Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

RectMamba combines hierarchical correspondence rectification and bidirectional state space fusion for multimodal remote sensing semantic segmentation

Jiaqi Xu
Yanshan University
Shengtao Sun
xysst@ysu.edu.cn
Yanshan University

Research Article

Keywords: multimodal semantic segmentation, remote sensing, cross-modal alignment, state-space model, Mamba, digital surface model, geometric rectification

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11371964/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

RectMamba combines hierarchical correspondence rectification and bidirectional state space fusion for multimodal remote sensing semantic segmentation

Jiaqi Xu[1] and Shengtao Sun[1*]

[1]School of Artificial Intelligence(School of Software), Hebei Key Laboratory of Computer Virtual Technology and System Integration, Yanshan University, Qinhuangdao, 066004, Hebei, China .

*Corresponding author(s). E-mail(s): xysst@ysu.edu.cn;

Abstract

Multimodal semantic segmentation benefits from complementary optical appearance and digital surface model (DSM) geometry. However, direct fusion remains sensitive to spatial discrepancies, while dense cross-modal attention is costly. We propose RectMamba, a dual-stream framework coupling Hierarchical Correspondence Rectification (HCR) with Cross-Conditioned Bidirectional State-Space Interaction (CC-BSSM). HCR constructs local correspondence volumes, estimates residual displacements by soft-argmax, refines offsets from coarse to fine, and warps DSM features through differentiable sampling. Correspondence entropy suppresses uncertain rectification. At the bottleneck, two CC-BSSM blocks condition each modality on relational descriptors from the other modality, model forward and reverse raster sequences, and adaptively aggregate directional states. A progressive decoder integrates the fused bottleneck with rectified shallow features. Under a unified full-tile protocol, RectMamba achieves mIoU scores of 85.57% on Potsdam and 81.66% on Vaihingen, outperforming the strongest re-trained comparator by 1.66 and 2.63 percentage points, respectively. On Potsdam, removing CC-BSSM or HCR reduces mIoU by 0.43 and 0.14 percentage points, and two blocks provide the best depth–accuracy trade-off. These consistent gains show that RectMamba preserves intramodal specificity while exploiting cross-modal complementarity more effectively than representative multimodal segmentation methods.

Keywords: multimodal semantic segmentation, remote sensing, cross-modal alignment, state-space model, Mamba, digital surface model, geometric rectification

1 Introduction

Semantic segmentation converts very-high-resolution remote sensing imagery into dense land-cover maps and supports applications such as urban mapping, environmental monitoring, and infrastructure inventory. Fully convolutional and encoder-decoder networks established effective local representation learning [1, 2], while remote-sensing architectures such as UNetFormer combined convolutional detail extraction with global context modeling [3]. Nevertheless, optical appearance alone remains ambiguous under shadows, spectral similarity, seasonal variation, and complex urban backgrounds.

Multisource observation offers a natural way to reduce this ambiguity. Optical imagery describes reflectance, color, and texture, whereas digital surface models (DSMs) provide elevation and shape cues; optical and synthetic aperture radar (SAR) imagery likewise exhibit complementary radiometric and scattering characteristics. Reviews of multisource remote sensing fusion show that these complementary measurements can improve scene interpretation, but they also introduce modality-dependent statistics and sensor-specific noise [4, 5]. Early dual-stream systems demonstrated the value of combining appearance with depth or elevation information [6, 7], and subsequent multiscale and gated methods strengthened cross-modal feature exchange [8, 9].

The central difficulty is therefore not simply whether two modalities should be fused, but whether the evidence being fused is spatially and semantically compatible. Even after preprocessing, residual displacement can remain because the source products differ in acquisition geometry, resolution, interpolation, or observation time. Direct feature addition, concatenation, or attention may then associate an optical pixel with an inconsistent height or scattering response. Xiao et al. explicitly addressed translational misalignment in heterogeneous optical-complex-SAR segmentation through dynamic feature alignment before latent-space fusion [21]. HDC-Net further separated shallow alignment-oriented processing from deep semantic interaction, arguing that fusion requirements vary with representation depth [22]. These developments motivate a more explicit treatment of correspondence reliability for optical-DSM segmentation.

Long-range interaction presents a second challenge. Hierarchical Transformer backbones and multimodal Transformer designs can capture nonlocal relationships and have produced strong remote-sensing results [14, 15], but dense token interaction increases computational and memory costs. Mamba replaces pairwise attention with an input-dependent selective state-space recurrence [16]. Visual and remote-sensing adaptations, including VMamba and RS3Mamba, demonstrate the potential of selective scanning for image understanding [17, 18]; Sigma extends the idea to multimodal semantic segmentation [19]. More recently, FMLSNet combined multistage Mamba fusion with large-strip convolution and disentangled representation learning for optical-DSM segmentation [20]. However, translating a one-dimensional state-space scan to multimodal images still requires careful treatment of scan direction, local spatial mixing, and the point at which cross-modal information enters the selective dynamics.

These studies leave a specific design space comparatively underexplored: explicit, uncertainty-aware geometric correspondence followed by efficient cross-conditioned state-space interaction. RectMamba is built around the premise that local spatial consistency should be established before long-range semantic exchange. Its Hierarchical Correspondence Rectification (HCR) module estimates a dense displacement field by normalized local matching and soft-argmax, propagates offsets from deep to shallow, and derives confidence from correspondence entropy. Its Cross-Conditioned Bidirectional State-Space Interaction (CC-BSSM) module then modulates each modality with relational cues from the other modality before bidirectional selective scans, adaptively aggregates forward and reverse states, and exchanges information through residual cross-state updates.

This study makes three main contributions:

- We introduce HCR, a coarse-to-fine rectification mechanism that explicitly predicts a spatially varying two-dimensional offset field from local cross-modal correspondence. Differentiable warping performs geometric correction, while entropy-derived confidence prevents ambiguous matches from dominating fusion.
- We develop CC-BSSM for deep multimodal interaction and stack two blocks at the bottleneck. Cross-conditioning affects the representation entering Mamba, forward and reverse raster states are adaptively aggregated, and residual cross-state exchange preserves two explicit modality streams until bottleneck fusion.
- We establish a controlled evaluation protocol on ISPRS Vaihingen and Potsdam. RectMamba reaches 85.57% and 81.66% mIoU, respectively, while four-way and internal ablations quantify the effects of HCR, CC-BSSM, confidence estimation, coarse-to-fine refinement, scan direction, and state-space depth.

2 Related work

2.1 From fusion placement to modality-aware interaction

Multimodal segmentation methods can first be distinguished by where information exchange occurs. Early fusion combines registered inputs before feature extraction, late fusion merges modality-specific predictions, and intermediate fusion connects feature hierarchies between these extremes [4, 5]. FuseNet established a representative intermediate-fusion design by injecting depth features into an RGB encoder through element-wise summation [6]. For very-high-resolution urban remote sensing, Audebert et al. compared early and late combinations of optical and elevation information and showed that fusion placement materially affects the result [7].

Later methods shifted from fixed fusion toward content-dependent interaction. Crossmodal multiscale fusion aggregates complementary information at several resolutions [8], whereas SA-Gate propagates information in both directions and uses separation-and-aggregation gating to control modality exchange [9]. CMX performs cross-modal correction followed by Transformer-based fusion for RGB-X segmentation [13]. These designs improve feature compatibility through reweighting, propagation, or multiscale aggregation. However, feature correction in the semantic or channel domain

does not necessarily establish geometric correspondence: when two feature vectors originate from different physical locations, a stronger fusion operator may still combine inconsistent evidence.

2.2 Global interaction with Transformers and state-space models

Transformers expanded the range of contextual interaction available to segmentation networks. Swin Transformer introduced hierarchical shifted-window attention [14], while the multilevel multimodal fusion Transformer of Ma et al. alternates modality-specific and cross-modal interaction within a remote-sensing encoder-decoder [15]. Such models are effective at associating spatially separated regions, but attention cost grows with the number of token pairs, and window restriction exchanges global range for efficiency.

