1510.
25. Li, Y.; Liu, M.-Y.; Li, X.; Yang, M.-H.; Kautz, J. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*; IEEE: Piscataway, NJ, USA, 2018; pp. 453–468.
26. Yang, Y.; Soatto, S. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 4085–4095.
27. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS J. Photogramm. Remote Sens.* **2012**, *I-3*, 293–298. [CrossRef]
28. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
29. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.
30. Peng, D.; Guan, H.; Zang, Y.; Bruzzone, L. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607317. [CrossRef]
31. Ma, X.; Zhang, X.; Ding, X.; Pun, M.-O.; Ma, S. Decomposition-based unsupervised domain adaptation for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5645118. [CrossRef]
32. 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*, 5400515. [CrossRef]

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.