scientific reports

Check for updates

OPEN

Projection Kernel regularization for diffusion-based multimodal remote sensing segmentation

Xu Tong[1], Fan Yang[2], Qiang Yang[2]✉, Huajun Wang[1]✉, Sen Wang[3]✉, Tao Zhang[2], Wuxueying Qiu[4] & Shujun Liu[5]

With the increasing availability of multimodal remote sensing (RS) data, semantic segmentation that leverages complementary information from true orthophotos (TOP) and digital surface models (DSM) has become essential for urban analysis. Diffusion-based segmentation provides an effective iterative refinement mechanism for modeling complex multimodal distributions; however, conventional pixel-wise supervision emphasizes local accuracy while overlooking global distribution alignment, often leading to inconsistent predictions and blurred object boundaries. Although maximum mean discrepancy (MMD) measures global statistical differences between predicted and ground-truth distributions, its effectiveness in high-dimensional class-probability spaces is limited by directional cancellation effects that reduce sensitivity to complex distribution shifts. To address this issue, we propose a projection-kernel regularized diffusion-based multimodal RS segmentation framework that enforces global statistical alignment through distribution-level regularization rather than modifying the intrinsic diffusion process. The proposed regularization performs multi-directional projections of high-dimensional class-probability vectors onto one-dimensional subspaces and derives a closed-form kernel integration to avoid numerical sampling across projection directions, enabling efficient and stable global distribution matching. In addition, a Cross-Attention Dual-Encoder Fusion (CADEF) module is introduced to alleviate geometry–texture misalignment, and a Hierarchical EMA-Gated Recursive Denoising (HERD) mechanism is designed to stabilize multiscale feature refinement. Experiments on the ISPRS Vaihingen and Potsdam benchmarks demonstrate that the proposed regularization consistently improves segmentation accuracy over state-of-the-art CNN-, Transformer-, and diffusion-based baselines, yielding enhanced global consistency and sharper boundary delineation. Code is available at: https://github.com/tonyy127/PKDiff

Keywords Projection, Diffusion models, Multimodal remote sensing, Image segmentation, Multimodal fusion

Recent advances in sensing technology have made it increasingly feasible to acquire multimodal RS data—including optical, multispectral, hyperspectral imagery, synthetic aperture radar (SAR), and light detection and ranging (LiDAR). In urban planning[1], land-use monitoring[2], environmental protection[3], and disaster response[4], such heterogeneous sources offer complementary evidence that can substantially improve semantic segmentation. In particular, combining TOP and DSM is attractive: TOP provides rich spectral and textural cues, whereas DSM contributes geometric and structural priors. Yet, the inherent distribution shift across modalities and the difficulties of cross modal alignment pose enduring challenges for building models that are both globally consistent and locally precise[5–7].

Deep learning has propelled RS segmentation forward, but key limitations remain. Convolutional neural networks (CNNs) excel at hierarchical local feature extraction but are constrained in long-range dependency modeling, which can hinder global coherence in complex scenes. In contrast, Transformer-based architectures alleviate this limitation by leveraging self-attention to capture global interactions; however, their computational cost is non-trivial, and adapting them to multimodal RS scenarios requires carefully designed fusion mechanisms to avoid over-smoothing or modality dominance[8–12]. A line of multimodal methods proposes cross-scale or

[1]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu 610059, China. [2]School of Computer Science and Technology, Yibin University, Yibin 644007, China. [3]Shanghai Shentong Metro Group Co., Ltd, Shanghai 201103, China. [4]College of Geophysics, Chengdu University of Technology, Chengdu 610059, China. [5]School of Computer Engineering, Chengdu Technological University, Chengdu 611730, China. ✉email: scyangqiang@163.com; wanghuajun@cdut.edu.cn; shentong_wangsen@163.com

cross-layer fusion to strengthen complementary cues among modalities, yet existing designs may still struggle to balance global structure and fine boundary details in dense prediction[11,13,14].

Compared to convolutional neural networks, which depend on local receptive fields, and Transformers, which are sensitive to shifts in data distributions, diffusion models[15] provide an effective iterative probabilistic refinement mechanism for modeling complex multimodal remote sensing data and preserving global structural information[16–19]. In semantic segmentation, their role is better understood as structured prediction refinement rather than strict likelihood-based generative modeling, particularly when discrete label distributions are involved.

However, conventional diffusion models exhibit notable shortcomings in semantic segmentation. First, their denoising process lacks appropriate distribution-level supervision, relying solely on a single probability flow path to generate feature representations, which struggles to fully capture fine-grained distributional details, resulting in deviations between predicted and ground-truth distributions. Second, diffusion models may fail to effectively handle the complexity of high-dimensional distributions in multimodal RS data, particularly in scenarios requiring precise boundary preservation and multimodal feature fusion.

To address the challenge of distribution alignment, conventional pointwise loss functions, such as mean squared error (MSE) and cross-entropy (CE), are widely employed to optimize pixel-level errors. However, MSE is sensitive to outliers and struggles to capture the overall structure of distributions; CE tends to prioritize dominant classes, lacking global perspective, and performs inadequately in high-dimensional or multimodal RS data. To this end, maximum mean discrepancy (MMD)[20], a kernel-based nonparametric distribution metric, has been introduced to quantify the overall differences between two distributions by comparing their mean embeddings, offering flexibility without requiring explicit distributional assumptions. Nevertheless, in high-dimensional or multimodal scenarios, MMD is limited by the "cancellation effect" in mean embeddings, where discrepancies in different directions may be diminished or partially offset, thereby reducing sensitivity to variations in location, scale, and covariance[21,22].

To overcome these limitations, we propose the Projection Kernel Discrepancy Regularizer (PKD-R) module, which achieves more stable and consistent distribution alignment by decomposing high-dimensional distributions into multiple one-dimensional subspaces and integrating kernel discrepancies. PKD-R projects high-dimensional class probability vectors onto multiple one-dimensional subspaces, computes one-dimensional Gaussian kernel MMD, and aggregates multidirectional discrepancies to form a stable global distribution metric, significantly enhancing sensitivity to complex distributional variations, including location, scale, and covariance changes. By enforcing global distributional alignment between predictions and labels, PKD-R improves the empirical stability and consistency of diffusion-based segmentation performance on standard multimodal RS benchmarks.

Building on this observation, we introduce a projection-kernel–regularized diffusion-based segmentation framework, termed PKDiff, in which PKD-R serves as the central distribution-level regularizer, supported by CADEF and HERD for multimodal fusion and stable denoising.Importantly, the proposed framework does not redefine the probabilistic diffusion process; instead, diffusion is employed as an iterative denoising and refinement mechanism, while PKD-R operates purely as a distribution-level regularizer in the training objective. PKD-R, as the core innovation, achieves stable global distribution alignment through multidirectional projection and kernel discrepancy integration, addressing the lack of explicit distributional constraints in diffusion models. HERD employs hierarchical recursion and stride-based exponential moving average (EMA) gating mechanisms to stabilize multiscale denoising, suppress error propagation, and balance global and local features, providing consistent feature inputs for PKD-R. CADEF leverages coarse-to-fine cross modal attention fusion, utilizing DSM's height-aware positional encodings to guide global structure while optimizing fine-scale texture details, mitigating geometry-texture mismatches and delivering fine-grained multimodal features for PKD-R.

The primary innovations and contributions of this study are summarized as follows:

- We propose a projection-kernel distribution regularization term (PKD-R) for diffusion-based multimodal remote sensing segmentation, which aligns predicted and ground-truth class-probability distributions at the global level. This formulation introduces projection-based maximum mean discrepancy into the segmentation objective and derives a closed-form solution for projection integration, significantly reducing computational complexity while improving stability and effectiveness in high-dimensional multimodal data spaces.
- We integrate PKD-R into a diffusion-based segmentation framework as a distribution-level regularizer without modifying the intrinsic forward or reverse diffusion dynamics, enabling efficient global statistical alignment during iterative denoising.
- To support multimodal feature interaction and stable multiscale refinement within this framework, we incorporate a cross-attention dual-encoder fusion module (CADEF) and a hierarchical EMA-gated recursive denoising mechanism (HERD), which enhance geometry–texture consistency and denoising stability.
- Extensive experiments on the ISPRS Vaihingen and Potsdam benchmarks, together with cross-architecture evaluations on CNN- and Transformer-based segmentation models, demonstrate consistent accuracy improvements and confirm empirically validated effectiveness and cross-architecture transferability on standard multimodal RS benchmarks.

Related works

CNN-based semantic segmentation in remote sensing

Before deep learning became dominant, RS segmentation widely relied on expert-designed descriptors (e.g., spectral–spatial indices, morphological profiles) coupled with classical classifiers such as SVMs and Random Forests, often regularized by graphical models (MRF/CRF) or Bayesian fusion[23–26]. These pipelines were effective but brittle under distribution shifts and modality heterogeneity; extensive feature engineering was required and

cross modal coupling remained ad hoc[6,7]. CNNs offered a decisive shift by learning hierarchical features end-to-end, reducing reliance on manual design and enabling more scalable representation learning.

Seminal CNN-based segmentation models–FCN and SegNet–pioneered dense prediction with fully convolutional decoders and efficient upsampling[27,28]. In RS, CNN variants tailored to very-high-resolution (VHR) imagery (e.g., attention-enhanced U-Net family) improved boundary localization and small-object delineation[29]. For multi-source inputs, early fusion and late fusion are straightforward yet limited; stronger gains emerged from architectures that explicitly leverage geometric priors (e.g., LiDAR/DSM) alongside optical texture through modality-aware encoders or interaction blocks[30]. cross modal multiscale fusion further strengthened complementary cues while mitigating resolution mismatches and alignment errors[13]. Beyond standard backbones, graph-based or hypergraph modeling has been explored to encode high-order relations across modalities and classes, complementing convolutional fields[31]. Despite progress, CNNs' locality bias makes global context aggregation expensive or indirect, and balancing global consistency with fine structures under multimodal distribution shifts remains challenging.

Transformer-based segmentation and multimodal fusion

To address the limited receptive field and long-range dependency modeling in CNNs, Vision Transformers introduce self-attention to capture global interactions directly. In RS segmentation, Transformer-based schemes demonstrate clear benefits on fine-resolution scenes, either in pure-Transformer forms or in hybrid designs that retain convolutional inductive biases[8,9]. In dense prediction, hierarchical backbones (e.g., Swin[32]) with U-Net–style decoders are widely adopted, and UNetFormer[33] further integrates windowed self-attention with skip connections into a lightweight, efficient encoder–decoder paradigm, achieving a favorable accuracy–efficiency trade-off for high-resolution remote sensing segmentation. Cross-branch designs (e.g., CNN–Transformer fusion) exploit complementary strengths in locality and non-local reasoning[34,35].

A crucial thread in RS is how to fuse heterogeneous modalities effectively within Transformer frameworks. Multimodal ViTs for land-cover classification show that modality-specific encoders and token-level interactions can improve robustness and scalability[12,36]. For segmentation, multilevel multimodal fusion Transformers explicitly integrate hierarchical interactions and cross-attention between optical and geometric streams to strengthen both global semantics and boundary fidelity[11]. Beyond segmentation, Transformer-based fusion for generic image fusion (e.g., SwinFusion) provides additional evidence that long-range modeling helps reconcile heterogeneous statistics across sources[37]. Nevertheless, naively coupling modalities can induce modality dominance or over-smoothing; it remains non-trivial to decide at which scales and directions cross modal attention should operate–coarse scales for geometry-aware structure versus fine scales for texture-preserving detail. Moreover, computational costs and data efficiency are practical concerns in RS where labeled, well-registered multimodal data are limited[38].

Generative models for dense prediction

While Transformers improve global context and flexible fusion, most RS segmentation pipelines remain fundamentally discriminative: they optimize per-pixel classification without explicitly modeling output distributional structure. Generative models such as GAN[39] provide a complementary view by learning data distributions and enabling stronger priors for reconstruction, translation, and dense labeling. In RS, adversarial learning has been leveraged for cloud removal and domain adaptation, improving realism and cross-domain robustness but often facing training instabilities and mode collapse[40–42].

Diffusion models (DMs) offer a likelihood-based alternative with stable training and excellent distribution modeling. Foundational works demonstrate progressive refinement and cascaded generation achieving high fidelity across resolutions, with broad surveys consolidating theoretical and practical advances[43,44]. Recent applications extend diffusion to structured prediction and change detection, hinting at advantages in uncertainty calibration and sample diversity; however, most diffusion-based RS segmentation still relies predominantly on pixel-level supervision, lacking explicit distribution-level alignment of semantic outputs. This gap is exacerbated in multimodal settings where TOP and DSM exhibit heterogeneous statistics and alignment noise. Consequently, there is growing interest in frameworks that (i) introduce distributional constraints during training, (ii) stabilize iterative refinement, and (iii) perform principled cross modal interaction under coarse-to-fine semantics–directions our work directly targets.

Distribution alignment and discrepancy measures

To address global distribution misalignment in high-dimensional spaces, Maximum Mean Discrepancy (MMD) has been widely adopted as a kernel-based nonparametric metric[20]. However, in high-dimensional class-probability spaces, standard MMD may exhibit reduced sensitivity to complex distributional shifts when discrepancies along different directions partially offset each other in the RKHS mean embedding[21,22].

Projection-based variants such as Sliced MMD and Sliced Wasserstein distance define discrepancies as expectations of one-dimensional projected measures, which in practice are approximated using Monte Carlo sampling of projection directions. This introduces approximation variance and requires a sufficiently large number of projections for stable estimates under a fixed computational budget, motivating analytic integration strategies[45].

In contrast, the proposed Projection Kernel Discrepancy (PKD) analytically integrates Gaussian-kernel MMD over projection directions drawn from a Gaussian measure $\Psi = N(0, I)$, yielding a closed-form pairwise kernel (Eq. 26) that avoids numerical integration and sampling noise. This closed-form formulation provides a deterministic and smooth estimate of projection-aggregated discrepancy while retaining the global statistical alignment capability of kernel-based two-sample measures.