Selective state-space models provide a different route to long-range modeling. Mamba makes the state-space parameters input dependent and evaluates the recurrence with a hardware-aware selective scan [16]. VMamba reorganizes scanning for visual representations [17], and RS3Mamba couples a state-space auxiliary branch with a convolutional main branch for remote-sensing segmentation [18]. For multimodal inputs, Sigma adopts a Siamese Mamba network [19]. FMLSNet further develops a multistage optical-DSM design in which Mamba-based fusion is combined with disentangled representation learning and large-strip convolution for boundary enhancement [20]. Collectively, these studies demonstrate that state-space models can support multimodal contextual reasoning without explicit all-pairs attention. They also show that image token ordering is a design variable: a single causal scan is not naturally symmetric, and multimodal interaction can occur before, within, or after state evolution.

2.3 Alignment-aware and stage-aware multimodal fusion

Recent research increasingly treats modality inconsistency as a first-class problem rather than an issue to be absorbed by fusion. In optical-complex-SAR segmentation, Xiao et al. use the optical stream as a reference, dynamically adjust displaced SAR features, and then map heterogeneous representations into a shared latent space [21]. This sequence distinguishes spatial correction from semantic fusion and is particularly relevant when acquisition geometries differ substantially.

A related direction assigns different responsibilities to different network depths. HDC-Net uses interactive shared attention in shallow encoding stages, hierarchical cross-token interaction for deep semantic modeling, and a pyramidal bridge for decoding [22]. FMLSNet similarly distributes processing across boundary-oriented large-strip convolution and multistage Mamba fusion [20]. These architectures support the broader principle that shallow features require detail-preserving compatibility, whereas deep features benefit from long-range semantic interaction.

RectMamba follows this stage-aware principle but adopts a different mechanism and supervision target. HCR constructs a local correspondence distribution at every spatial position, converts its expectation into an explicit displacement, and exposes

both the refined offset and its entropy-derived confidence. CC-BSSM is then applied only after hierarchical rectification and allows cross-modal relations to condition the input to bidirectional selective scans. The resulting separation between *where to sample* and *how to interact* is the main distinction from attention-only correction, implicit alignment, and state-space fusion without an explicit confidence-aware displacement field.

3 Proposed method

3.1 Overall architecture

Let the optical input be $I^o \in \mathbb{R}^{H\times W\times 3}$ and the DSM input be $I^d \in \mathbb{R}^{H\times W\times 1}$. Two independent ResNetV2 encoders extract four hierarchical feature sets, denoted by $\{F_l^o\}_{l=0}^3$ and $\{F_l^d\}_{l=0}^3$, where $l = 0$ and $l = 3$ indicate the shallowest and deepest levels, respectively. The encoders do not share parameters, thereby preserving modality-specific statistics. Both branches can nevertheless be initialized from the same ResNet-compatible pretrained model; the DSM stem is initialized by averaging the optical stem weights along the input-channel dimension and is subsequently optimized independently.

As illustrated in Fig. 1, RectMamba follows an alignment–interaction–decoding pipeline. HCR is executed in the order $l = 3 \to 2 \to 1 \to 0$. At every level, the optical feature is treated as the spatial reference and the DSM feature is rectified toward it. The refined displacement estimated at level l is transferred to level $l-1$ as a coarse prior, while the aligned shallow representations $\{S_l\}_{l=0}^2$ are retained as decoder skip features. At the bottleneck, the deepest optical feature and aligned DSM feature are projected into $D = 768$-dimensional tokens, supplemented with a learnable two-dimensional positional embedding, and passed through two sequential CC-BSSM blocks. The resulting modality streams are fused and progressively decoded to produce the six-class segmentation map.

3.2 Hierarchical Correspondence Rectification

Figure 2 presents the forward propagation of HCR at level l. Its inputs are the reference optical feature F_l^o, the source DSM feature F_l^d, and the displacement $\mathbf{\Delta}_{l+1}$ estimated at the next deeper level. The deepest module has no preceding estimate and therefore uses a zero prior. HCR proceeds through four ordered stages: prior-guided initialization, local correspondence estimation, residual-offset refinement and differentiable warping, and confidence-aware fusion.

Prior-guided initialization.

For $l < 3$, the deeper displacement is bilinearly resized to the current resolution and its horizontal and vertical components are scaled according to the corresponding spatial-size ratios. At $l = 3$, the prior is initialized to zero. The source DSM feature is first pre-warped with this coarse estimate, so that the subsequent local search only needs to predict a residual correction:

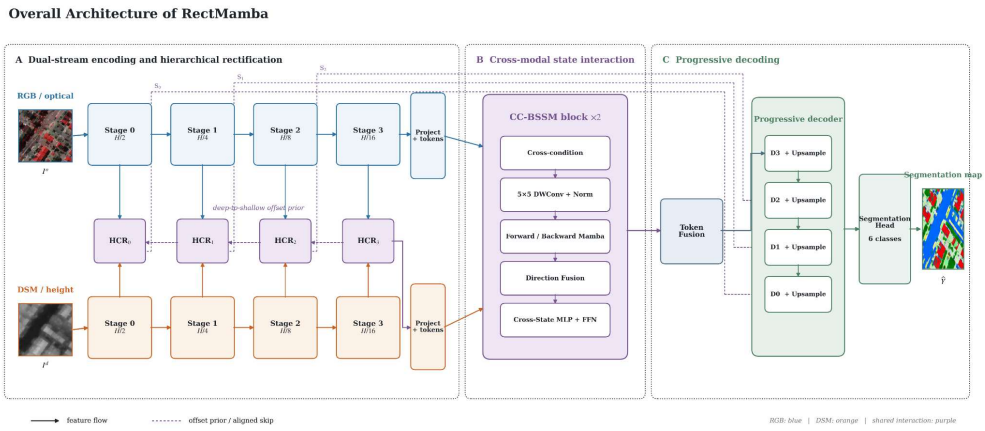


Fig. 1 Overall architecture of RectMamba. The dual-stream encoder extracts optical and DSM features at four scales. HCR progressively rectifies DSM features from deep to shallow, two CC-BSSM blocks perform bottleneck interaction, and the progressive decoder combines the fused tokens with aligned shallow skips to predict the segmentation map.

$$\bar{\boldsymbol{\Delta}}_l = \begin{cases} \mathbf{0}, & l = 3, \\ \mathcal{S}_l\big(\mathrm{Up}(\boldsymbol{\Delta}_{l+1})\big), & l < 3, \end{cases} \qquad F^d_{l,\mathrm{pre}}(\mathbf{p}) = \mathcal{W}\big(F^d_l, \bar{\boldsymbol{\Delta}}_l\big)(\mathbf{p}), \tag{1}$$

where $\mathrm{Up}(\cdot)$ denotes bilinear interpolation, $\mathcal{S}_l(\cdot)$ rescales displacement components, and $\mathcal{W}(\cdot)$ denotes bilinear grid sampling.

Local correspondence estimation.

Separate 1×1 projections map the reference and pre-warped source features to a compact correlation space. Channel-wise ℓ_2 normalization reduces sensitivity to modality-dependent feature magnitude:

$$\mathbf{q}_l(\mathbf{p}) = \frac{\phi^o_l(F^o_l(\mathbf{p}))}{\|\phi^o_l(F^o_l(\mathbf{p}))\|_2 + \varepsilon}, \qquad \mathbf{k}_l(\mathbf{p}) = \frac{\phi^d_l(F^d_{l,\mathrm{pre}}(\mathbf{p}))}{\|\phi^d_l(F^d_{l,\mathrm{pre}}(\mathbf{p}))\|_2 + \varepsilon}. \tag{2}$$

For each position $\mathbf{p}$, HCR evaluates candidate displacements $\boldsymbol{\delta}$ in a local window $\mathcal{D}_l$. The search radii are $(\rho_0, \rho_1, \rho_2, \rho_3) = (1, 1, 2, 3)$, giving $K_l = (2\rho_l + 1)^2$ candidates. Correlations are converted to a probability distribution using temperature $\tau = 0.07$:

$$\begin{aligned} C_l(\mathbf{p}, \boldsymbol{\delta}) &= \mathbf{q}_l(\mathbf{p})^{\top} \mathbf{k}_l(\mathbf{p} + \boldsymbol{\delta}), \\ P_l(\mathbf{p}, \boldsymbol{\delta}) &= \frac{\exp(C_l(\mathbf{p}, \boldsymbol{\delta})/\tau)}{\sum_{\boldsymbol{\delta}' \in \mathcal{D}_l} \exp(C_l(\mathbf{p}, \boldsymbol{\delta}')/\tau)}. \end{aligned} \tag{3}$$

Offset refinement and differentiable warping.