From a theoretical perspective, PKD can be interpreted as a projection-integrated kernel discrepancy related to sliced kernel methods, but with exact integration instead of Monte Carlo approximation. Its unbiased U-statistic estimator preserves the computational order of standard MMD ($\mathcal{O}(N^2)$), while empirically demonstrating improved stability and effectiveness for dense prediction in multimodal remote sensing segmentation. These properties make PKD particularly suitable for softmax probability distributions, where accurate global class-balance and boundary alignment are essential.

Preliminaries

Diffusion models

Diffusion models model data distributions through a progressive noising-to-denoising paradigm: the forward process corrupts clean samples into an isotropic Gaussian prior, while the reverse process reconstructs data.

Basic definitions

Let $\mathrm{x}_0 \sim p_{\mathrm{data}}(\mathrm{x})$ denote a clean sample. The forward Markov chain of length T is

$$q(\mathrm{x}_{1:T} \mid \mathrm{x}_0) = \prod_{t=1}^{T} q(\mathrm{x}_t \mid \mathrm{x}_{t-1}), \quad \mathrm{x}_T \approx N(0, \mathrm{I}) \tag{1}$$

The parameterized reverse process is

$$p_\theta(\mathrm{x}_{0:T}) = p(\mathrm{x}_T) \prod_{t=1}^{T} p_\theta(\mathrm{x}_{t-1} \mid \mathrm{x}_t) \tag{2}$$

Forward diffusion

With variance schedule $\{\beta_t\}_{t=1}^{T}$, define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The forward kernel is

$$q(\mathrm{x}_t \mid \mathrm{x}_{t-1}) = N\left(\sqrt{\alpha_t}\,\mathrm{x}_{t-1},\ \beta_t \mathrm{I}\right) \tag{3}$$

Any timestep t sample is given by

$$\mathrm{x}_t = \sqrt{\bar{\alpha}_t}\,\mathrm{x}_0 + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \mathrm{I}) \tag{4}$$

Noise prediction objective

A U-Net predicts noise $\hat{\boldsymbol{\epsilon}}_\theta(\mathrm{x}_t, t) \approx \boldsymbol{\epsilon}$. Training minimizes the simplified objective:

$$L_{\mathrm{noise}}(\theta) = \mathbb{E}_{t, \mathrm{x}_0, \boldsymbol{\epsilon}} \left\| \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta(\mathrm{x}_t, t) \right\|_2^2 \tag{5}$$

where t is uniformly sampled from $\{1, \ldots, T\}$.

Maximum mean discrepancy

Setup and Notation Let $\mathrm{x} \in X$ follow distribution P. Consider a positive definite kernel $k : X \times X \to \mathbb{R}$ with Reproducing Kernel Hilbert Space (RKHS) H and feature map $\varphi : X \to H$, where $k(\mathrm{x}, \mathrm{x}') = \langle \varphi(\mathrm{x}), \varphi(\mathrm{x}') \rangle_H$.

The distribution P is embedded into H as:

$$\mu_P := \mathbb{E}_{\mathrm{x} \sim P}\big[\varphi(\mathrm{x})\big] = \mathbb{E}_{\mathrm{x} \sim P}\big[k(\mathrm{x}, \cdot)\big] \in H \tag{6}$$

Similarly, μ_Q for distribution Q. The MMD is:

$$\mathrm{MMD}(P, Q; H) := \left\| \mu_P - \mu_Q \right\|_H \tag{7}$$

$$\mathrm{MMD}^2(P, Q) = \left\| \mu_P - \mu_Q \right\|_H^2 \tag{8}$$

For a characteristic kernel (e.g., Gaussian RBF), $\mathrm{MMD}(P, Q) = 0 \iff P = Q$, enabling two-sample testing and distribution alignment.

MMD^2 is expressed as:

$$\begin{aligned} \mathrm{MMD}^2(P, Q) = {} & \mathbb{E}_{\mathrm{x}, \mathrm{x}' \sim P} k(\mathrm{x}, \mathrm{x}') + \mathbb{E}_{\mathrm{y}, \mathrm{y}' \sim Q} k(\mathrm{y}, \mathrm{y}') \\ & - 2\, \mathbb{E}_{\mathrm{x} \sim P, \mathrm{y} \sim Q} k(\mathrm{x}, \mathrm{y}) \end{aligned} \tag{9}$$

This form scales to high-dimensional spaces by only requiring kernel evaluations.

Treating pixelwise softmax vectors as points in $\mathbb{R}^C$, MMD measures global statistical differences between predicted and ground-truth distributions, capturing class imbalance and multi-modality beyond per-pixel errors.

Given i.i.d. samples $\{x_i\}_{i=1}^{m} \sim P$ and $\{y_j\}_{j=1}^{n} \sim Q$, the unbiased U-statistic estimator is:

$$\widehat{\mathrm{MMD}}_{\mathrm{U}}^{2} = \frac{1}{m(m-1)} \sum_{i \neq i'} k(x_i, x_{i'}) + \frac{1}{n(n-1)} \sum_{j \neq j'} k(y_j, y_{j'}) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) \tag{10}$$

Method

Overall framework overview

The proposed PKDiff framework aims to address the challenges in multimodal (RS) semantic segmentation by integrating global consistency and local precision from multimodal data. The framework's core comprises three innovative components: the PKD-R, HERD, and CADEF. The first step of this method is to use pure noise. The overall workflow is as follows: Input images, including the fused TOP and DSM features (Img), are subjected to a diffusion process with pure noise to generate the noisy state P_{noised}. These are fed into the noise prediction network along with the time step t to predict noise and gradually denoise to obtain fine segmentation results. The CADEF module within the denoising network fuses TOP and DSM features, while the HERD module stabilizes multiscale denoising through hierarchical recursion and EMA gating mechanisms. The PKD-R module computes the L_{PKD} loss by performing multidirectional one-dimensional projections of seg_{pred} and seg_{gt}, enhancing global distribution alignment. The overall structure is illustrated in Fig. 1, where the diffusion backbone is explicitly decomposed into encoder and decoder stages, and the placements of CADEF (cross-modal fusion), HERD (recursive denoising stabilization), and PKD-R (distribution alignment loss) are clearly annotated to improve architectural interpretability.

Cross-Attention Dual-Encoder Fusion (CADEF)

In multi-source remote sensing, TOP provide rich texture and spectral information but are prone to geometric distortions, while DSM encode height and structural boundaries yet may suffer from noise or coarseness in shadowed regions. Simple feature concatenation often leads to: (i) geometry-texture misalignment, where global layouts are disrupted by insufficient geometric guidance; and (ii) reduced detail fidelity, where boundaries and textures compete, causing over-smoothing. CADEF addresses these issues through cross modal attention in a coarse-to-fine manner: DSM provides geometric priors to guide global structure at coarse scales, while at fine scales, DSM refines boundaries to enhance texture details. The CADEF architecture is shown in Fig. 2.

Given a registered pair $(I_{\mathrm{top}}, I_{\mathrm{dsm}})$, multi-scale patch embeddings yield token sequences for scales $s \in \{1/8\}$ (coarse) and $s \in \{1/4, 1/2\}$ (fine):

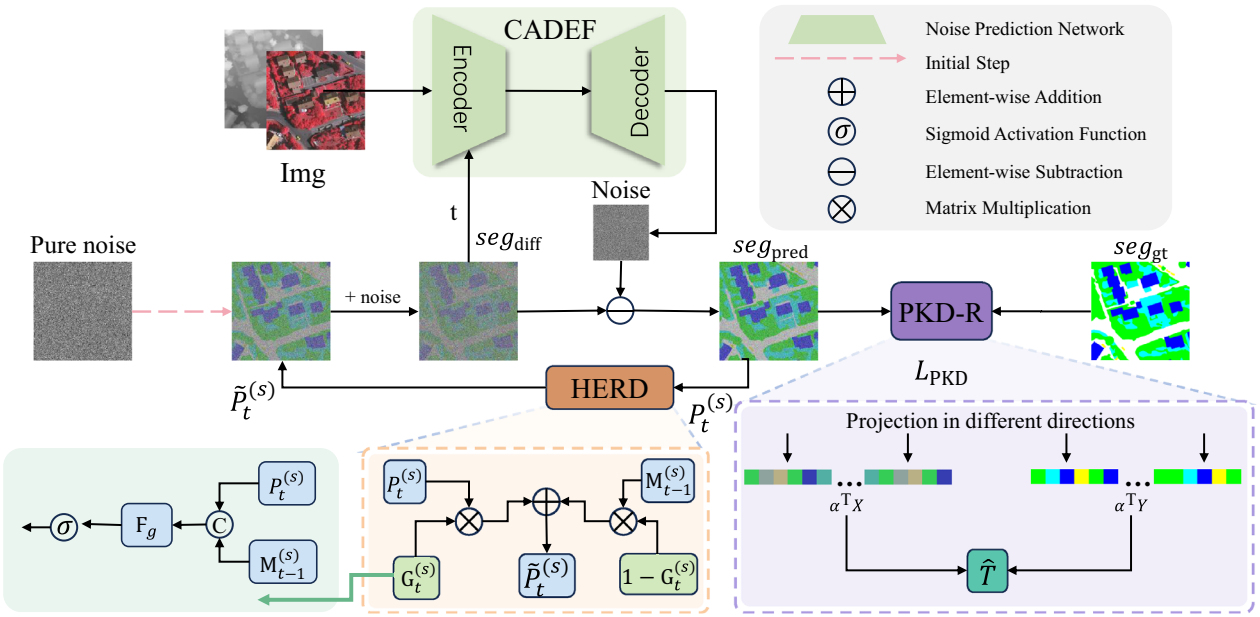

Fig. 1. Overview of the diffusion-based segmentation framework with projection-kernel regularization. The input image (Img), incorporating fused TOP and DSM features, undergoes a diffusion process to generate the noisy state P_{noised}. Together with the timestep t, these are input into the noise prediction network (highlighted in green). The CADEF module fuses TOP and DSM features, while the HERD module stabilizes denoising through hierarchical recursion, outputting $\widetilde{P}_t^{(s)}$. The PKD-R module computes the L_{PKD} loss by performing multidirectional one-dimensional projections of seg_{pred} and seg_{gt}.

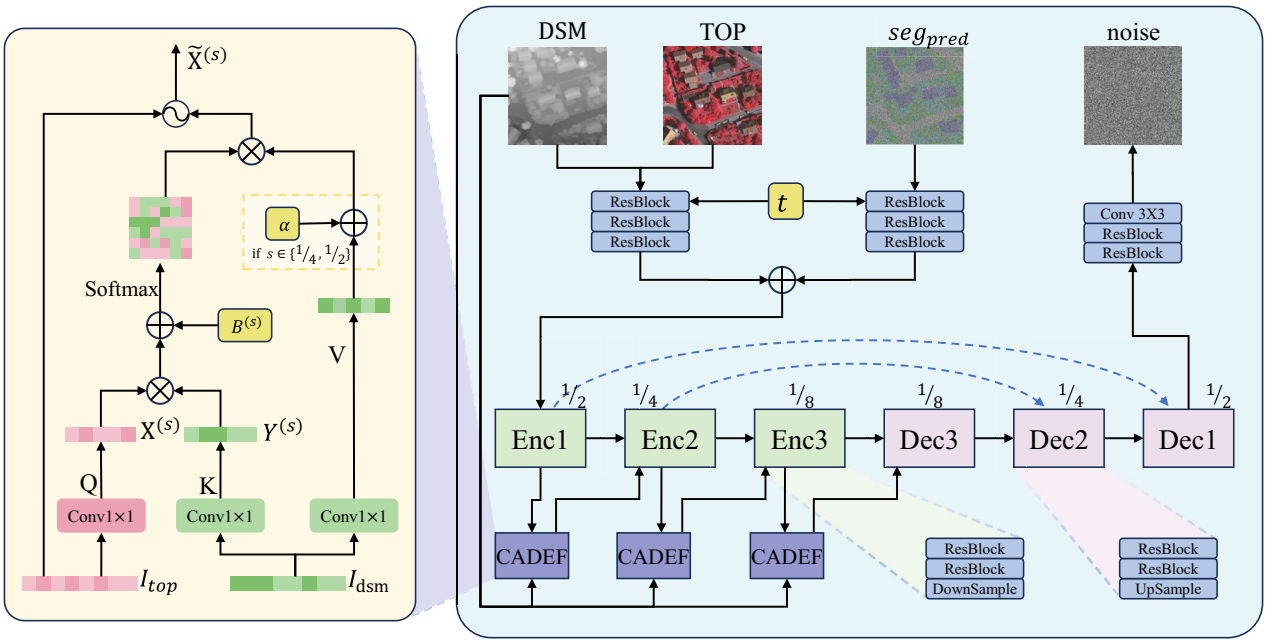

Fig. 2. It illustrates the overall structure of the CADEF, including multi-scale patch embedding, cross modal attention mechanisms at coarse (1/8) and fine (1/4, 1/2) scales, and DSM-guided global structure alignment and boundary refinement processes. Additionally, the architecture incorporates time-conditioned cross modal fusion, where timestep embedding is generated via sinusoidal mapping and MLP, and a U-shaped top-down fusion strategy in the decoder produces DSM-informed skip connections. Finally, the high-resolution feature is transformed into a noise through residual blocks. The symbols $\oplus$, $\otimes$, and $\curvearrowright$ represent element-wise addition, element-wise multiplication, and residual connection operations, respectively.

$$\mathrm{X}^{(s)} = \mathrm{PE}^{(s)}_{\mathrm{top}}(I_{\mathrm{top}}), \quad \mathrm{Y}^{(s)} = \mathrm{PE}^{(s)}_{\mathrm{dsm}}(I_{\mathrm{dsm}}) \tag{11}$$

where N_s is the number of tokens and d_s the channel dimension at scale s. The TOP branch acts as the query stream, with DSM providing key/value streams for cross modal fusion.