Soft-argmax computes the expected local displacement, yielding a differentiable residual field $\mathbf{e}_l$. Adding it to the coarse prior produces the refined displacement:

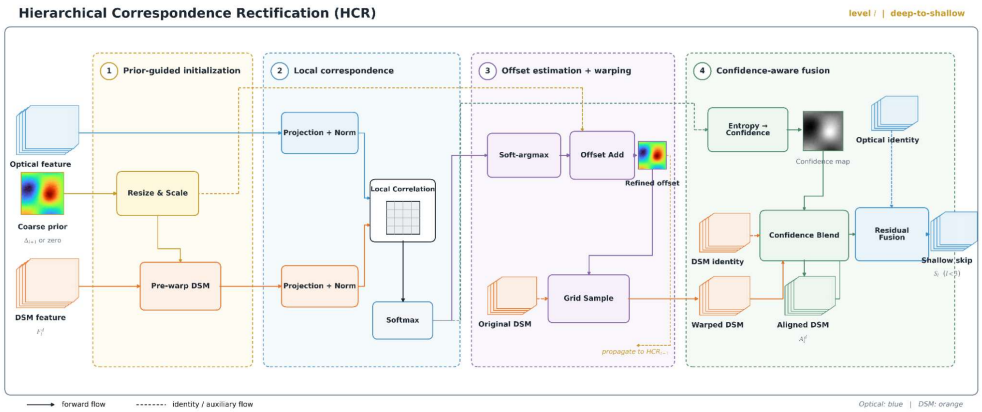


Fig. 2 Forward propagation of HCR at level l. A resized deep-level displacement initializes a coarse pre-warp; normalized local correlation estimates a residual offset; the refined displacement warps the original DSM feature; and entropy-derived confidence controls conservative blending and residual skip fusion.

$$\mathbf{e}_l(\mathbf{p}) = \sum_{\boldsymbol{\delta}\in\mathcal{D}_l} P_l(\mathbf{p},\boldsymbol{\delta})\boldsymbol{\delta}, \qquad \boldsymbol{\Delta}_l(\mathbf{p}) = \bar{\boldsymbol{\Delta}}_l(\mathbf{p}) + \mathbf{e}_l(\mathbf{p}). \tag{4}$$

The refined field is then applied to the original DSM feature rather than to the already pre-warped feature, preventing the coarse displacement from being counted twice:

$$\widetilde{F}_l^d(\mathbf{p}) = \mathcal{W}(F_l^d, \boldsymbol{\Delta}_l)(\mathbf{p}) = \mathrm{BilinearSample}\big(F_l^d, \mathbf{p} + \boldsymbol{\Delta}_l(\mathbf{p})\big)\,. \tag{5}$$

Confidence-aware fusion.

Not every local match is reliable. HCR measures the normalized entropy of P_l and assigns high confidence only to sharply peaked correspondence distributions:

$$H_l(\mathbf{p}) = -\sum_{\boldsymbol{\delta}\in\mathcal{D}_l} P_l(\mathbf{p},\boldsymbol{\delta}) \log P_l(\mathbf{p},\boldsymbol{\delta}), \qquad c_l(\mathbf{p}) = 1 - \frac{H_l(\mathbf{p})}{\log K_l}. \tag{6}$$

The confidence map conservatively blends the warped and unwarped DSM features. Thus, reliable matches use the rectified representation, whereas ambiguous regions retain the original DSM evidence:

$$A_l^d = c_l \odot \widetilde{F}_l^d + (1 - c_l) \odot F_l^d. \tag{7}$$

For $l < 3$, the aligned DSM feature is projected and injected into an optical identity path to construct the shallow decoder skip. The learnable residual strength α_l is initialized to 0.1, so the module initially behaves close to the optical identity mapping:

$$S_l = \psi_l\big(F_l^o + \alpha_l G_l(A_l^d)\big)\,, \qquad l \in \{0,1,2\}. \tag{8}$$

3.3 Cross-Conditioned Bidirectional State-Space Interaction

After HCR, the deepest optical feature F_3^o and aligned DSM feature A_3^d are projected and flattened into token sequences $\mathbf{X}^o, \mathbf{X}^d \in \mathbb{R}^{N\times D}$. For a 256×256 input patch, the 16×16 bottleneck grid contains $N = 256$ tokens, and the embedding dimension is $D = 768$. A learnable two-dimensional positional embedding is added to both streams:

$$\widehat{\mathbf{X}}^o = \mathbf{X}^o + \mathbf{E}_{\text{pos}}, \qquad \widehat{\mathbf{X}}^d = \mathbf{X}^d + \mathbf{E}_{\text{pos}}, \qquad \mathbf{E}_{\text{pos}} \in \mathbb{R}^{N\times D}. \tag{9}$$

As shown in Fig. 3, the final network stacks two identical CC-BSSM blocks. Each block contains four operations: cross-conditioning, local convolutional mixing, bidirectional selective state modeling, and cross-state interaction followed by feed-forward refinement.

Cross-conditioning.

For each modality, the conditioner constructs a relation descriptor from the two normalized streams and their signed difference. It predicts bounded scale and shift terms that modulate the current stream before state-space processing:

$$\mathbf{R}^o = [\mathrm{LN}(\widehat{\mathbf{X}}^o); \mathrm{LN}(\widehat{\mathbf{X}}^d); \mathrm{LN}(\widehat{\mathbf{X}}^d) - \mathrm{LN}(\widehat{\mathbf{X}}^o)], \tag{10a}$$

$$\mathbf{R}^d = [\mathrm{LN}(\widehat{\mathbf{X}}^d); \mathrm{LN}(\widehat{\mathbf{X}}^o); \mathrm{LN}(\widehat{\mathbf{X}}^o) - \mathrm{LN}(\widehat{\mathbf{X}}^d)], \tag{10b}$$

$$[\mathbf{s}^m, \mathbf{b}^m] = 0.1 \tanh(g_m(\mathbf{R}^m)), \qquad \mathbf{Z}^m = \widehat{\mathbf{X}}^m \odot (1 + \mathbf{s}^m) + \mathbf{b}^m, \quad m \in \{o, d\}. \tag{10c}$$

The final layer of each conditioner is zero-initialized. Consequently, $\mathbf{s}^m = \mathbf{b}^m = \mathbf{0}$ at initialization and the modulation begins as an identity transformation. Cross-modal information therefore influences the input-dependent selective dynamics without directly replacing Mamba's internal parameterization.

Local mixing and bidirectional state modeling.

The conditioned tokens are restored to the 16×16 grid, mixed by a 5×5 depthwise convolution, flattened, and normalized:

$$\mathbf{U}^m = \mathrm{LN}(\mathrm{Flatten}(\mathrm{DWConv}_{5\times5}(\mathrm{Reshape}(\mathbf{Z}^m)))), \qquad m \in \{o, d\}. \tag{11}$$

The same modality-specific Mamba operator is applied to the forward raster sequence and its reversal; hence the two directions share parameters within each modality, whereas the optical and DSM operators remain distinct. The reversed output is restored to the original token order before adaptive direction fusion. This design implements two one-dimensional raster directions rather than a four-direction SS2D scan:

$$\mathbf{H}^m_{\rightarrow} = \mathrm{Mamba}_m(\mathbf{U}^m), \qquad \mathbf{H}^m_{\leftarrow} = \mathrm{Rev}(\mathrm{Mamba}_m(\mathrm{Rev}(\mathbf{U}^m))), \tag{12a}$$

$$[\omega^m_{\rightarrow}, \omega^m_{\leftarrow}] = \mathrm{softmax}(a_m([\mathbf{H}^m_{\rightarrow}; \mathbf{H}^m_{\leftarrow}])),$$

$$\mathbf{S}^m = \omega^m_{\rightarrow} \odot \mathbf{H}^m_{\rightarrow} + \omega^m_{\leftarrow} \odot \mathbf{H}^m_{\leftarrow}, \qquad m \in \{o, d\}. \tag{12b}$$