Height-Aware Positional Encoding (HAPE): Standard 2D positional encoding $\Pi^{(s)}$ is augmented with normalized DSM height and its spatial derivatives:

$$h = \mathrm{norm}(I_{\mathrm{dsm}}), \quad \mathrm{g} = \nabla h = \left(\partial_x h, \partial_y h\right) \tag{12}$$

forming the height-aware encoding:

$$\Pi^{(s)}_{\mathrm{H}} = \mathrm{MLP}\left(\left[\ \Pi^{(s)} \parallel h^{(s)} \parallel \mathrm{g}^{(s)}\ \right]\right) \tag{13}$$

$\Pi^{(s)}_{\mathrm{H}}$ introduces learnable relative positional biases in attention, enhancing sensitivity to height discontinuities (e.g., eaves, facades) for boundary-aware modeling.

cross modal Attention (Coarse-to-Fine Staging):

1) Coarse scales (1/8): DSM guides global structure. Using TOP as queries and DSM as keys/values:

$$\mathrm{Q}_X = \mathrm{W}^{(s)}_Q \mathrm{X}^{(s)} \quad \mathrm{K}_Y = \mathrm{W}^{(s)}_K \mathrm{Y}^{(s)} \quad \mathrm{V}_Y = \mathrm{W}^{(s)}_V \mathrm{Y}^{(s)} \tag{14}$$

with learnable bias $\mathrm{B}^{(s)} = \mathrm{B}^{(s)}\left(\Pi^{(s)}, \Pi^{(s)}_{\mathrm{H}}\right)$. The cross-attention and residual update are:

$$\mathrm{A}^{(s)}_{X \leftarrow Y} = \mathrm{softmax}\left(\frac{\mathrm{Q}_X \mathrm{K}_Y^{\top}}{\sqrt{d_s}} + \mathrm{B}^{(s)}\right) \tag{15}$$

$$\widetilde{\mathrm{X}}^{(s)} = \mathrm{X}^{(s)} + \mathrm{A}^{(s)}_{X \leftarrow Y} \mathrm{V}_Y \tag{16}$$

Coarse scales leverage DSM priors to align global layouts, reducing texture-induced errors.

2) Fine scales (1/4, 1/2): DSM refines boundaries. At fine scales, TOP dominates, with DSM enhancing boundaries:

$$\widehat{\mathrm{A}}^{(s)}_{X\leftarrow Y} = \mathrm{softmax}\left(\frac{\mathrm{Q}_X \mathrm{K}_Y^{\top}}{\sqrt{d_s}} + \mathrm{B}^{(s)}\right) \tag{17}$$

$$\widetilde{\mathrm{X}}^{(s)} = \mathrm{X}^{(s)} + \alpha^{(s)} \cdot \widehat{\mathrm{A}}^{(s)}_{X\leftarrow Y} \mathrm{V}_Y \tag{18}$$

$$\alpha^{(s)} = \sigma\left(\mathrm{Conv}_{3\times 3}\left(\sqrt{(\partial_x h^{(s)})^2 + (\partial_y h^{(s)})^2 + \epsilon}\right)\right) \tag{19}$$

where $\partial_x h^{(s)}$ and $\partial_y h^{(s)}$ are the spatial gradients of the normalized DSM at scale s computed using Sobel filters, $\mathrm{Conv}_{3\times 3}$ is a 3x3 convolution, σ is the sigmoid activation, and $\epsilon = 10^{-12}$ ensures numerical stability. $\alpha^{(s)} \in [0, 1]$ is a per-pixel weight at scale s that enhances DSM feature contributions at structural boundaries (high gradient regions) and reduces them in flat areas to improve boundary accuracy while maintaining TOP texture details.

CADEF's coarse-to-fine cross-attention integrates DSM's geometric priors with TOP's texture details, addressing geometry-texture misalignment and improving boundary precision. Height-aware positional encoding enhances attention to structural discontinuities. Compared to simple concatenation, CADEF better balances global consistency and local detail, supporting diffusion-based denoising and segmentation.

Hierarchical EMA-Gated Recursive Denoising (HERD)

In iterative denoising, naïve recursion (feeding previous outputs directly to the next step) often leads to: (i) error propagation, where early biases amplify across iterations, destabilizing the solution; and (ii) global-local imbalance, where coarse-scale estimates prioritize structure but lack detail, while fine-scale estimates enhance details but are prone to noise. HERD addresses these issues by: (i) using hierarchical recursion across scales, (ii) applying a cross-step EMA for stable memory, and (iii) employing a learnable gate to balance historical and current predictions per pixel and class. This stabilizes denoising, enhances cross-scale consistency, and improves detail preservation.

Single-scale gating with EMA memory

Let the network output at iteration t and scale s be $\mathrm{P}_t^{(s)} \in \mathbb{R}^{B\times C\times H_s\times W_s}$, with a cross-step memory $\mathrm{M}_t^{(s)}$ per scale.

Gate generation (pixel- and class-wise): A lightweight module F_g produces the gate:

$$\mathrm{G}_t^{(s)} = \sigma\Big(F_g\big([\mathrm{P}_t^{(s)} \parallel \mathrm{M}_{t-1}^{(s)}]\big)\Big) \tag{20}$$

where $\mathrm{G}_t^{(s)} \in [0, 1]^{B\times C\times H_s\times W_s}$, $[\cdot\|\cdot]$ denotes channel-wise concatenation and σ is the sigmoid. The gate balances current predictions and historical memory.

Gated fusion:

$$\widetilde{\mathrm{P}}_t^{(s)} = \mathrm{G}_t^{(s)} \odot \mathrm{P}_t^{(s)} + \big(1 - \mathrm{G}_t^{(s)}\big) \odot \mathrm{M}_{t-1}^{(s)} \tag{21}$$

where $\odot$ is the Hadamard product. The gate prioritizes current predictions when confident, otherwise relying on historical memory to reduce noise propagation.

EMA memory update:

$$\mathrm{M}_t^{(s)} = \gamma \mathrm{M}_{t-1}^{(s)} + (1 - \gamma)\, \widetilde{\mathrm{P}}_t^{(s)} \tag{22}$$

where $\gamma \in [0, 1)$, γ controls the forgetting rate, acting as a low-pass filter to suppress noise while maintaining responsiveness.

Hierarchical recursion and cross-scale propagation

HERD operates on a coarse-to-fine pyramid of scales $s = 1, \ldots, S$. Cross-scale flow is achieved via top-down prior injection.

Prior upsampling injection (coarse → fine):

$$\widehat{\mathrm{M}}_t^{(s+1)} = \alpha\, U^{s\to s+1}\big(\widetilde{\mathrm{P}}_t^{(s)}\big) + (1 - \alpha)\, \mathrm{M}_{t-1}^{(s+1)} \tag{23}$$

where $\alpha \in [0, 1]$ is a fixed scalar (set to 0.6), $U^{s\to s+1}$ upsamples the coarse-scale output. The prior $\widehat{\mathrm{M}}_t^{(s+1)}$ initializes fine-scale memory, balancing coarse-scale structure with fine-scale history.

Fine-scale gated alignment: At scale $s + 1$, replace $\mathrm{M}_{t-1}^{(s+1)}$ with $\widehat{\mathrm{M}}_t^{(s+1)}$ in (20)–(22), aligning coarse priors with fine-scale evidence via the gate–EMA pipeline. This recursion ensures global structure propagation while refining details.

HERD provides: (i) error attenuation through EMA-based memory to mitigate noise and cascading errors; (ii) cross-scale coherence via top-down prior injection, ensuring consistent global structure; and (iii) adaptive refinement through pixel-/class-wise gates to enhance boundaries and textures. Integrated with PKD-R and CADEF, HERD supports stable denoising in multimodal RS tasks.

Projection Kernel Discrepancy Regularizer (PKD-R)

Standard MMD, as described in Section "Maximum Mean Discrepancy", quantifies distribution differences via mean embeddings in a RKHS. However, in high-dimensional or multimodal RS data, MMD's sensitivity is limited by the "cancellation effect," where discrepancies in different directions offset each other. To address this, we propose the PKD-R, which projects high-dimensional class-probability vectors onto multiple one-dimensional subspaces, computes Gaussian-kernel MMD per projection, and integrates discrepancies across directions. A key innovation is our derivation of a closed-form solution for the integrated MMD, detailed in Appendix A, which simplifies computation and enhances sensitivity to shifts in location, scale, and covariance. This complements pointwise losses (e.g., MSE, CE) by enforcing global statistical alignment, improving consistency in complex RS segmentation tasks.

Let $\mathrm{x}\sim P$ and $\mathrm{y}\sim Q$ be class-probability vectors in $\mathbb{R}^C$. For a direction $\boldsymbol{\alpha}\in\mathbb{R}^C$, we compute the one-dimensional MMD between projections $\boldsymbol{\alpha}^\top\mathrm{x}$ and $\boldsymbol{\alpha}^\top\mathrm{y}$. Integrating over directions with a Gaussian measure $\Psi = N(0,\mathrm{I})$ yields the PKD-R statistic:

$$D = \int_{\mathbb{R}^C} \mathrm{MMD}^2\big(\boldsymbol{\alpha}^\top\mathrm{x},\ \boldsymbol{\alpha}^\top\mathrm{y}\big)\,\mathrm{d}\Psi(\boldsymbol{\alpha}) \tag{24}$$

Using a Gaussian kernel $k(u,v)=\exp\big(-\|u-v\|^2/(2\sigma^2)\big)$, we derive a closed-form expression (see Appendix A):

$$D = \mathbb{E}\Big[d(\mathrm{x}_1,\mathrm{x}_2) - 2\,d(\mathrm{x}_1,\mathrm{y}_2) + d(\mathrm{y}_1,\mathrm{y}_2)\Big] \tag{25}$$

$$d(\mathrm{z}_1,\mathrm{z}_2) = \left(1+\frac{\|\mathrm{z}_1-\mathrm{z}_2\|^2}{\sigma^2}\right)^{-\frac{1}{2}} \tag{26}$$

where $(\mathrm{x}_1,\mathrm{x}_2)\overset{\text{i.i.d.}}{\sim}P$, $(\mathrm{y}_1,\mathrm{y}_2)\overset{\text{i.i.d.}}{\sim}Q$, and pairs are mutually independent. This closed-form solution eliminates numerical integration, enabling efficient computation.

Given i.i.d. samples $\{\mathrm{x}^{(i)}\}_{i=1}^N\sim P$ and $\{\mathrm{y}^{(j)}\}_{j=1}^N\sim Q$ (with $N=BHW$ for RS images), the unbiased U-statistic estimator, consistent with Section "Maximum Mean Discrepancy", is:

$$\begin{aligned}\widehat{D} = {} & \frac{1}{N(N-1)}\sum_{i\neq i'} d\big(\mathrm{x}^{(i)},\mathrm{x}^{(i')}\big) + \frac{1}{N(N-1)}\sum_{j\neq j'} d\big(\mathrm{y}^{(j)},\mathrm{y}^{(j')}\big) \\ & - \frac{2}{N^2}\sum_{i=1}^N\sum_{j=1}^N d\big(\mathrm{x}^{(i)},\mathrm{y}^{(j)}\big)\end{aligned} \tag{27}$$

In a diffusion-based segmentation framework, let $seg_{\mathrm{pred}}\in\mathbb{R}^{B\times C\times H\times W}$ be the denoised class-probability map at each training timestep. Reshape it to $\mathrm{P}_{\mathrm{pred}}\in\mathbb{R}^{N\times C}$ and the one-hot ground truth to $\mathrm{P}_{\mathrm{gt}}\in\mathbb{R}^{N\times C}$. Using (26), compute the empirical statistic $\widehat{D}$ as:

$$\begin{aligned}\widehat{D} = {} & \frac{1}{N(N-1)}\sum_{i\neq i'} d\big(\mathrm{P}^{(i)}_{\mathrm{pred}},\mathrm{P}^{(i')}_{\mathrm{pred}}\big) + \frac{1}{N(N-1)}\sum_{j\neq j'} d\big(\mathrm{P}^{(j)}_{\mathrm{gt}},\mathrm{P}^{(j')}_{\mathrm{gt}}\big) \\ & - \frac{2}{N^2}\sum_{i=1}^N\sum_{j=1}^N d\big(\mathrm{P}^{(i)}_{\mathrm{pred}},\mathrm{P}^{(j)}_{\mathrm{gt}}\big)\end{aligned} \tag{28}$$

The PKD-R loss is set as $L_{\mathrm{PKD}}\approx\widehat{D}$. The total training objective combines:

$$L_{\mathrm{total}} = L_{\mathrm{MSE}} + \lambda\,L_{\mathrm{PKD}} \tag{29}$$

where $L_{\mathrm{MSE}} = L_{\mathrm{noise}}(\theta)$ [Eq. (5) in Sec. 3.1] optimizes local noise prediction errors in the diffusion process, and L_{PKD}, driven by the closed-form solution, serves as the segmentation prediction loss, enforcing global distribution alignment to enhance consistency in remote sensing (RS) data distributions.

The closed-form solution in (26), derived in Appendix A, is a core innovation of PKD-R, enabling efficient computation of projection-integrated without numerical integration. By addressing the cancellation effect, PKD-R improves sensitivity to high-dimensional and multimodal distribution shifts, complementing pointwise losses to balance local detail correction with global statistical alignment. Integrated with HERD's stable multiscale denoising and CADEF's multimodal feature fusion, PKD-R achieves consistent global structures and precise boundaries in complex RS tasks.

Finally, the complete training procedure at each timestep from the current prediction and ground true, is summarized in Algorithm 1.

Require: Images I; one-hot labels Y_{onehot}; timesteps T; kernel bandwidth σ; weights γ, λ

Initialize previous segmentation S_{prev} with random values; initialize HERD memory $M_0 = \mathbf{0}$.
for $t = 1$ to T **do**
  $P_{\text{prev}} = \text{softmax}(S_{\text{prev}})$.
  CADEF fusion: $I_{\text{fused}} = \text{CADEF}(I, t)$.
  Diffuse: $seg_{\text{diff}} = \text{diffuse}(P_{\text{prev}}, t)$; $noise_{\text{gt}} = seg_{\text{diff}} - Y_{\text{onehot}}$.
  Denoise: $noise_{\text{pred}} = \text{model}(seg_{\text{diff}}, I_{\text{fused}}, t)$.
  Define the current time step prediction result: $seg_{\text{pred}} = seg_{\text{diff}} - noise_{\text{pred}}$.
  HERD stabilization: $G_t = \sigma\big(F_g([seg_{\text{pred}}, M_{t-1}])\big)$; $\widetilde{seg}_{\text{pred}} = G_t \odot seg_{\text{pred}} + (1 - G_t) \odot M_{t-1}$; $M_t = \gamma M_{t-1} + (1 - \gamma)\widetilde{seg}_{\text{pred}}$; set $seg_{\text{pred}} = \widetilde{seg}_{\text{pred}}$.
  Noise loss compute: $L_{\text{mse}} = \|noise_{\text{pred}} - noise_{\text{gt}}\|_2^2$.
  Segmentation loss compute:
  $\mathbf{P}_{\text{pred}} = \text{Flatten}(seg_{\text{pred}})$; $\mathbf{P}_{\text{gt}} = \text{Flatten}(Y_{\text{onehot}})$.
  $d(\mathbf{z}_1, \mathbf{z}_2) = \left(1 + \|\mathbf{z}_1 - \mathbf{z}_2\|_2^2/\sigma^2\right)^{-1/2}$.
  $\widehat{D} = \frac{1}{N(N-1)}\sum_{i \neq i'} d(\mathbf{P}_{\text{pred}}^{(i)}, \mathbf{P}_{\text{pred}}^{(i')}) + \frac{1}{N(N-1)}\sum_{j \neq j'} d(\mathbf{P}_{\text{gt}}^{(j)}, \mathbf{P}_{\text{gt}}^{(j')}) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} d(\mathbf{P}_{\text{pred}}^{(i)}, \mathbf{P}_{\text{gt}}^{(j)})$.
  $L_{\text{PKD}} = \widehat{D}$.
  Total loss : $L_{\text{total}} = L_{\text{mse}} + \lambda L_{\text{PKD}}$.
  $S_{\text{prev}} = seg_{\text{pred}}$.
end for

Algorithm 1. PKDiff: Training with Per-Timestep Loop

Experiment

Datasets and evaluation protocol

We evaluate the proposed method on two public high–resolution aerial semantic labeling benchmarks released by the International Society for Photogrammetry and Remote Sensing (ISPRS): the *Vaihingen* and *Potsdam* datasets. Both benchmarks target challenging urban-scene understanding.

Vaihingen. This dataset comprises 33 orthoimages acquired over the German town of Vaihingen, each with an average size of approximately 2500×2000 pixels. The ground sampling distance (GSD) is 9 cm, and three spectral bands are provided: near-infrared (NIR), red (R), and green (G). The scenes feature complex residential areas with numerous small buildings, structurally intricate rooftops, and densely packed objects, which collectively stress fine boundary delineation and inter-class discrimination. Under the standard evaluation protocol, 12 images are used for training and 4 images for testing.

Potsdam. This dataset contains 38 larger orthoimages of size 6000×6000 pixels captured over the city of Potsdam. The spatial resolution is higher, with a GSD of 5 cm, and four bands are available: RGB and NIR. The scenes include large-scale urban landscapes, a dense historic center with complexly shaped buildings, and extensive shadows cast by tall structures, which occlude other objects and increase semantic labeling difficulty. The standard train–test split uses 18 images for training and 5 images for testing.

Both datasets share the same six semantic classes for evaluation: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background (misc.). Their very high spatial resolution emphasizes fine structural details; combined with widely adopted train–test splits and a principled tiling strategy, these benchmarks provide a fair and reproducible platform for assessing boundary precision, separation of spectrally similar categories (e.g., vegetation types), and structural consistency.

Some data samples from the two datasets are illustrated in Fig. 3. To handle the large image sizes during training and testing, we adopt a sliding-window strategy with 256×256 patches. During training, the stride is set to 256 (non-overlapping) to maximize data independence and improve efficiency. During testing, the stride is reduced to 64 to introduce substantial overlap; predictions in overlapping regions are then aggregated by simple averaging, which effectively mitigates border artifacts and improves the coherence and accuracy of the final segmentation maps.

Evaluation metrics

We evaluate performance using three standard metrics: F1-score, mean Intersection over Union (mIoU), and Overall Accuracy (OA). Let $M \in \mathbb{N}^{K \times K}$ denote the confusion matrix for K semantic classes (excluding void regions), where M_{ij} counts pixels with ground truth class i predicted as class j. For class k,

$$\text{TP}_k = M_{kk}, \quad \text{FP}_k = \sum_{i \neq k} M_{ik}, \quad \text{FN}_k = \sum_{j \neq k} M_{kj} \tag{30}$$

F1-score. Precision and recall for class k are

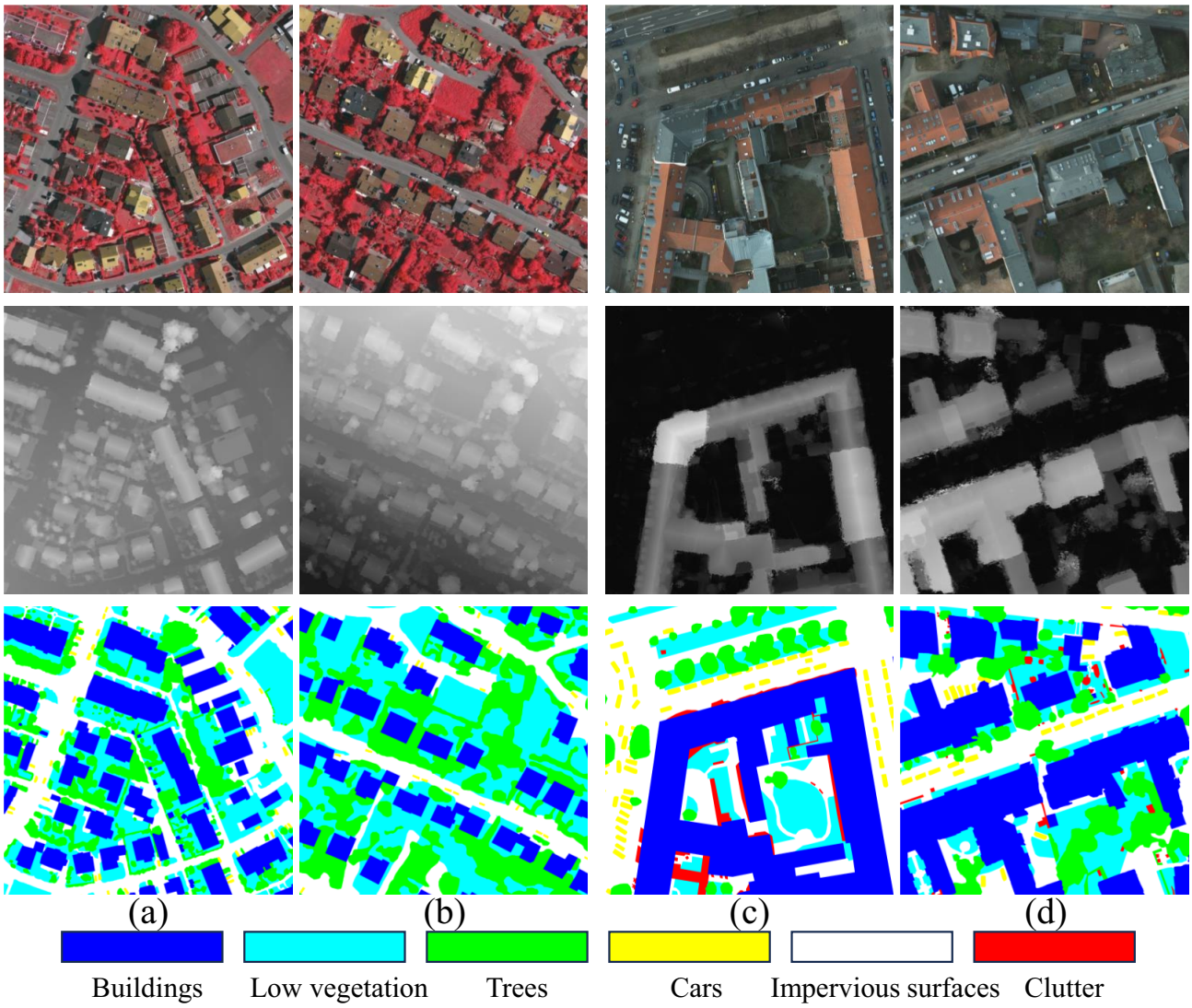

Fig. 3. Examples from the Vaihingen and Potsdam datasets are illustrated. (**a**) and (**b**) are two samples from Vaihingen, (**c**) and (**d**) are two samples from Potsdam. The size is 2048×2048.The first row shows the orthophotos with three channels (NIRRG for Vaihingen and RGB for Potsdam). The second and third rows present the corresponding pixel-wise depth information and ground truth labels. They show the individual and complementary characteristics of remote sensing data from different sources.

$$\mathrm{Prec}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \quad \mathrm{Rec}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{31}$$

with per-class F1-score

$$\mathrm{F1}_k = \frac{2 \cdot \mathrm{Prec}_k \cdot \mathrm{Rec}_k}{\mathrm{Prec}_k + \mathrm{Rec}_k} = \frac{2 \cdot \mathrm{TP}_k}{2 \cdot \mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{32}$$

We report the macro-average F1-score, $\mathrm{F1} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{F1}_k$.

Mean Intersection over Union (mIoU). The IoU for class k is

$$\mathrm{IoU}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{33}$$

with mean IoU as

Type	Model	Imp	Building	Low	Trees	Car	OA	F1	mIoU
CNN-based	MANet	92.28	94.09	72.66	**94.44**	83.54	90.05	88.55	79.91
	ABCNet	91.24	96.53	76.42	91.70	81.19	90.43	87.90	78.96
	PSPNet	90.02	94.65	77.47	90.48	72.23	89.31	86.23	76.39
Transformer-based	FTransUNet	89.22	95.86	76.17	88.07	74.44	88.45	85.68	75.55
	ASMFNet	89.43	95.39	73.12	89.64	34.69	88.14	78.51	67.82
	CMFNet	90.76	95.45	80.09	89.90	83.20	90.14	88.45	79.76
	UNetFormer	91.28	95.81	77.35	91.31	**88.42**	90.33	89.03	80.68
Diffusion-based	SegDiff	70.66	83.76	58.82	81.33	30.82	74.77	74.75	52.52
	RNDiff	**92.29**	96.34	79.25	90.75	87.32	90.89	88.70	80.19
	PKDiff	91.65	**96.84**	**81.46**	90.54	88.21	**91.19**	**89.51**	**81.46**

Table 1. Quantitative results on the ISPRS **Vaihingen** dataset. Best results are shown in bold, and second-best results are underlined (%). Imp: impervious surface; Low: low vegetation. Imp: impervious surface; Low: low vegetation; OA: overall accuracy; mIoU: mean IoU.

Type	Model	Imp	Building	Low	Trees	Car	OA	F1	mIoU
CNN-based	MANet	90.61	97.58	87.34	**85.45**	**95.61**	89.82	**91.05**	83.91
	ABCNet	89.88	96.52	85.23	84.10	93.72	88.45	89.51	81.35
	PSPNet	87.53	96.04	84.61	78.99	95.14	86.70	88.01	79.07
Transformer-based	FTransUNet	90.13	97.03	84.69	82.16	92.68	88.04	88.80	80.32
	ASMFNet	86.46	95.78	83.03	63.63	79.78	82.36	81.15	69.19
	CMFNet	91.05	97.51	86.14	79.24	95.04	88.22	88.98	80.62
	UNetFormer	91.10	97.36	**88.63**	85.07	95.58	89.94	91.03	83.87
Diffusion-based	SegDiff	50.48	22.83	20.71	5.51	0.21	28.03	27.39	10.71
	RNDiff	90.12	96.80	83.22	83.31	95.30	87.61	88.51	79.90
	PKDiff	**95.12**	**97.72**	85.76	80.17	95.33	**90.54**	91.02	**84.03**

Table 2. Quantitative results on the ISPRS **Potsdam** dataset. Best results are shown in bold, and second-best results are underlined (%). Imp: impervious surface; Low: low vegetation. Imp: impervious surface; Low: low vegetation; OA: overall accuracy; mIoU: mean IoU.

$$\mathrm{mIoU} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{IoU}_k \tag{34}$$

Overall Accuracy (OA). OA is the fraction of correctly classified pixels:

$$\mathrm{OA} = \frac{\sum_{k=1}^{K} M_{kk}}{\sum_{i=1}^{K}\sum_{j=1}^{K} M_{ij}} \tag{35}$$

All metrics are computed using official dataset evaluation masks.

Implementation details

We train the network for 100 epochs using the AdamW optimizer with an initial learning rate of 3×10^{-4} and a weight decay of 0.01. The batch size is set to 32. The weight parameter λ for the segmentation loss L_{PKD} is set to 0.05, and the α in the HERD module is set to 0.6. All experiments are conducted on four NVIDIA RTX 4090 GPUs (24 GB memory each).

Performance comparison

We benchmark the proposed PKDiff against nine representative methods spanning three paradigms: CNN-based (MANet[46], ABCNet[47], PSPNet[48]), Transformer-based (FTransUNet[11], ASMFNet[49], CMFNet[13], UNetFormer[33]), and diffusion-based (SegDiff[50], RNDiff[51]). Most competing approaches are tailored for remote sensing. In our implementation, ABCNet, PSPNet, and UNetFormer are trained on optical imagery only (RGB), in order to highlight the contribution of DSM cues and the advantages of multimodal learning over single-modality counterparts. Comparative quantitative results on Vaihingen and Potsdam are reported in Table 1 and Table 2, respectively.