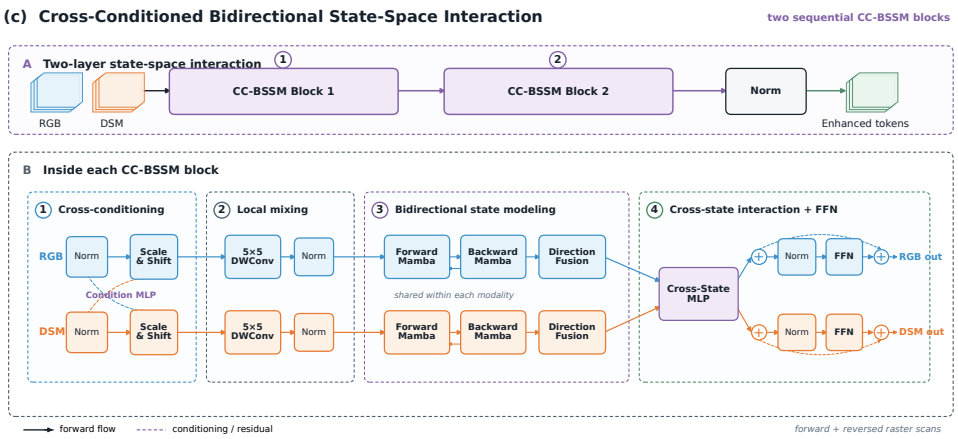


Fig. 3 Cross-Conditioned Bidirectional State-Space Interaction with two sequential CC-BSSM blocks. The other modality conditions tokens before Mamba; forward and reversed raster scans are adaptively combined, followed by additive cross-state interaction and residual feed-forward refinement.

Cross-state interaction and block output.

The direction-fused states form a shared relation descriptor. A cross-state MLP predicts modality-specific residual updates, whose learnable strengths λ_o and λ_d are initialized to 0.1:

$$\mathbf{R}^s = [\mathbf{S}^o; \mathbf{S}^d; \mathbf{S}^d - \mathbf{S}^o], \qquad [\boldsymbol{\delta}^o, \boldsymbol{\delta}^d] = h(\mathbf{R}^s), \qquad \widetilde{\mathbf{S}}^m = \mathbf{S}^m + \lambda_m \boldsymbol{\delta}^m. \tag{13}$$

Each stream is finally updated through a state residual and a modality-specific pre-normalized feed-forward network:

$$\mathbf{Y}^m = \widehat{\mathbf{X}}^m + \widetilde{\mathbf{S}}^m, \qquad \mathbf{X}^m_{\text{out}} = \mathbf{Y}^m + \text{FFN}_m(\text{LN}(\mathbf{Y}^m)). \tag{14}$$

The output pair of the first block becomes the input pair of the second block. This preserves two explicit modality streams throughout state-space interaction. The final configuration uses two CC-BSSM blocks, $d_{\text{state}} = 16$, $d_{\text{conv}} = 4$, and expansion factor 2; the two-block depth is selected by the completed depth experiment.

3.4 Progressive decoder and full-tile reconstruction

The final optical and DSM token streams are concatenated, layer-normalized, and linearly projected back to D channels. An optical residual is retained to stabilize fusion. The token map is reshaped to 16×16, expanded to 512 channels, and progressively upsampled through decoder stages with output channels $(256, 128, 64, 16)$. The first three decoder stages receive HCR-derived shallow skip features. A 3×3 segmentation head predicts six ISPRS classes. The bottleneck fusion is

$$\mathbf{X}_{fuse} = W_f \ \text{LN}\left(\left[\mathbf{X}^o; \mathbf{X}^d\right]\right) + \mathbf{X}^o \tag{15}$$

High-resolution tiles are not resized globally. Training data are pre-cut into synchronized 256×256 optical/DSM/label patches with stride 128 (50% overlap), and

Table 1 Locked dataset partition used by all methods.

Dataset	GSD	Training	Validation	Test	Total
Vaihingen	~ 9 cm	9	3	4	16
Potsdam	~ 5 cm	14	4	6	24

each sample receives synchronized horizontal/vertical flips and 90° rotations. DSM normalization is computed at the full-tile level before patch extraction. During full-tile inference, overlapping windows are processed in mini-batches; class logits are accumulated at their original coordinates and divided by a count map before the final argmax:

$$\bar{\mathbf{z}}(\mathbf{p}) = \frac{\sum_{j:\,\mathbf{p}\in\Omega_j} \mathbf{z}_j(\mathbf{p})}{\sum_{j:\,\mathbf{p}\in\Omega_j} 1}, \qquad \widehat{y}(\mathbf{p}) = \underset{c}{arg\,max}\ \bar{z}_c(\mathbf{p}) \tag{16}$$

3.5 Computational characteristics

The local HCR correlation volume contains $K_l = (2\rho_l + 1)^2$ candidates per spatial position, so its matching cost is $\mathcal{O}(K_l N_l)$ with fixed local radii rather than all-pairs $\mathcal{O}(N_l^2)$ correspondence. The selective SSM bottleneck scales linearly with token length N for a fixed hidden dimension. However, RectMamba also contains convolutional encoders, local correlation, MLPs, and a decoder; therefore the complete network should not be described as strictly $\mathcal{O}(N)$. Section 4.5 complements this asymptotic characterization with measured parameter count, operator-counted GFLOPs, peak GPU memory, throughput, and wall-clock latency.

$$T_{HCR} = \mathcal{O}\left(\sum_{l=0}^{3} K_l N_l\right), \qquad T_{SSM} = \mathcal{O}(ND) \quad \text{for fixed state dimensions} \tag{17}$$

4 Experiments

4.1 Datasets and evaluation protocols

We evaluate RectMamba on the ISPRS Vaihingen and Potsdam 2-D semantic labeling benchmarks. Both datasets provide very-high-resolution optical imagery, DSM data, and six semantic classes: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. To prevent tile leakage, the partition is fixed before training and shared by RectMamba and every re-trained comparator. Table 1 summarizes the locked partition.

The training tiles are cropped into synchronized 256×256 optical/DSM/label patches with stride 128, corresponding to 50% overlap. Random horizontal and vertical flips and random 90° rotations are applied synchronously to all modalities. Validation is performed on the original validation tiles using 256×256 sliding windows with stride 128. Logits in overlapping regions are averaged before the final label decision, and the

Table 2 Implementation settings used in the reported experiments.

Setting	Value
Patch size	256×256
Training stride	128 (50% overlap)
Epochs / samples per epoch	50 / 4,000
Optimizer	AdamW
Initial learning rate / weight decay	2×10^{-4} / 10^{-2}
Warm-up	5 epochs
HCR radii / temperature	$(1, 1, 2, 3)$ / 0.07
CC-BSSM depth	2 blocks (selected by depth ablation)
SSM d_{state} / d_{conv} / expand	16 / 4 / 2

best checkpoint is selected only by validation mIoU. Final testing is performed once on the locked test tiles using the same window size, stride 32, overlap-logit averaging, and the no-boundary/eroded ground truth. All values in the controlled comparison follow this protocol.

Following the project convention, the primary mIoU and mF1 are macro averages over the first five foreground classes; clutter/background is excluded. The six-class mIoU, denoted by mIoU_6, is additionally reported in the ablation study so that behavior on clutter/background remains visible. Overall accuracy (OA), class-wise IoU, F1 score, and Cohen's kappa are computed from the accumulated full-tile confusion matrix.

4.2 Implementation details

The model is implemented in PyTorch and trained on an NVIDIA RTX 4090 GPU. The final configuration uses R50-ViT-B/16-compatible hybrid dimensions: hidden size 768, MLP dimension 3072, three decoder skip connections, HCR search radii $(1, 1, 2, 3)$, HCR correlation dimensions $(24, 32, 48, 64)$, temperature $\tau = 0.07$, two CC-BSSM blocks, $d_{\text{state}} = 16$, $d_{\text{conv}} = 4$, and expansion factor 2. The optical and DSM encoders are initialized from the same R50-compatible pretrained model and then optimized independently. The single-channel DSM stem is initialized by averaging the optical stem kernel over its input-channel dimension.

The current training script uses AdamW with learning rate 2×10^{-4}, weight decay 10^{-2}, five warm-up epochs followed by cosine decay, gradient clipping at 1.0, and 50 total epochs. Each epoch samples 4,000 patches from the pre-cut training set unless otherwise noted. Batch size is selected according to available GPU memory and kept identical across ablations. All four ablation variants use the same seed, data pipeline, optimizer, scheduler, validation protocol, and checkpoint rule; only HCR and CC-BSSM are toggled.

4.3 Comparison with state-of-the-art methods

We compare RectMamba with two RGB-only models, TransUNet [11] and GLOTS [12], and three representative optical–DSM competitors, FuseNet [6], SA-Gate [9], and CMGFNet [10]. All competitors are re-trained using the locked split and evaluated by the same full-tile reconstruction protocol; the entries are therefore directly comparable rather than copied from papers using different partitions or boundary conventions. Tables 3 and 4 group the methods by input modality and identify their original publication venues. In addition to OA, mF1, and mIoU, the tables retain the five foreground class IoUs so that improvements cannot be attributed to one dominant category.

Table 3 Quantitative comparison on the ISPRS Vaihingen test set. Class columns report IoU (%); the best result in each column is in bold.

Input	Method	Pub.	Imp.	Bldg.	Low	Tree	Car	OA	mF1	mIoU
RGB	TransUNet	arXiv 2021	82.19	89.14	65.45	82.45	71.29	89.63	87.44	78.10
	GLOTS	TGRS 2023	78.98	86.26	61.65	80.26	50.41	87.66	82.65	71.51
RGB+DSM	FuseNet	ACCV 2016	74.07	81.91	53.97	76.79	53.36	84.81	80.34	68.02
	SA-Gate	ECCV 2020	75.18	82.76	47.22	71.96	37.66	83.12	75.79	62.95
	CMGFNet	ISPRS J. 2022	80.99	88.05	64.86	81.75	79.47	89.10	88.07	79.02
	RectMamba (ours)	–	**84.30**	**91.05**	**68.63**	**84.16**	**80.17**	**90.95**	**89.72**	**81.66**

Table 4 Quantitative comparison on the ISPRS Potsdam test set. Class columns report IoU (%); the best result in each column is in bold.

Input	Method	Pub.	Imp.	Bldg.	Low	Tree	Car	OA	mF1	mIoU
RGB	TransUNet	arXiv 2021	81.57	89.36	69.07	65.98	88.44	86.78	87.86	78.88
	GLOTS	TGRS 2023	82.51	89.60	70.78	66.95	87.95	87.53	88.32	79.56
RGB+DSM	FuseNet	ACCV 2016	77.71	91.45	62.84	67.80	82.29	84.70	86.25	76.42
	SA-Gate	ECCV 2020	65.13	83.08	45.74	47.29	60.87	73.99	74.46	60.42
	CMGFNet	ISPRS J. 2022	85.27	**95.22**	72.96	75.21	90.92	89.85	91.01	83.91
	RectMamba (ours)	–	**87.26**	94.23	**75.46**	**77.32**	**93.59**	**90.88**	**92.03**	**85.57**

RectMamba obtains the highest primary mIoU on both benchmarks. On Vaihingen, it reaches 81.66% mIoU, 89.72% mF1, and 90.95% OA, improving over the strongest comparator, CMGFNet, by 2.63, 1.65, and 1.85 percentage points (pp), respectively. Its gains are distributed across all five foreground classes, with particularly clear improvements for impervious surfaces, buildings, and low vegetation. On Potsdam, RectMamba achieves 85.57% mIoU, 92.03% mF1, and 90.88% OA, exceeding CMGFNet by 1.66, 1.02, and 1.03 pp. CMGFNet retains the highest building IoU (95.22%), whereas RectMamba is best on impervious surfaces, low vegetation, trees, and cars. This pattern suggests that hierarchical geometric correction is especially helpful for boundary-sensitive and small-object categories, although building performance remains close to the strongest competing result.

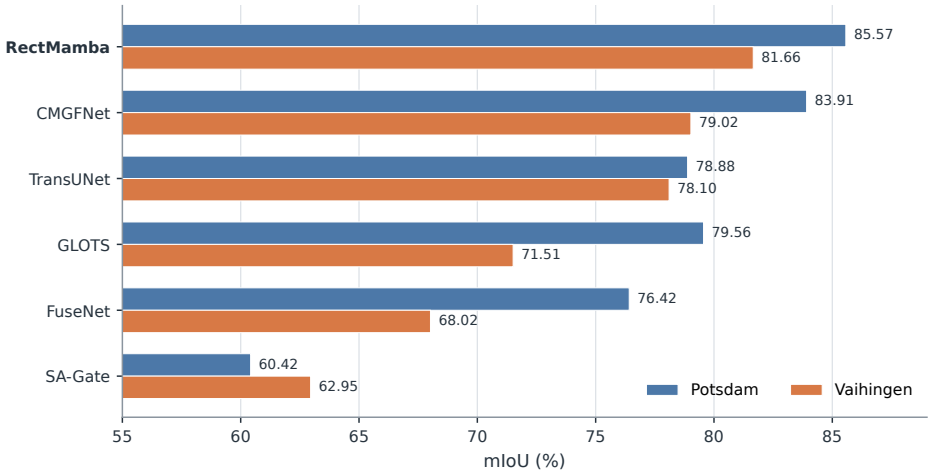


Fig. 4 Primary mIoU comparison under the unified testing protocol. Each bar is the mean over the five foreground classes; clutter/background is excluded.

Figure 4 provides a compact cross-dataset view. RectMamba ranks first on both datasets, while the larger margin on Vaihingen indicates that the benefit is not tied to one city or ground-sampling distance.

4.4 Ablation study

The core ablation uses the same RectMamba-v2 scaffold and training pipeline while toggling only the two proposed components. "Baseline" disables HCR and sets the state-block depth to zero; "HCR only" keeps rectification but removes CC-BSSM; "CC-BSSM only" removes HCR but retains the state blocks; and "Full" activates both. This design isolates the contribution of alignment and state-space interaction without changing the data loader, optimizer, decoder, or checkpoint-selection rule.

Both proposed components provide substantial improvements over the simple dual-stream baseline. HCR alone raises mIoU from 75.67% to 85.14%, confirming that explicit spatial compatibility is important before fusion. CC-BSSM alone reaches 85.43%, demonstrating the value of cross-conditioned long-range interaction even without explicit rectification. Combining the modules produces the best result, 85.57% mIoU. Relative to HCR only, adding CC-BSSM yields 0.43 pp; relative to CC-BSSM only, adding HCR yields 0.14 pp. The modest incremental gain over the stronger single-module variant, compared with the large gain over the baseline, indicates partially overlapping benefits rather than purely additive effects.

4.4.1 HCR and CC-BSSM design analysis

Table 6 examines the operations inside the two modules. Within HCR, removing entropy-derived confidence reduces mIoU by 0.31 pp, while removing coarse-to-fine offset propagation reduces it by 0.17 pp. Confidence estimation also improves mIoU_6, suggesting that uncertain matches disproportionately affect the clutter/background class. Within CC-BSSM, cross-conditioning is the most influential tested operation:

Table 5 Core component ablation on Potsdam. All variants share the same training and testing protocol.

Variant	HCR	CC-BSSM	OA	mF1	mIoU	$mIoU_6$	Kappa
Baseline	✗	✗	84.91	85.85	75.67	63.06	0.7984
HCR only	✓	✗	90.49	91.78	85.14	77.67	0.8742
CC-BSSM only	✗	✓	90.82	91.95	85.43	78.89	0.8789
Full	✓	✓	**90.88**	**92.03**	**85.57**	**79.35**	**0.8797**

Table 6 Internal module ablation on Potsdam. The full model is repeated within each group for direct comparison.

Module	Configuration	OA	mF1	mIoU	$mIoU_6$	Kappa
HCR	w/o confidence estimation	90.77	91.85	85.27	79.31	0.8784
	w/o coarse-to-fine prior	90.75	91.93	85.40	78.76	0.8780
	Full HCR	**90.88**	**92.03**	**85.57**	**79.35**	**0.8797**
CC-BSSM	w/o cross-conditioning	90.57	91.73	85.04	78.78	0.8759
	Unidirectional scan	90.76	91.83	85.23	79.04	0.8782
	Fixed direction fusion	90.44	91.81	85.19	77.80	0.8739
	Full CC-BSSM	**90.88**	**92.03**	**85.57**	**79.35**	**0.8797**

deleting it causes a 0.54 pp mIoU loss. Replacing bidirectional scanning with a unidirectional scan costs 0.34 pp, and replacing adaptive direction fusion with fixed fusion costs 0.39 pp. These observations support the intended forward process: relational information conditions state evolution, complementary scan directions reduce raster-order bias, and learned aggregation selects their contributions.