Performance Comparison on the Vaihingen Dataset Considering the design emphases of the compared families (multi-scale context aggregation, cross modal attention, and generative denoising), Table I shows that *PKDiff* attains the best scores on all three primary metrics: F1 = 89.51%, mIoU = 81.46%, and OA = 91.19%.

Against the strongest RGB-only baseline UNetFormer (F1 = 89.03%, mIoU = 80.68%, OA = 90.33%), *PKDiff* improves by $+0.48/+0.78/+0.86\%$. Relative to the diffusion baseline RNDiff (F1 = 88.70%, mIoU = 80.19%, OA = 90.89%), *PKDiff* yields $+0.81/+1.27/+0.30\%$. These gains align with our modeling rationale: *PKDiff* augments a recursive diffusion backbone with an integral-MMD distribution alignment (mitigating prediction bias and residual noise) and injects DSM geometry via a cross modal dual-encoder at coarse–to–fine scales, thereby achieving a stronger balance between global consistency and boundary fidelity.

At the class level, *PKDiff* excels on geometry/boundary-sensitive categories. For *Building*, it achieves 96.84%, surpassing ABCNet (96.53%) and RNDiff (96.34%); for *Low vegetation*, it reaches 81.46%, exceeding CMFNet (80.09%) and RNDiff (79.25%) by +1.37 and +2.21 %, respectively. On the small-object *Car* class, *PKDiff* obtains 88.21%, slightly below UNetFormer (88.42%, −0.21 %) yet clearly above RNDiff (87.32%), indicating that DSM height cues, coupled with diffusion-based detail refinement, improve contour preservation. Meanwhile, RNDiff attains the best *Impervious surface* score (92.29%) over *PKDiff* (91.65%), reflecting diffusion's innate advantage on large homogeneous regions; MANet leads *Trees* with 94.44% vs. 90.54% for *PKDiff*, consistent with CNNs' strong local texture discrimination on heavily texture-dependent classes. Overall, *PKDiff* leads on the key difficult categories (buildings, low vegetation, and small vehicles) while remaining competitive on large homogeneous and strongly textured classes, thereby delivering the best combined F1/mIoU/OA.

A broader cross-method view further quantifies the breadth and stability of *PKDiff*'s advantages. Compared with CMFNet, *PKDiff* improves by (F1/mIoU/OA $+1.06/+1.70/+1.05$)%; versus MANet by $+0.96/+1.55/+1.14$%; versus ABCNet by $+1.61/+2.50/+0.76$%; and relative to PSPNet and FTransUNet, it gains $+3.28/+3.83$% in F1, $+5.07/+5.91$% in mIoU, and $+1.88/+2.74$% in OA. Even against the strong diffusion baseline RNDiff, *PKDiff* maintains consistent positive margins (see above). These trends concur with each method family's stated strengths: RGB-only UNetFormer/PSPNet/ABCNet remain competitive where optical textures are clean, but lack geometric priors under deep shadows and height variations; multimodal CMFNet/ASMFNet/FTransUNet leverage cross-scale alignment and attention for stronger structural modeling, yet exhibit room for improvement in distribution-level constraints and small-object boundaries; and diffusion approaches such as SegDiff (F1/mIoU/OA = 74.75%/52.52%/74.77%) and ASMFNet's weakness on extreme categories (e.g., *Car* = 34.69%) further underscore the necessity of unifying generative detail refinement, distribution alignment, and cross modal geometric injection within a single framework. Collectively, the evidence indicates that multimodal structural priors and distribution-level regularization are crucial for boundary continuity, and small-object segmentation in complex urban scenes.

Visual Comparison on the Vaihingen Dataset: Figure 4 presents a visual comparison of semantic segmentation results on two 1024 × 1024 samples from the Vaihingen test set. For each sample, subfigure (a) displays the IRRG image, (b) the DSM height map, and (c) the ground truth labels. Subfigures (d) to (l) illustrate the predictions from competing models: MANet (d), ABCNet (e), PSPNet (f), FTransUNet (g), ASMFNet (h), CMFNet (i), UNetFormer (j), RNDiff (k), and the proposed PKDiff (l). Purple boxes are added to all subfigures to highlight regions with notable differences, such as complex boundaries, shadowed areas, and small objects.

In the first sample (subscript 1), the proposed PKDiff (l1) achieves superior boundary fidelity for buildings and cars within the purple boxes, accurately separating them from impervious surfaces and low vegetation, whereas MANet (d1) and ABCNet (e1) exhibit over-segmentation artifacts, leading to fragmented building contours. PSPNet (f1) and FTransUNet (g1) struggle with low vegetation, often misclassifying it as trees due to texture similarities, resulting in blurred transitions. ASMFNet (h1) shows severe under-detection of cars, missing small vehicles entirely, which aligns with its quantitative weakness on small-object classes (e.g., Car IoU=34.69%). CMFNet (i1) and UNetFormer (j1) perform better on homogeneous regions like impervious surfaces but introduce noise in shadowed low vegetation areas. RNDiff (k1), as a diffusion baseline, reduces some noise compared to earlier models but still exhibits residual artifacts around building edges. In contrast, PKDiff leverages cross modal DSM injection and distribution alignment to refine details, yielding cleaner delineations and fewer misclassifications.

For the second sample (subscript 2), PKDiff (l2) excels in distinguishing trees from low vegetation in densely vegetated regions, preserving intricate boundaries that are merged in PSPNet (f2) and FTransUNet (g2). Cars are detected with high precision, avoiding the false positives seen in UNetFormer (j2) and the omissions in ASMFNet (h2). MANet (d2) overemphasizes local textures, leading to clutter misclassification as low vegetation, while RNDiff (k2) handles large homogeneous areas well but falters on fine-grained geometry, such as vehicle contours under shadows. Overall, the visual results corroborate the quantitative metrics: PKDiff consistently outperforms baselines in geometry-sensitive categories (e.g., buildings and low vegetation) by integrating generative denoising with multimodal priors, achieving better global consistency and local detail preservation across complex urban scenes.

Performance Comparison on the Potsdam Dataset From Table 2, the proposed PKDiff exhibits the most balanced overall performance on Potsdam, attaining the best Overall Accuracy (OA) and mean IoU (mIoU) with OA = 90.54% and mIoU = 84.03%. Its F1-score reaches 91.02%, effectively on par with the strongest baselines (MANet: 91.05%; UNetFormer: 91.03%; gaps of 0.03 and 0.01 percentage points, respectively). Against the diffusion baseline RNDiff, mIoU = 79.90%), PKDiff delivers sizeable gains of $+2.93/+2.51/+4.13$ %, indicating that, within a generative framework, introducing distribution-level regularization and DSM-driven geometric priors can systematically strengthen generalization and discrimination. Compared with the strongest RGB-only competitor UNetFormer, PKDiff improves OA by +0.60 % (90.54% vs. 89.94%) and mIoU by +0.16 % (84.03% vs. 83.87%), while essentially matching F1 (−0.01 %). Relative to classical CNN baselines, PKDiff yields consistent positives over MANet (OA/mIoU = 89.82%/83.91%), PSPNet (86.70%/79.07%), and ABCNet (88.45%/81.35%), with the advantages being more apparent under Potsdam's larger image scale and complex shadowing.

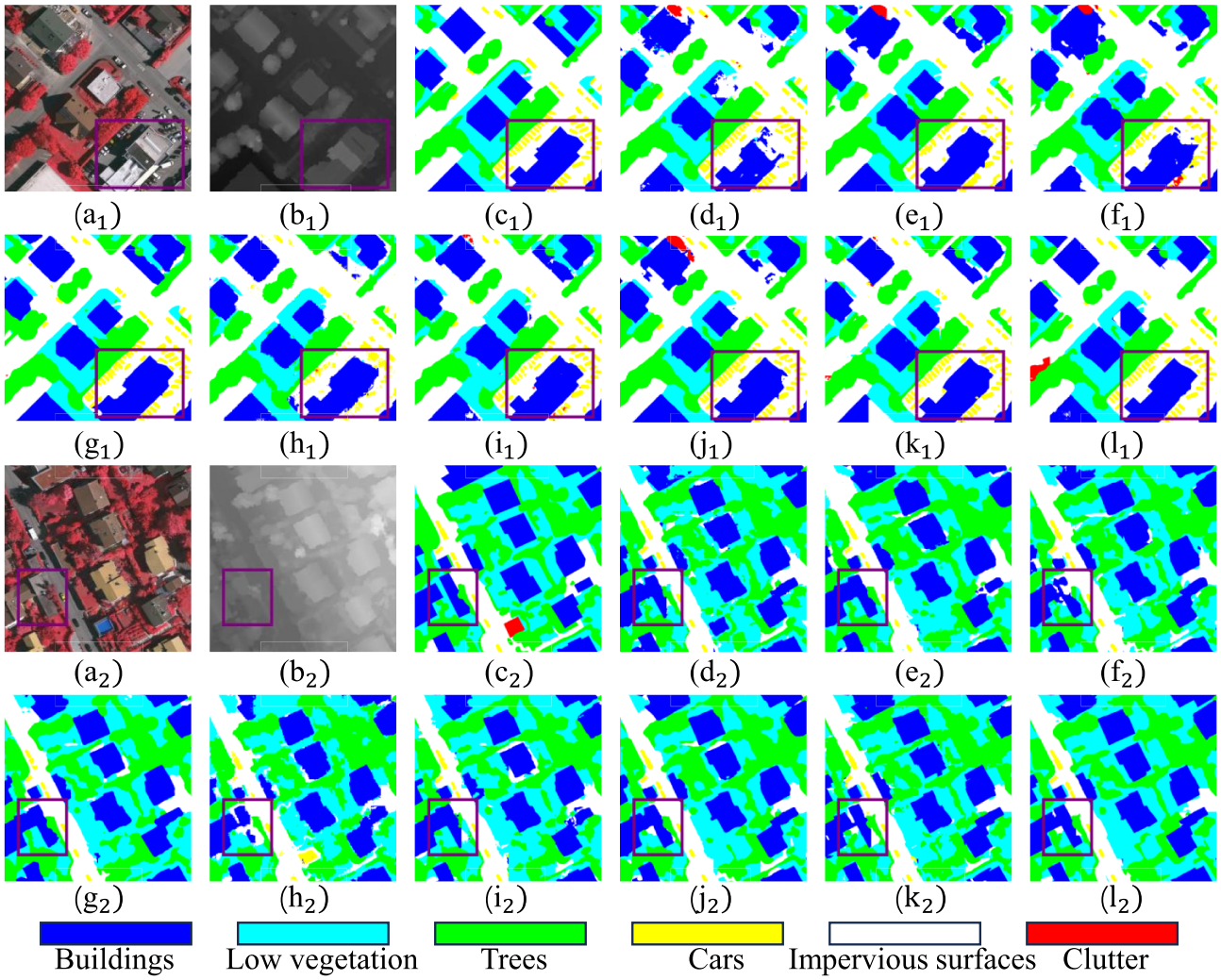

Fig. 4. Visualization comparison on the 1024 × 1024 Vaihingen test set. (**a**) IRRG image, (**b**) DSM, (**c**) groundtruth, (**d**) MANet, (**e**) ABCNet, (**f**) PSPNet, (**g**) FTransUNet, (**h**) ASMFNet, (**i**) CMFNet, (**j**) UNetFormer, (**k**) RNDiff, (**l**) proposed, respectively. To highlight differences, purple boxes are added to all subfigures. Subscripts 1 and 2 denote sample indices.

From a class-wise perspective, PKDiff leads on *Impervious surface* and *Building* with 95.12% and 97.72%, respectively, both exceeding MANet (90.61%, 97.58%) and UNetFormer (91.10%, 97.36%). In particular, its *Impervious surface* score shows a +5.00 % advantage over RNDiff (95.12% vs. 90.12%), evidencing stronger consistency on large homogeneous regions; we attribute this to the combination of recursive diffusion for detail reconstruction and CADEF's coarse-to-fine geometric injection, yielding more complete recovery of roads, squares, and roof contours. For the small-object *Car* class, PKDiff reaches 95.33%, ranking in the first tier–slightly below MANet (95.61%) and UNetFormer (95.58%)–yet clearly above RNDiff (95.30%) and the multimodal CMFNet (95.04%), indicating that diffusion-based boundary refinement plus DSM height cues help separate adjacent instances and preserve object contours. On vegetation-related categories, PKDiff attains 85.76% for *Low vegetation* and 80.17% for *Trees*, trailing UNetFormer (88.63%, 85.07%) and MANet (87.34%, 85.45%). This can be explained by two factors consistent with method characteristics: (i) RGB-only UNetFormer/MANet leverage strong global receptive fields or local texture modeling that are highly sensitive to canopy/grass textures; and (ii) DSM can be noisy over vegetation, making the geometry–texture trade-off in cross modal alignment more conservative for these classes, which also leads to some regression relative to RNDiff on *Trees* (83.31%). Nevertheless, PKDiff employs PKD– R at the distribution level to suppress class bias, maintaining strong overall balance and thus yielding the best mIoU/OA.

The performances of Transformer and multimodal baselines further substantiate these findings. CMFNet and FTransUNet leverage cross-scale attention to mitigate geometry–texture misalignment, yet still face challenges in maintaining consistency for small objects and fine boundaries on Potsdam's large scenes with heavy shadows. ASMFNet shows notably low scores on *Trees* (63.63%) and *Car* (79.78%), suggesting higher sensitivity to scale and cross modal alignment in complex urban imagery. By contrast, SegDiff exhibits training/adaptation

instability on this dataset (F1/mIoU/OA = 27.39%/10.71%/28.03%), while RNDiff, although more stable and competitive, is still surpassed by PKDiff across OA/F1/mIoU.

In summary, Potsdam mirrors the conclusions drawn on Vaihingen: deeply integrating cross modal geometric priors and distribution alignment (PKD– R) into a diffusion framework yields pronounced benefits on structural classes and trades for stronger global consistency, culminating in the best overall OA/mIoU with F1 on par with the top baselines. Fine-grained vegetation separation remains a promising direction (e.g., canopy-aware positional encoding, boundary-sensitive class reweighting, or dedicated DSM denoising/refinement branches), which we leave for future work.

Visual Comparison on the Potsdam Dataset: Figure 5 illustrates a visual comparison of semantic segmentation outcomes on two 1024 × 1024 samples from the Potsdam test set. Purple boxes are overlaid on all subfigures to emphasize areas with significant discrepancies, including shadowed regions, intricate boundaries, and small objects like cars.

In the first sample (subscript 1), PKDiff (l1) excels at delineating impervious surfaces and buildings within the purple boxes, accurately capturing roof edges and road layouts under heavy shadows, where MANet (d1) and UNetFormer (j1) show fragmentation and misclassify parts as low vegetation. ABCNet (e1) and PSPNet (f1) exhibit over-smoothing, blurring transitions between trees and low vegetation. FTransUNet (g1) and ASMFNet (h1) struggle with car detection, often merging them with surrounding surfaces or omitting them, consistent with their lower quantitative scores (e.g., ASMFNet Car IoU=79.78%). CMFNet (i1) improves multimodal alignment but retains noise in vegetated areas, while RNDiff (k1) reduces artifacts but shows residual errors around building edges. PKDiff's integration of DSM-driven geometric priors and distribution alignment yields sharper boundaries.

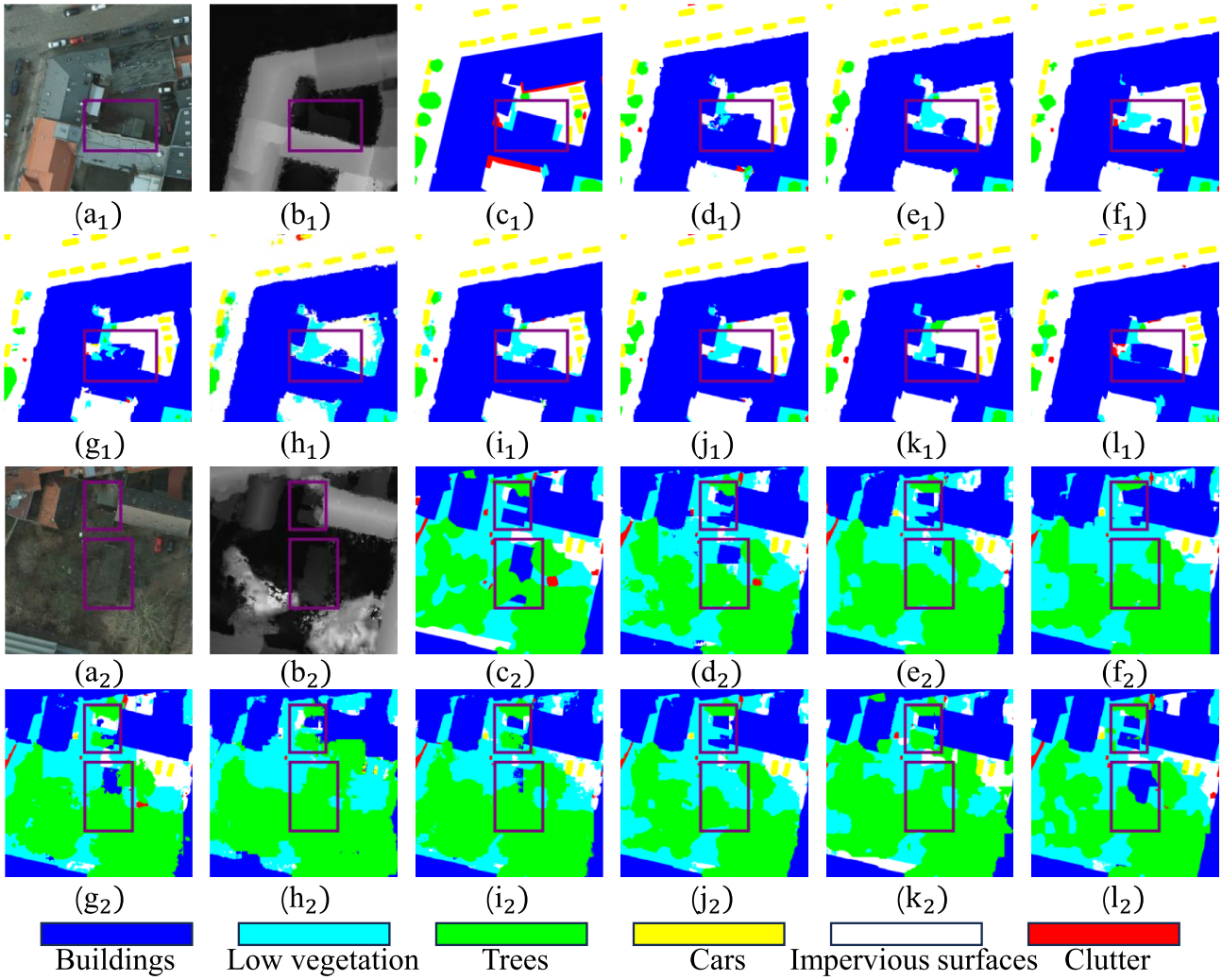

Fig. 5. Visualization comparison on the 1024 × 1024 Potsdam test set. (**a**) IRRG image, (**b**) DSM, (**c**) groundtruth, (**d**) MANet, (**e**) ABCNet, (**f**) PSPNet, (**g**) FTransUNet, (**h**) ASMFNet, (**i**) CMFNet, (**j**) UNetFormer, (**k**) RNDiff, (**l**) proposed, respectively. To highlight differences, purple boxes are added to all subfigures. Subscripts 1 and 2 denote sample indices.

CADEF	HERD	OA	F1	mIoU
–	–	89.43	86.99	77.52
–	✓	89.68	87.58	78.36
✓	–	90.86	88.45	79.86
✓	✓	**91.19**	**89.51**	**81.46**

Table 3. Ablation on Vaihingen. ✓ indicates the component is enabled.

L_{MSE}	L_{CE}	L_{PKD}	OA	F1	mIoU
✓			90.92	88.66	80.23
✓	✓		90.83	88.67	80.20
✓	✓	✓	91.11	88.57	80.15
✓		✓	**91.19**	**89.51**	**81.46**

Table 4. Ablation on loss functions (Vaihingen). ✓ means the loss is enabled.

For the second sample (subscript 2), the primary distinction lies in the recognition of buildings. PKDiff (l2) effectively identifies building structures within the purple boxes, accurately delineating their outlines and edges, where PSPNet (f2) and FTransUNet (g2) fail to separate buildings from surrounding vegetation, resulting in conflated regions. UNetFormer (j2) over-detects clutter as buildings, while ASMFNet (h2) misses several building instances. MANet (d2) overemphasizes local textures, leading to misclassification of clutter as impervious surfaces, and RNDiff (k2) struggles with precise geometric fidelity in shadowed building areas. CMFNet (i2) shows improved alignment but retains noise in complex structural regions. These visual results align with the quantitative data, highlighting PKDiff's superior consistency on structural classes (e.g., buildings) through the integration of generative refinement and cross modal cues, though it trades some texture sensitivity on vegetation, achieving the best balanced OA and mIoU in complex, large-scale urban environments. This improved detection precision, particularly for buildings, may be attributed to the influence of PKD-R, which enhances global distribution alignment by mitigating the cancellation effect in high-dimensional data through multidirectional projection MMD, ensuring better performance of complex structural features as observed in the purple boxed areas.

Ablation study on components

To verify the independent contributions and synergies of each component, we conduct ablation experiments on the Vaihingen dataset for **CADEF** and **HERD**. The results are shown in Table 3: the baseline without any components is $\mathrm{OA} = 89.43\%$, $\mathrm{F1} = 86.99\%$, and $\mathrm{mIoU} = 77.52\%$; when adding only HERD (removing CADEF), the performance is $\mathrm{OA} = 89.68\%$, $\mathrm{F1} = 87.58\%$, and $\mathrm{mIoU} = 78.36\%$; when using only CADEF (without HERD), the metrics are $90.86\%/88.45\%/79.86\%$; using CADEF+HERD (the full model) achieves the best $91.19\%/89.51\%/81.46\%$.

Comparing each "remove-one" setting to the full model quantifies *marginal contributions*: (i) Removing **CADEF** causes notable drops of $\mathrm{OA}\ -1.51$ %, $\mathrm{F1}\ -1.93$ %, and $\mathrm{mIoU}\ -3.10$ %, indicating that cross-modal geometric injection and coarse-to-fine interactions are crucial for global structural consistency and boundary alignment, as stated in the introduction, CADEF utilizes DSM's height-aware positional encodings to guide global structure while optimizing fine-scale texture details, thereby alleviating geometry-texture mismatches and providing more features for PKD to facilitate distribution extraction; the larger decline in mIoU highlights CADEF's direct gain on inter-class boundaries and fine-grained regions. (ii) Removing **HERD** leads to $\mathrm{OA}\ -0.33$ %, $\mathrm{F1}\ -1.06$ %, and $\mathrm{mIoU}\ -1.60$ %, showing that recursive denoising's temporal consistency and boundary refinement effectively improve recall and IoU, especially for small objects and high-frequency details, as emphasized in the introduction, HERD stabilizes multiscale denoising through hierarchical recursion and stride-based exponential moving average (EMA) gating mechanisms, suppresses error propagation, and dynamically balances global and local features, providing more features for PKD to facilitate distribution extraction.

Lateral comparisons before/after adding CADEF further reveal its *dominant* role: from baseline to CADEF, mIoU improves by $+2.34$ % (77.52→79.86); from HERD to CADEF, mIoU improves by $+1.50$ % (78.36→79.86). This indicates that even without denoising mechanisms, cross-modal structural alignment alone provides stable gains, as stated in the introduction, CADEF enhances geometry-guided global consistency while preserving fine-grained texture details. Furthermore, given CADEF, adding HERD brings additional $+0.33/+1.06/+1.60$ % (OA/F1/mIoU), reflecting the *complementarity* between "geometric alignment–recursive denoising": CADEF stabilizes global layout and boundaries, HERD enhances temporal stability and detail restoration–jointly producing the full model's best results ($\mathrm{OA} = 91.19\%$, $\mathrm{F1} = 89.51\%$, $\mathrm{mIoU} = 81.46\%$). Overall, the cumulative gain in mIoU ($+3.94$ %) from baseline is notably larger than that in OA ($+1.76$ %), further confirming the framework's advantage on difficult boundaries and fine structures, in line with the design intent of multimodal geometric priors and recursive denoising.

Ablation study on loss functions

To validate the effectiveness of the proposed PKD-R in addressing the limitations of conventional pointwise losses (MSE and CE) as discussed in the introduction, we conduct an ablation study on the Vaihingen dataset. We evaluate the impact of combining the noise prediction loss (L_{MSE}), the segmentation cross-entropy loss (L_{CE}), and the segmentation PKD-R loss (L_{PKD}) on Overall OA, F1-score, and mIoU. Results are reported in Table 4.

The baseline with only L_{MSE} achieves OA=90.92%, F1=88.66%, and mIoU=80.23%, reflecting the diffusion model's ability to capture multimodal distributions but lacking explicit semantic supervision, as noted in the introduction. Adding L_{CE} slightly adjusts performance (OA=90.83%, F1=88.67%, mIoU=80.20%), indicating that CE's focus on dominant classes may introduce bias in high-dimensional RS data, consistent with its limitations in capturing global structure. Including all losses ($L_{\mathrm{MSE}} + L_{\mathrm{CE}} + L_{\mathrm{PKD}}$) improves OA to 91.11% but yields marginal gains in F1 (88.57%) and mIoU (80.15%), suggesting potential conflicts between CE and PKD-R. The best performance is obtained with $L_{\mathrm{MSE}} + L_{\mathrm{PKD}}$ (OA=91.19%, F1=89.51%, mIoU=81.46%), with gains of +0.27 %, +0.85 %, and +1.23 % over the baseline, respectively. This underscores PKD-R's ability to enhance global distribution alignment, mitigating the cancellation effect and improving sensitivity to complex distributional shifts, as described in the introduction. These results confirm that PKD-R, when paired with noise prediction, consistently improves global distribution alignment and boundary delineation on the standard multimodal urban RS benchmark, demonstrating empirically validated performance gains.

Ablation study on alternative distribution discrepancies

To clarify the distinction between PKD-R and existing discrepancy measures, we compare PKD-R with several representative kernel- and projection-based alternatives, including standard Gaussian MMD, sliced-MMD with different numbers of projection directions ($S = 10$ and $S = 20$), and sliced Wasserstein distance (SWD with $S = 100$). All methods are implemented within the same PKDiff framework while replacing only the distribution regularization term, ensuring a fair comparison.

Table 5 reports segmentation accuracy and computational overhead on the ISPRS Vaihingen dataset. Standard MMD significantly degrades performance, indicating that naive high-dimensional kernel matching is insufficient for probability-space alignment in diffusion-based segmentation. Projection-based Monte Carlo methods (sliced-MMD and SWD) partially recover performance but remain inferior to PKD-R and incur noticeably higher training cost due to multiple projection evaluations.

Notably, increasing the number of projections in sliced-MMD may not lead to monotonic performance improvement. For example, sliced-MMD with $S = 20$ performs slightly worse than $S = 10$. This behavior is expected because projection-based discrepancies rely on Monte Carlo sampling of directions, which introduces stochastic gradient variability and may alter optimization trajectories during diffusion training. Larger projection counts can reduce estimation variance but also increase computational burden.Increasing S from 10 to 20 leads to significant consumption of GPU memory. Due to the limitations of our experimental equipment, we did not conduct experiments with a larger S.

In contrast, PKD-R analytically integrates Gaussian-kernel discrepancies over projection directions and yields a deterministic closed-form estimator. This formulation achieves the best segmentation accuracy while maintaining training efficiency comparable to standard MMD. These results demonstrate that PKD-R provides a more stable and computationally efficient distribution alignment mechanism for diffusion-based remote sensing segmentation.