4.4.2 Depth sensitivity

The number of stacked CC-BSSM blocks is varied while retaining the same HCR and decoder. As shown in Table 7 and Fig. 5, two blocks achieve the best mIoU. One block improves only slightly over HCR alone, whereas four blocks do not provide further benefit. The two-block model exceeds the one- and four-block configurations by 0.32 pp in both cases, suggesting that moderate state-space depth is sufficient at the bottleneck and that deeper interaction can introduce redundant transformation.

4.5 Efficiency and model-size analysis

Table 8 evaluates accuracy–efficiency trade-offs at an input size of 256×256, batch size 1, and FP32 precision on the same NVIDIA RTX 4090 GPU. The first panel distinguishes the RGB-only methods from the RGB+DSM methods, while the second panel isolates the incremental cost of the two RectMamba components. We report total parameters, FVCore-counted GFLOPs, peak allocated GPU memory,

Table 7 Sensitivity to the number of CC-BSSM blocks on Potsdam.

Blocks	OA	mF1	mIoU	Kappa
0 (HCR only)	90.49	91.78	85.14	0.8742
1	90.70	91.84	85.25	0.8776
2	**90.88**	**92.03**	**85.57**	**0.8797**
4	90.75	91.84	85.25	0.8780

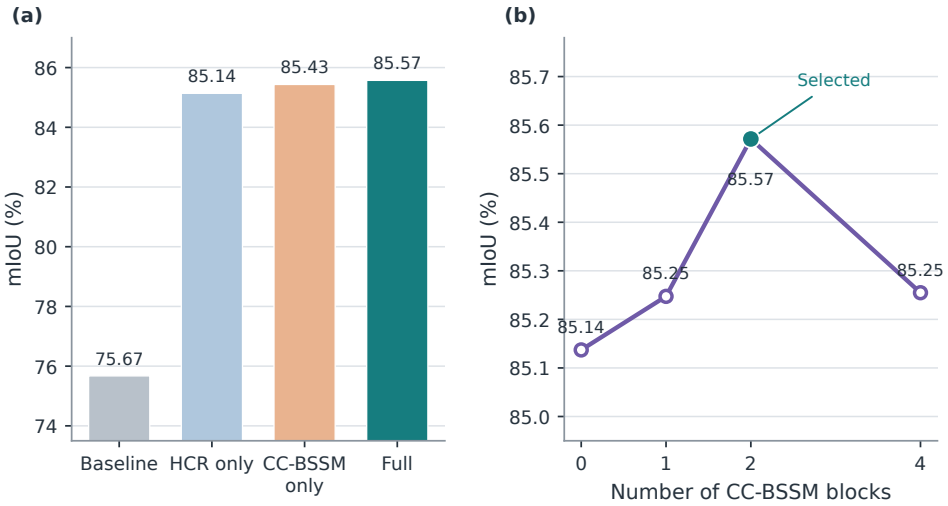


Fig. 5 Ablation summary on Potsdam: (a) contribution of HCR and CC-BSSM; (b) mIoU sensitivity to CC-BSSM depth. The dashed line marks the selected two-block configuration.

mean patch latency, throughput, and Potsdam mIoU. Because the released implementations require different validated PyTorch/CUDA environments, the timing results should be interpreted as hardware-controlled implementation measurements rather than kernel-normalized theoretical comparisons.

RectMamba attains the highest mIoU in panel (a). Compared with CMGFNet, the strongest multimodal accuracy baseline, it improves mIoU by 1.66 percentage points, but increases the parameter count from 43.68 M to 103.65 M, peak allocated memory from 324.4 to 591.1 MB, and latency from 16.95 to 33.14 ms. RectMamba is therefore an accuracy-oriented model rather than a globally lightweight one. Relative to TransUNet, RectMamba uses 1.59% fewer parameters and has a similar reported operator count, while improving mIoU by 6.69 points; nevertheless, its dual-stream processing produces 36.79% higher measured latency. Relative to GLOTS, it reduces parameters by 32.57% and peak memory by 25.79%, but remains 19.29% slower in wall-clock inference. The mismatch between reported GFLOPs and latency reinforces the need to report both: fused kernels, unsupported operations, memory traffic, and framework-specific execution overhead are not captured by the partial operator count.

Table 8 Accuracy–efficiency comparison on Potsdam and cost decomposition of RectMamba. All measurements use a 256 × 256 input, batch size 1, FP32 precision, and an NVIDIA RTX 4090.

Input	Method	Params (M)	GFLOPs[a]	Mem. (MB)	Lat. (ms)	Patch/s	mIoU
(a) Comparison with re-trained segmentation models							
RGB	TransUNet	105.32	38.57	546.8	24.23	41.28	78.88
	GLOTS	153.71	98.09	796.5	27.78	36.00	79.56
RGB+DSM	FuseNet	44.17	60.27	368.5	**12.88**	**77.66**	76.42
	CMGFNet	**43.68**	**10.34**	**324.4**	16.95	58.99	83.91
	SA-Gate	110.85	38.74	691.9	35.45	28.21	60.42
	RectMamba (ours)	103.65	37.33	591.1	33.14	30.18	**85.57**
(b) RectMamba component-level cost breakdown							
RGB+DSM	Baseline	35.00	22.43	**310.6**	**20.67**	**48.37**	75.67
	HCR only	35.00	22.65	323.3	25.77	38.80	85.14
	CC-BSSM only	103.65	37.11	577.4	28.09	35.60	85.43
	Full, 2 blocks	103.65	37.33	591.1	33.14	30.18	**85.57**

[a] FVCore produced partial operator counts for every implementation. Unsupported operators are omitted; in particular, the fused Mamba selective-scan kernel is not included. GFLOPs should therefore be read as implementation-dependent lower bounds, not exact cross-architecture complexity.
Peak memory denotes maximum allocated memory during inference. Latency is the measured mean per patch; Patch/s is its reciprocal throughput. Bold values mark the best result within each panel and efficiency criterion, except that panel (b) emphasizes the full model only for accuracy.

Panel (b) clarifies where RectMamba spends its computation. Activating HCR increases latency from 20.67 to 25.77 ms and peak memory by only 12.7 MB, while improving mIoU by 9.47 points. Adding the two CC-BSSM blocks accounts for most of the parameter and memory increase: the CC-BSSM-only configuration uses 103.65 M parameters and 577.4 MB, compared with 35.00 M and 310.6 MB for the baseline. The full model adds geometric rectification to this state-space configuration, increasing latency by 5.05 ms and memory by 13.7 MB over CC-BSSM only for a further 0.14-point mIoU gain. Thus, HCR provides a favorable standalone accuracy–cost trade-off, whereas CC-BSSM supplies the highest final accuracy at substantially greater model capacity.

4.6 Qualitative analysis

Figures 6 and 7 present qualitative comparisons on representative regions from the Vaihingen and Potsdam test sets, respectively. Each figure includes the RGB image, ground truth, predictions from the internal baseline and five competing methods, and the result of RectMamba. To facilitate spatially consistent comparison, boxes 1–3 indicate identical locations in all panels. The selected regions contain challenging building boundaries, tree–low-vegetation transitions, narrow impervious surfaces, and small vehicles. These cases complement the quantitative evaluation by illustrating differences in boundary preservation, regional consistency, and small-object recognition.

On the Vaihingen example in Fig. 6, the internal baseline produces extensive class confusion and fragmented regions, indicating that direct dual-stream encoding alone is insufficient for reliable optical–DSM fusion. In box 1, where a building is partially surrounded by dense tree crowns, several competing methods either erode the building region or introduce irregular vegetation predictions along its boundary. RectMamba

preserves a more compact building footprint while maintaining a clearer separation from the adjacent trees. Box 2 contains another building embedded in heterogeneous vegetation. Although the competing multimodal models recover its approximate location, their predictions exhibit different degrees of boundary expansion, erosion, and local low-vegetation leakage. The RectMamba result is visually closer to the ground-truth geometry and retains a more coherent surrounding impervious region. In box 3, the dense tree canopy is adjacent to rooftops and narrow ground surfaces. RectMamba better maintains the continuity of the tree region and reduces scattered confusion among trees, low vegetation, and buildings. These observations suggest that hierarchical rectification is beneficial where optical appearance and height discontinuities must be reconciled at object boundaries.