Generality beyond diffusion-based segmentation

To investigate whether PKD-R is intrinsically tied to diffusion training, we integrate the PKD regularization term into representative non-diffusion segmentation backbones, including PSPNet, MANet (CNN-based), and UNetFormer (Transformer-based), while keeping their original training settings unchanged.

As shown in Table 6, PKD consistently improves segmentation accuracy across all architectures. Specifically, PKD increases mIoU from 78.96% to 79.39% on PSPNet, from 79.91% to 80.72% on MANet, and from 80.68% to 81.54% on UNetFormer, yielding gains of 0.55%, 1.01%, and 1.07%, respectively.

These consistent improvements demonstrate that PKD-R is not inherently dependent on diffusion-based optimization, but instead acts as a general distribution-level regularizer that can enhance prediction consistency

Method	OA (%)	F1 (%)	mIoU (%)	Training Time/batch
MSE only	90.92	88.66	80.23	1.00×
+ Standard MMD	89.29	85.62	76.45	1.06×
+ Sliced-MMD ($S = 10$)	89.58	86.03	77.11	1.64×
+ Sliced-MMD ($S = 20$)	88.49	84.91	75.43	1.65×
+ SWD ($S = 100$)	90.06	85.73	77.06	1.27×
+ PKD-R (ours)	**91.19**	**89.51**	**81.46**	1.07×

Table 5. Comparison of PKD-R with alternative distribution discrepancy measures on the ISPRS Vaihingen dataset. All methods are evaluated within the same PKDiff framework by replacing only the distribution regularization term.

Method	Baseline mIoU	+PKD	Gain (%)
PSPNet	78.96	79.39	+0.55
MANet	79.91	80.72	+1.01
UNetFormer	80.68	81.54	+1.07

Table 6. Generality of PKD-R on non-diffusion segmentation backbones. PKD consistently improves mIoU across CNN-based (PSPNet, MANet) and Transformer-based (UNetFormer) architectures, demonstrating that the proposed distribution regularization is not intrinsically tied to diffusion training.

λ	OA	F1	mIoU
0.01	90.81	88.90	80.58
0.025	90.92	89.07	80.81
0.05	**91.19**	**89.51**	**81.46**
0.075	91.06	89.02	80.75
0.10	90.91	88.68	80.22

Table 7. Effect of weighting parameter λ for PKD-R on Vaihingen.

in both CNN and Transformer segmentation frameworks. This observation further supports the architectural generality and practical applicability of the proposed PKD formulation.

Impact of weighting parameter for PKD-R loss

To evaluate the impact of the weighting parameter λ in the PKD-R loss on model performance, we conducted an ablation study on the Vaihingen dataset. This study examines the effect of different λ values in the total loss function $L_{\text{total}} = L_{\text{MSE}} + \lambda L_{\text{PKD}}$ on the model's performance metrics. The results, presented in Table 7, show the Overall Accuracy (OA), F1-score, and mean Intersection over Union (mIoU) for λ values in {0.01, 0.025, 0.05, 0.075, 0.10}.

As shown in Table 7, model performance improves steadily across all three metrics as λ increases from 0.01 to 0.05. At $\lambda = 0.05$, the model achieves the best performance: OA of 91.19%, F1 of 89.51%, and mIoU of 81.46%. This result suggests that a moderate weighting of the PKD-R loss effectively balances global distribution alignment with local noise prediction optimization, enhancing both global consistency and boundary precision in complex remote sensing scenes.

However, as λ increases beyond 0.05 to 0.075 and 0.10, performance begins to decline. For instance, from $\lambda = 0.05$ to $\lambda = 0.075$, OA decreases by 0.13 percentage points, F1 drops by 0.49 percentage points, and mIoU decreases by 0.71 percentage points. At $\lambda = 0.10$, performance further degrades, with mIoU falling to 80.22%. This trend indicates that excessively high λ values may overemphasize global distribution alignment, suppressing the optimization of local details, which is critical for precise boundary and texture recovery in high-dimensional, multimodal remote sensing data.

Specifically, lower λ values (e.g., 0.01 and 0.025) result in insufficient contribution from L_{PKD}, limiting its ability to correct biases between predicted and ground-truth distributions. This leads to weaker global consistency, particularly in regions with complex geometric structures, such as buildings and small objects. Conversely, higher λ values (e.g., 0.075 and 0.10) may cause the model to over-prioritize distribution-level constraints, diminishing the role of L_{MSE} in local noise prediction, which negatively impacts boundary details and small-object segmentation (e.g., cars). In contrast, $\lambda = 0.05$ strikes an optimal balance between global distribution alignment and local feature optimization, aligning with the design goal of PKD-R to enhance sensitivity to complex distributional shifts via projection MMD while preserving the detail recovery capabilities of the diffusion model.

Furthermore, comparing these results with the loss ablation study in Table 4, the configuration with $\lambda = 0.05$ (OA=91.19%, F1=89.51%, mIoU=81.46%) significantly outperforms the baseline using only L_{MSE} (OA=90.92%, F1=88.66%, mIoU=80.23%). This further validates PKD-R's effectiveness in improving global consistency and boundary precision. In summary, $\lambda = 0.05$ is the optimal weight for the Vaihingen dataset, leveraging PKD-R's distribution alignment capabilities while synergizing with HERD's recursive denoising and CADEF's cross modal fusion to achieve superior segmentation performance. Future work could explore adaptive λ scheduling strategies to dynamically adjust the weight, accommodating varying dataset and scene characteristics.

Effect of training timesteps

As illustrated in Table 8 and Fig. 6, we evaluate the performance of PKDiff as the number of training timesteps increases from 2 to 64 on the Vaihingen dataset. The Overall Accuracy (OA) improves monotonically from 88.54% (2 steps) to 90.59% (16 steps), reaching a peak of **91.19%** at 64 steps. Similarly, the F1-score rises from 85.46% to 88.43% and culminates at **89.51%**, while the mean Intersection over Union (mIoU) increases from 75.31% to 79.76% and achieves **81.46%** at 64 steps. The most substantial gains occur between 2 and 16 steps, with improvements of +2.05 % in OA, +2.97 % in F1, and +4.45 % in mIoU. Beyond 32 steps, the enhancements

Timestep	OA	F1	mIoU
2	88.54	85.46	75.31
4	88.94	86.85	77.17
8	89.78	87.50	78.28
16	90.59	88.43	79.76
32	90.72	89.44	81.32
64	**91.19**	**89.51**	**81.46**

Table 8. Performance of PKDiff with different training timesteps on Vaihingen.

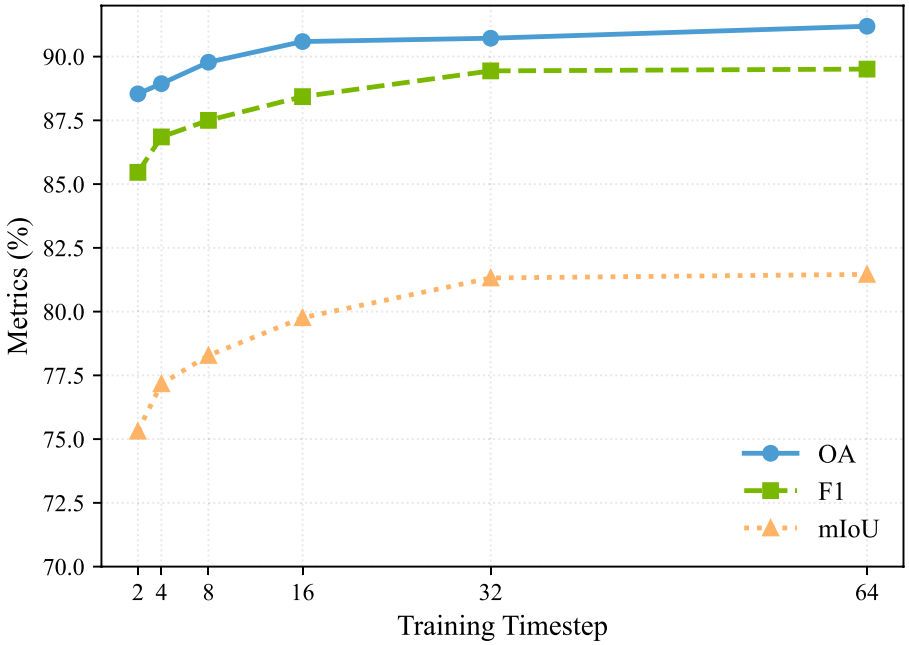

Fig. 6. Performance metrics of PKDiff on Vaihingen dataset across training timesteps (2–64). Curves show monotonic gains, with largest improvements from 2 to 16 steps; gains saturate beyond 32 steps, indicating 32 steps as an optimal accuracy-efficiency tradeoff. Best at 64 steps.

diminish: transitioning from 32 to 64 steps yields only +0.47 % in OA, +0.07 % in F1, and +0.14 % in mIoU, highlighting diminishing returns relative to increased computational overhead. Notably, the 32-step model captures approximately 97.7% of the total mIoU improvement observed at 64 steps, positioning 32 steps as an effective tradeoff between accuracy and efficiency. These trends demonstrate the effectiveness and efficiency of PKDiff's PKD-R and HERD modules. They enable strong performance and rapid convergence even at moderate timestep counts, while striking a favorable balance between global consistency and local precision in multimodal remote sensing segmentation.

Computational complexity

As summarized in Table 9, the proposed PKDiff achieves the best overall accuracy on Vaihingen with **OA/F1/mIoU = 91.19/89.51/81.46**, surpassing the strongest Transformer baseline *UNetFormer* by $+\mathbf{0.78}$ mIoU (80.68 $\rightarrow$ 81.46) and the multimodal *CMFNet* by $+\mathbf{1.70}$ mIoU (79.76 $\rightarrow$ 81.46). Despite its diffusion nature, PKDiff remains highly compact at 6.4 M parameters and 24.54MB, second only to *RNDiff* (6.29M/24.09MB), and notably smaller than *UNetFormer* (11.72M/44.87MB), reflecting roughly $\sim$45% fewer parameters and disk size.

In terms of throughput, Transformer/CNN baselines dominate: *UNetFormer* reaches the highest 171.36 FPS, followed by *ABCNet* at 163.27 FPS. Diffusion models trade speed for accuracy due to iterative denoising: PKDiff runs at 4.08 FPS, $\sim 42\times$ slower than *UNetFormer* (171.36/4.08) and $\sim 40\times$ slower than *ABCNet* (163.27/4.08). Nevertheless, compared with prior diffusion baselines, PKDiff not only improves accuracy over RNDiff by $+\mathbf{1.27}$ mIoU (80.19 $\rightarrow$ 81.46) but also retains a similarly minimal footprint. Meanwhile, classic diffusion segmentation (*SegDiff*) is both heavy (157.69M/601.55MB) and slow (1.77 FPS) with significantly inferior mIoU (52.52), underscoring the efficacy of our diffusion architecture. Overall, PKDiff delivers a favorable compactness–accuracy Pareto point, and its latency can be further reduced via step truncation, sampler acceleration, or diffusion distillation without increasing parameters.

Type	Model	Params (M)	Size (MB)	FPS	OA	F1	mIoU
CNN-based	MANet	35.86	137.05	55.55	90.05	88.55	79.91
	ABCNet	13.67	52.19	<u>163.27</u>	90.43	87.90	78.96
	PSPNet	49.07	187.42	35.38	89.31	86.23	76.39
Transformer-based	FTransUNet	203.40	775.93	10.96	88.45	85.68	75.55
	ASMFNet	83.48	321.60	32.43	88.14	78.51	67.82
	CMFNet	104.07	397.13	10.92	90.14	88.45	79.76
	UNetFormer	11.72	44.87	**171.36**	90.33	<u>89.03</u>	<u>80.68</u>
Diffusion-based	SegDiff	157.69	601.55	1.77	74.77	74.75	52.52
	RNDiff	**6.29**	**24.09**	7.17	<u>90.89</u>	88.70	80.19
	PKDiff	<u>6.43</u>	<u>24.54</u>	4.08	**91.19**	**89.51**	**81.46**

Table 9. Comparison of model complexity, inference speed, and segmentation accuracy on the ISPRS **Vaihingen** dataset with 256×256 input resolution. **Params** and **Size** denote the number of learnable parameters and storage footprint, respectively, while **FPS** indicates inference throughput. Best results are shown in bold, and second-best results are underlined.

Conclusion

This work presents a projection-kernel–regularized diffusion-based framework for multimodal remote sensing semantic segmentation. By integrating three core components–PKD-R for global distribution alignment, CADEF for coarse-to-fine multimodal integration, and HERD for stable multiscale refinement–PKDiff demonstrates consistent and competitive performance on the ISPRS Vaihingen and Potsdam benchmarks, matching or surpassing representative CNN-, Transformer-, and diffusion-based methods.

Despite these advances, challenges remain in fine-grained vegetation segmentation due to DSM noise, canopy variations, and the inherent trade-off between geometric and textural cues in multimodal alignment. Future work will investigate canopy-aware positional encodings and boundary-sensitive optimization strategies to further improve vegetation delineation, together with enhanced DSM denoising techniques for more reliable height representation. In addition, improving diffusion inference efficiency through adaptive timestep scheduling and model distillation, as well as exploring cross-domain and weakly supervised learning paradigms, may extend the empirical applicability of projection-kernel regularization to broader remote sensing settings. Extensions toward 3D spatiotemporal multimodal fusion with LiDAR, SAR, and time-series observations also constitute promising future research directions.

Data availability

Publicly available in a repository: The Vaihingen and Potsdam datasets used in the study is available at the following URL: https://isprs.org/resources/datasets/benchmarks/UrbanSemLab/default.aspx.

A appendix: derivation of PKD-R closed-form solution

This appendix derives the closed-form solution for the PKD-R statistic D in Section "Projection Kernel Discrepancy Regularizer (PKD-R)", a key theoretical contribution that enables efficient computation of projection-integrated MMD.