The Potsdam comparison in Fig. 7 further demonstrates the behavior of the models in a denser urban scene. Box 1 contains closely spaced buildings and small intervening regions. The baseline and several competing approaches generate fragmented building interiors, irregular contours, or local class leakage, whereas RectMamba produces more coherent building structures and preserves a greater proportion of the small surrounding objects. Box 2 focuses on regularly distributed roadside trees and the adjacent low-vegetation strip. SA-Gate shows substantial tree-to-low-vegetation confusion in this area, while FuseNet also produces visibly noisy boundaries. RectMamba retains more distinct tree instances and a clearer separation among the tree crowns, low vegetation, and impervious road surface. In box 3, buildings, impervious surfaces, vegetation, and small vehicles occur within a narrow spatial region. Some methods smooth out the small car regions or distort the nearby building boundaries. RectMamba preserves more of the small yellow car predictions while maintaining comparatively regular building and road contours.

Overall, the two examples show that RectMamba does not merely improve large homogeneous regions; its main qualitative advantages are observed at cross-class boundaries, in spatially fragmented urban structures, and around small objects. This behavior is consistent with the intended roles of the proposed modules: HCR improves local correspondence between optical and DSM features before fusion, while CC-BSSM propagates complementary contextual information across spatially separated regions. Nevertheless, residual errors remain at ambiguous tree–low-vegetation transitions and in extremely small clutter regions. These failure cases are consistent with the class-wise quantitative results and indicate that fine-grained vegetation discrimination and rare-class recognition remain directions for further improvement.

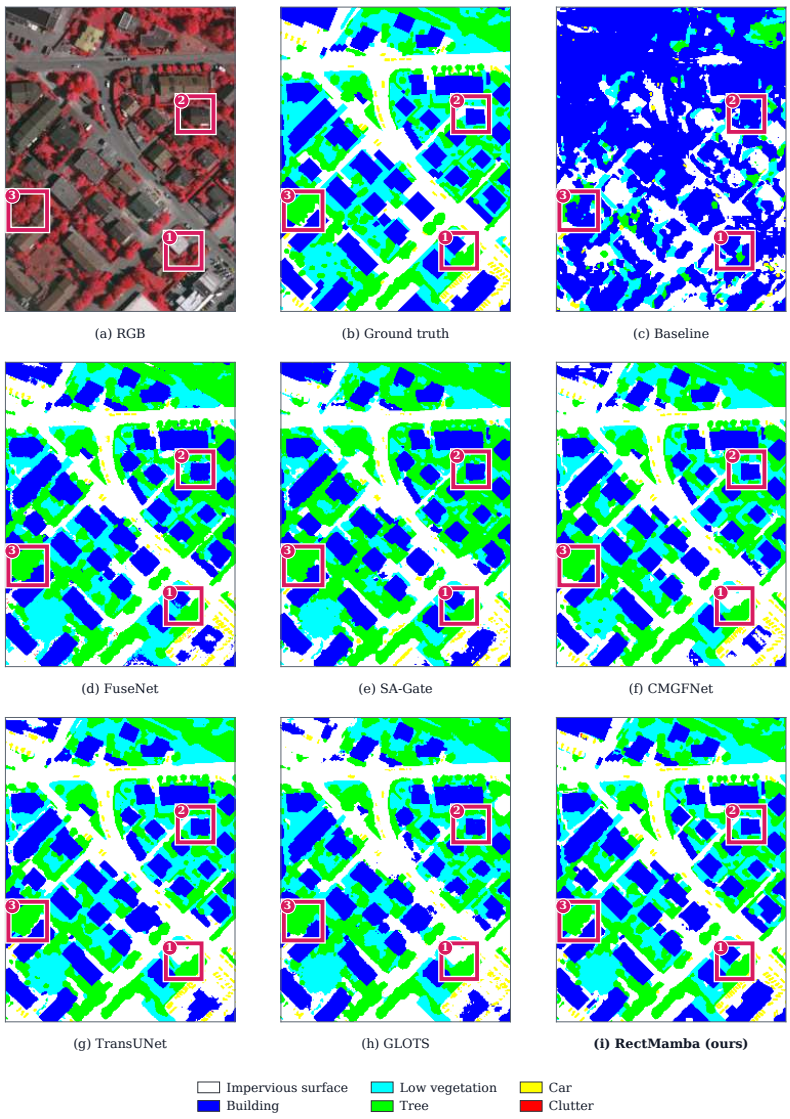


Fig. 6 Qualitative comparison on a representative Vaihingen test region: (a) RGB image, (b) ground truth, (c) baseline, (d) FuseNet, (e) SA-Gate, (f) CMGFNet, (g) TransUNet, (h) GLOTS, and (i) RectMamba. Boxes 1–3 mark identical spatial locations in every panel and highlight building–vegetation boundaries and dense tree-canopy regions.

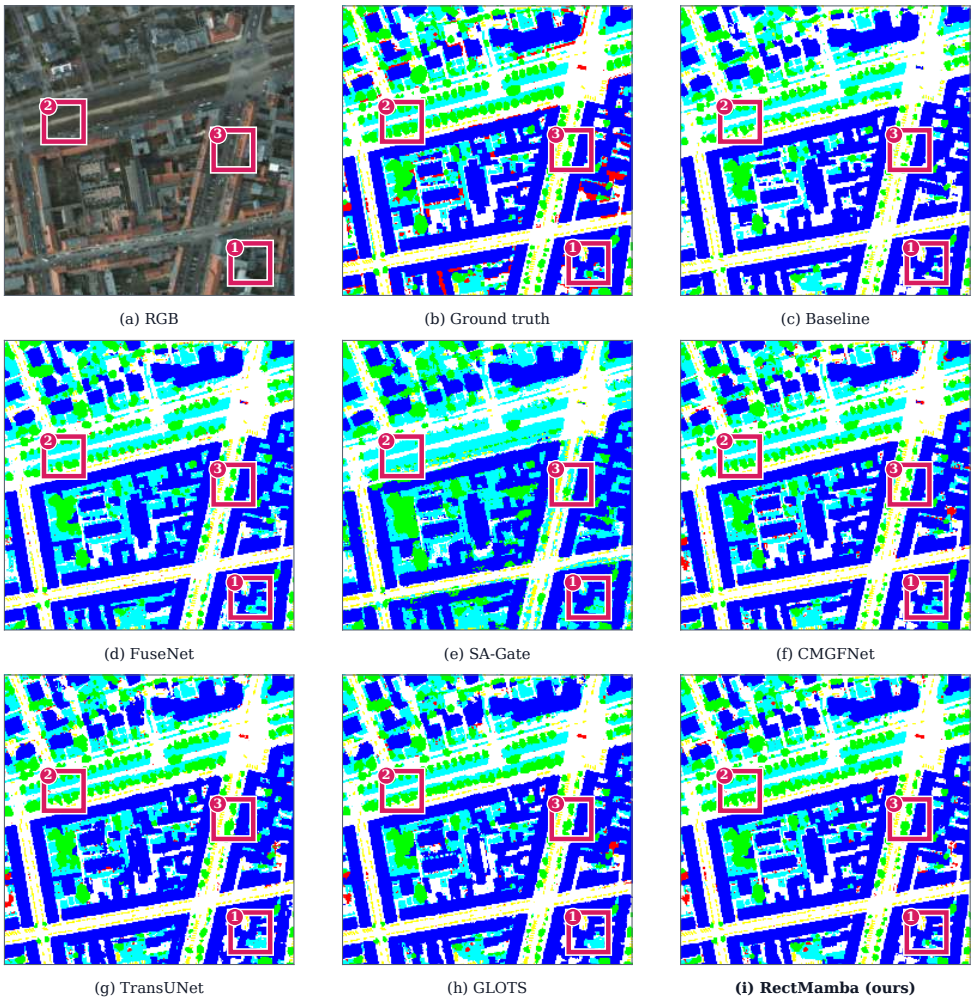


Fig. 7 Qualitative comparison on a representative Potsdam test region: (a) RGB image, (b) ground truth, (c) baseline, (d) FuseNet, (e) SA-Gate, (f) CMGFNet, (g) TransUNet, (h) GLOTS, and (i) RectMamba. Boxes 1–3 mark identical spatial locations in every panel and emphasize dense buildings and small objects, roadside tree–low-vegetation transitions, and building/parking/car boundaries.

5 Discussion

The controlled comparison demonstrates that RectMamba improves over both RGB-only and RGB–DSM alternatives rather than benefiting from a weak single baseline. On Vaihingen, it exceeds the best RGB-only result by 3.56 pp mIoU and the strongest multimodal competitor, CMGFNet, by 2.63 pp. On Potsdam, the corresponding gains are 6.01 and 1.66 pp. Relative to CMGFNet, RectMamba improves all five foreground categories on Vaihingen and four of five categories on Potsdam; only the Potsdam building IoU is 0.99 pp lower. The consistent cross-dataset gains support the intended division of labor in RectMamba: independent encoders retain intramodal appearance and height-specific representations, HCR improves spatial compatibility, and CC-BSSM captures complementary long-range relations without prematurely collapsing the two streams.

RectMamba is intentionally structured as alignment before interaction. The controlled results support both stages: HCR and CC-BSSM independently improve the baseline by 9.47 and 9.76 pp mIoU, while the full model reaches 85.57%. The remaining 0.14 pp gain obtained by adding HCR to the stronger CC-BSSM-only model is positive but small, indicating that both modules resolve some common cross-modal incompatibilities. The internal ablations are more diagnostic: confidence-aware rectification, coarse-to-fine propagation, cross-conditioning, bidirectional scanning, and adaptive direction fusion each contribute to the final result.

The measured efficiency results qualify the architectural motivation. The two-block RectMamba configuration contains 103.65 M parameters, requires 591.1 MB peak allocated memory, and processes a patch in 33.14 ms on the RTX 4090. It is more parameter- and latency-intensive than CMGFNet and FuseNet, although it delivers the highest mIoU. At the same time, it uses fewer parameters and less memory than GLOTS, and fewer resources with lower latency than SA-Gate. Consequently, the appropriate claim is not that RectMamba is globally lightweight, but that its selective SSM bottleneck avoids quadratic all-pairs attention while offering a measurable accuracy–resource trade-off within a complete dual-stream architecture.

Class-wise results show that vegetation remains more difficult than buildings and cars. On Potsdam, low vegetation and trees reach 75.46% and 77.32% IoU, well below buildings (94.23%) and cars (93.59%). On Vaihingen, low vegetation is again the weakest foreground class at 68.63%. Similar spectral appearance, variable canopy height, and mixed boundary pixels remain plausible sources of confusion. Consequently, the two qualitative figures should include vegetation transitions rather than showing only visually easy building examples.

6 Conclusion

This paper presents RectMamba, a multimodal remote sensing segmentation framework that combines explicit correspondence-guided geometric rectification with cross-conditioned bidirectional state-space interaction. HCR estimates dense local displacements through normalized correlation and soft-argmax, refines them hierarchically, and uses entropy-derived confidence to limit unreliable fusion. CC-BSSM conditions each modality on the other before Mamba, adaptively combines forward and

reverse raster scans, and exchanges information through residual cross-state interaction. Under a unified full-tile protocol, RectMamba achieves 85.57% mIoU on Potsdam and 81.66% on Vaihingen, improving over the strongest re-trained comparator by 1.66 and 2.63 pp. The core ablation shows that HCR and CC-BSSM both substantially improve the baseline, while their combination gives the best result; the internal ablations further validate confidence estimation, coarse-to-fine propagation, cross-conditioning, and adaptive bidirectional scanning. Two CC-BSSM blocks provide the highest accuracy among the evaluated depths. The measured efficiency analysis positions RectMamba as an accuracy-oriented model: it uses fewer parameters and less peak memory than GLOTS and SA-Gate, but incurs higher latency and memory than compact multimodal CNN baselines. Future work will investigate more compact state interaction and stronger discrimination between low vegetation and trees.

Declarations

Funding: Innovation Capability Improvement Plan Project of Hebei Province (22567626H)

Competing interests: The authors declare no competing interests.

Author contributions: Jiaqi Xu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, and Writing—original draft. Shengtao Sun: Supervision, Project administration, Resources, and Writing—review and editing. All authors reviewed and approved the final manuscript.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability: The datasets analyzed in this study are publicly available through the ISPRS 2D Semantic Labeling Benchmark. The Vaihingen dataset is available at https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx, and the Potsdam dataset is available at https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx. Access to the datasets is subject to the terms and conditions specified by the data provider. No new datasets were generated in this study.

Materials availability: Not applicable.

Code availability: https://github.com/nxu6326-boop/RectMamba

References

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 3431–3440.

[2] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015. p. 234–241.

[3] Wang L, Li R, Zhang C, Fang S, Duan C, Meng X, Atkinson PM. UNetFormer: a UNet-like Transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J Photogramm Remote Sens. 2022;190:196–214.

[4] Ghamisi P, Rasti B, Yokoya N, Wang Q, Höfle B, Bruzzone L, et al. Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art. IEEE Geosci Remote Sens Mag. 2019;7:6–39.

[5] Li J, et al. Deep learning in multimodal remote sensing data fusion: a comprehensive review. Int J Appl Earth Obs Geoinf. 2022;112:102926.

[6] Hazirbas C, Ma L, Domokos C, Cremers D. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Proceedings of the Asian Conference on Computer Vision; 2016. p. 213–228.

[7] Audebert N, Le Saux B, Lefèvre S. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS J Photogramm Remote Sens. 2018;140:20–32.

[8] Ma X, Zhang X, Pun M-O. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. IEEE J Sel Topics Appl Earth Observ Remote Sens. 2022;15:3463–3474.

[9] Chen X, Lin K-Y, Wang J, Wu W, Qian C, Li H, Zeng G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In: Proceedings of the European Conference on Computer Vision; 2020. p. 561–577.

[10] Hosseinpour H, Samadzadegan F, Javan FD. CMGFNet: a deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J Photogramm Remote Sens. 2022;184:96–115.

[11] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306; 2021.

[12] Liu Y, Zhang Y, Wang Y, Mei S. Rethinking Transformers for semantic segmentation of remote sensing images. IEEE Trans Geosci Remote Sens. 2023;61:1–15. Art. no. 5617515. https://doi.org/10.1109/TGRS.2023.3302024.

[13] Zhang J, Liu H, Yang K, Hu X, Liu R, Stiefelhagen R. CMX: cross-modal fusion for RGB-X semantic segmentation with Transformers. IEEE Trans Intell Transp Syst. 2023;24(12):14679–14694.

[14] Liu Z, et al. Swin Transformer: hierarchical vision Transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 10012–10022.

[15] Ma X, Zhang X, Pun M-O, Liu M. A multilevel multimodal fusion Transformer for remote sensing semantic segmentation. IEEE Trans Geosci Remote Sens. 2024;62:5403215.

[16] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. arXiv:2312.00752; 2023.

[17] Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, et al. VMamba: visual state space model. Adv Neural Inf Process Syst. 2024;37:103031–103063.

[18] Ma X, Zhang X, Pun M-O. RS3Mamba: visual state space model for remote sensing image semantic segmentation. IEEE Geosci Remote Sens Lett. 2024;21:6011405.

[19] Wan Z, Zhang P, Wang Y, Yong S, Stepputtis S, Sycara K, Xie Y. Sigma: Siamese Mamba network for multi-modal semantic segmentation. In: Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision; 2025.

[20] Yan L, Feng Q, Wang J, Cao J, Feng X, Tang X. A multilevel multimodal hybrid Mamba-large strip convolution network for remote sensing semantic segmentation. Remote Sens. 2025;17:2696. https://doi.org/10.3390/rs17152696.

[21] Xiao S, Wang P, Diao W, Fu K, Sun X. A multimodal semantic segmentation framework for heterogeneous optical and complex SAR data. IEEE J Sel Topics Appl Earth Observ Remote Sens. 2025;18:8083–8098. https://doi.org/10.1109/JSTARS.2025.3542487.

[22] Li Z, Zhang Y, Hu J, Wu K, Miao J, Wu J, Zhang X, Yang B, Xia Z. HDC-Net: a multimodal remote sensing semantic segmentation network with hierarchical dual-stream fusion and cross-token interaction. Knowl Based Syst. 2026;337:115416. https://doi.org/10.1016/j.knosys.2026.115416.