Given $\mathrm{x} \sim F$, $\mathrm{y} \sim G$ in $\mathbb{R}^p$, and a Gaussian kernel $k(u,v) = \exp\left(-\frac{(u-v)^2}{2\sigma^2}\right)$, the PKD-R statistic is:

$$D = \int_{\mathbb{R}^p} \mathrm{MMD}^2(\boldsymbol{\alpha}^T\mathrm{x}, \boldsymbol{\alpha}^T\mathrm{y})\,(2\pi)^{-p/2} \exp\left(-\frac{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}{2}\right) d\boldsymbol{\alpha}, \tag{36}$$

where $\mathrm{MMD}^2(\boldsymbol{\alpha}^T\mathrm{x}, \boldsymbol{\alpha}^T\mathrm{y}) = \mathbb{E}[k(\boldsymbol{\alpha}^T\mathrm{x}_1, \boldsymbol{\alpha}^T\mathrm{x}_2)] + \mathbb{E}[k(\boldsymbol{\alpha}^T\mathrm{y}_1, \boldsymbol{\alpha}^T\mathrm{y}_2)] - 2\mathbb{E}[k(\boldsymbol{\alpha}^T\mathrm{x}, \boldsymbol{\alpha}^T\mathrm{y})]$.
By linearity of expectation:

$$\begin{aligned} D = {} & \mathbb{E}_{\mathrm{x}_1,\mathrm{x}_2\sim F}\left[\mathbb{E}_{\boldsymbol{\alpha}}\left[k(\boldsymbol{\alpha}^T\mathrm{x}_1, \boldsymbol{\alpha}^T\mathrm{x}_2)\right]\right] \\ & + \mathbb{E}_{\mathrm{y}_1,\mathrm{y}_2\sim G}\left[\mathbb{E}_{\boldsymbol{\alpha}}\left[k(\boldsymbol{\alpha}^T\mathrm{y}_1, \boldsymbol{\alpha}^T\mathrm{y}_2)\right]\right] \\ & - 2\mathbb{E}_{\mathrm{x}\sim F,\mathrm{y}\sim G}\left[\mathbb{E}_{\boldsymbol{\alpha}}\left[k(\boldsymbol{\alpha}^T\mathrm{x}, \boldsymbol{\alpha}^T\mathrm{y})\right]\right]. \end{aligned} \tag{37}$$

For a pair a, b, let $\mathrm{d} = \mathrm{a} - \mathrm{b}$. The inner expectation is:

$$\mathbb{E}_{\boldsymbol{\alpha}\sim N(0,I_p)}\left[\exp\left(-\frac{\boldsymbol{\alpha}^T\mathrm{d}\mathrm{d}^T\boldsymbol{\alpha}}{2\sigma^2}\right)\right] = \int (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^T\mathrm{A}\boldsymbol{\alpha}\right) d\boldsymbol{\alpha}, \tag{38}$$

where $\mathrm{A} = I_p + \frac{1}{\sigma^2}\mathrm{d}\mathrm{d}^T$. Combine exponents: $-\frac{1}{2}\boldsymbol{\alpha}^T\left(I_p + \frac{1}{\sigma^2}\mathrm{d}\mathrm{d}^T\right)\boldsymbol{\alpha}$. The integral yields:

$$\left(\det\left(I_p + \frac{1}{\sigma^2}\mathrm{d}\mathrm{d}^T\right)\right)^{-1/2} = \left(1 + \frac{\|\mathrm{d}\|^2}{\sigma^2}\right)^{-1/2}, \tag{39}$$

using the matrix determinant lemma: $\det(I_p + \frac{1}{\sigma^2}\mathrm{d}\mathrm{d}^T) = 1 + \frac{\|\mathrm{d}\|^2}{\sigma^2}$. Thus:

$$\mathbb{E}_{\boldsymbol{\alpha}}\left[k(\boldsymbol{\alpha}^T\mathrm{a}, \boldsymbol{\alpha}^T\mathrm{b})\right] = \left(1 + \frac{\|\mathrm{a}-\mathrm{b}\|^2}{\sigma^2}\right)^{-1/2}. \tag{40}$$

Define $d(\mathrm{z}_1, \mathrm{z}_2) = \left(1 + \frac{\|\mathrm{z}_1-\mathrm{z}_2\|^2}{\sigma^2}\right)^{-1/2}$. Then:

$$\begin{aligned} D = {} & \mathbb{E}_{\mathrm{x}_1,\mathrm{x}_2\sim F}\left[d(\mathrm{x}_1, \mathrm{x}_2)\right] + \mathbb{E}_{\mathrm{y}_1,\mathrm{y}_2\sim G}\left[d(\mathrm{y}_1, \mathrm{y}_2)\right] \\ & - 2\mathbb{E}_{\mathrm{x}\sim F,\mathrm{y}\sim G}\left[d(\mathrm{x}, \mathrm{y})\right]. \end{aligned} \tag{41}$$

This closed-form solution (Eq. (40), (41)) corresponds to Eq. (26), (25), enabling efficient computation and enhancing sensitivity to complex distribution shifts in RS segmentation.

Received: 12 November 2025; Accepted: 12 March 2026
Published online: 21 March 2026

References

1. Karan, S. K., Borchsenius, B. T., Debella-Gilo, M. & Rizzi, J. Mapping urban green structures using object-based analysis of satellite imagery: A review. *Ecol. Indic.* **170**, 113027 (2025).
2. Chen, J. et al. Ctseg: Cnn and Vit collaborated segmentation framework for efficient land-use/land-cover mapping with high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **139**, 104546 (2025).
3. Zhu, Q., Weng, N., Fan, L. & Cai, Y. Enhancing environmental monitoring through multispectral imaging: The wastems dataset for semantic segmentation of lakeside waste. In *International Conference on Multimedia Modeling*, 362–372 (Springer, 2025).
4. He, Y., Wang, J., Zhang, Y. & Liao, C. An efficient urban flood mapping framework towards disaster response driven by weakly supervised semantic segmentation with decoupled training samples. *ISPRS J. Photogramm. Remote Sens.* **207**, 338–358 (2024).
5. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019).
6. Ghamisi, P. et al. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **7**, 6–39. https://doi.org/10.1109/MGRS.2018.2890023 (2019).
7. Gómez-Chova, L., Tuia, D., Moser, G. & Camps-Valls, G. Multimodal classification of remote sensing images: A review and future directions. *Proc. IEEE* **103**, 1560–1584. https://doi.org/10.1109/JPROC.2015.2449668 (2015).
8. Xu, Y., Yuan, M., Li, X., Zhang, L. & Zhang, L. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2022).
9. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022 (2021).
10. Lin, A. et al. Ds-transunet: Dual Swin transformer U-net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **71**, 1–15 (2022).
11. Ma, X., Zhang, X., Pun, M.-O. & Liu, M. A multilevel multimodal fusion transformer for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–15. https://doi.org/10.1109/TGRS.2024.3373033 (2024).
12. Yao, J., Zhang, B., Li, C., Hong, D. & Chanussot, J. Extended vision transformer (exvit) for land use and land cover classification: A multimodal deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–15. https://doi.org/10.1109/TGRS.2023.3284671 (2023).
13. Ma, X., Zhang, X. & Pun, M.-O. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **15**, 3463–3474. https://doi.org/10.1109/JSTARS.2022.3165005 (2022).
14. Zhang, X. et al. Cimfnet: Cross-layer interaction and multiscale fusion network for semantic segmentation of high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–16 (2022).
15. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
16. Liu, S. & Chang, L. Conditional dual diffusion for multimodal clustering of optical and SAR images. *IEEE Trans. Circuits Syst. Video Technol.* https://doi.org/10.1109/tcsvt.2025.3533301 (2025).
17. Jiang, F. et al. D3pm: Dual-stream denoising diffusion probabilistic model for change detection in multimodal remote sensing images. *IEEE Transactions on Geosci. Remote. Sens.* (2025).
18. Yue, C. et al. Diffusion mechanism and knowledge distillation object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* https://doi.org/10.1109/tgrs.2025.3561133 (2025).
19. Zhang, W., Mei, J. & Wang, Y. Dmdiff: A dual-branch multimodal conditional guided diffusion model for cloud removal through sar-optical data fusion. *Remote Sens.* **17**, 965 (2025).
20. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–773 (2012).
21. Song, H. & Chen, H. Generalized kernel two-sample tests. *Biometrika* **111**, 755–770. https://doi.org/10.1093/biomet/asad068 (2023).
22. Yan, J. & Zhang, X. Kernel two-sample tests in high dimensions: Interplay between moment discrepancy and dimension-and-sample orders. *Biometrika* **110**, 411–430. https://doi.org/10.1093/biomet/asac049 (2022).
23. Wei, Q. et al. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Trans. Geosci. Remote Sens.* **53**, 3658–3668. https://doi.org/10.1109/TGRS.2014.2381272 (2015).
24. Chen, Y. & Pock, T. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1256–1272. https://doi.org/10.1109/TPAMI.2016.2596743 (2016).
25. Fauvel, M., Tarabalka, Y., Benediktsson, J. A., Chanussot, J. & Tilton, J. C. Advances in spectral-spatial classification of hyperspectral images. *Proc. IEEE* **101**, 652–675. https://doi.org/10.1109/JPROC.2012.2197589 (2013).
26. Song, J., Gao, S., Zhu, Y. & Ma, C. A survey of remote sensing image classification based on CNNs. *Big Earth Data* **3**, 232–254 (2019).
27. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440, https://doi.org/10.1109/CVPR.2015.7298965 (2015).

28. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615 (2017).
29. Sun, W., Tian, Z., Qi, J., Tao, R. & Peng, Y. Maresu-net: A multi-stage attention resu-net for semantic segmentation of high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **18**, 4315–4319. https://doi.org/10.1109/LGRS.2020.3048949 (2021).
30. Jaritz, M., Vu, T.-H., de Charette, R., Wirbel, E. & Perez, P. Cross-modal learning for cross-domain vision-based driving: RGB and lidar. *IEEE Trans. Intell. Transp. Syst.* **22**, 1172–1182. https://doi.org/10.1109/TITS.2020.2991515 (2021).
31. Zhang, Y., Liu, M., He, J., Pan, F. & Guo, Y. Affinity fusion graph-based framework for natural image segmentation. *IEEE Trans. Multimedia* **24**, 440–450. https://doi.org/10.1109/TMM.2021.3053393 (2022).
32. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022 (2021).
33. Wang, L. et al. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **190**, 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008 (2022).
34. Cao, H. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218 (Springer, 2022).
35. Ji, Y., Shi, W., Lei, J. & Ding, J. Dbrsnet: A dual-branch remote sensing image segmentation model based on feature interaction and multi-scale feature fusion. *Sci. Rep.* **15**, 27786 (2025).
36. Roy, S. K. et al. Multimodal fusion transformer for remote sensing image classification. *IEEE Transactions on Geosci. Remote. Sens* **61**, 1–20. https://doi.org/10.1109/TGRS.2023.3286826 (2023).
37. Ma, J. et al. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA J. Autom. Sinica* **9**, 1200–1217. https://doi.org/10.1109/JAS.2022.105686 (2022).
38. Zhu, X. X. et al. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote. Sens. Mag.* **5**, 8–36. https://doi.org/10.1109/MGRS.2017.2762307 (2017).
39. Dhariwal, P. & Nichol, A. Diffusion models beat GANs on image synthesis. *Adv. Neural Inf. Process. Syst.* **34**, 8780–8794 (2021).
40. Riaz, R. et al. A novel ensemble Wasserstein GAN framework for effective anomaly detection in industrial Internet of Things environments. *Sci. Rep.* **15**, 26786 (2025).
41. Ma, X., Huang, Y., Zhang, X., Pun, M.-O. & Huang, B. Cloud-egan: Rethinking cyclegan from a feature enhancement perspective for cloud removal by combining cnn and transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **16**, 4999–5012. https://doi.org/10.1109/JSTARS.2023.3280947 (2023).
42. Wang, L., Xiao, P., Zhang, X. & Chen, X. A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **16**, 4109–4121. https://doi.org/10.1109/JSTARS.2023.3270302 (2023).
43. Ho, J. et al. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research* **23**, 1–33 (2022).
44. Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* **45**, 10850–10869 (2023).
45. Nadjahi, K., Durmus, A., Jacob, P. E., Badeau, R. & Simsekli, U. Fast approximation of the sliced-Wasserstein distance using concentration of random projections. *Adv. Neural Inf. Process. Syst.* **34**, 12411–12424 (2021).
46. Li, R. et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–13. https://doi.org/10.1109/TGRS.2021.3093977 (2022).
47. Li, R. et al. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS J. Photogramm. Remote Sens.* **181**, 84–98. https://doi.org/10.1016/j.isprsjprs.2021.09.005 (2021).
48. Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pspnet, 6230–6239, https://doi.org/10.1109/CVPR.2017.660 (IEEE, 2017).
49. Ma, X., Xu, X., Zhang, X. & Pun, M.-O. Adjacent-scale multimodal fusion networks for semantic segmentation of remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **17**, 20116–20128. https://doi.org/10.1109/JSTARS.2024.3486906 (2024).
50. Amit, T., Shaharbany, T., Nachmani, E. & Wolf, L. SegDiff: Image segmentation with diffusion probabilistic models, https://doi.org/10.48550/arXiv.2112.00390 (2022). arXiv:2112.00390 [cs].
51. Kolbeinsson, B. & Mikolajczyk, K. Multi-class segmentation from aerial views using recursive noise diffusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8439–8449 (2024).

Author contributions

Conceptualization, X.T. and Q.Y.; methodology, X.T.; software, X.T., F.Y. and T.Z.; validation, X.T., F.Y. and W.Q.; formal analysis, X.T. and Q.Y.; investigation, X.T. and F.Y.; resources, H.W. and S.W.; data curation, F.Y. and T.Z.; writing—original draft preparation, X.T.; writing—review and editing, Q.Y., H.W., S.W. and S.L.; visualization, X.T. and T.Z.; supervision, Q.Y., H.W. and S.W.; project administration, Q.Y.; funding acquisition, H.W., S.W. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported in part by the Scientific Research Foundation of Yibin University Grant 2022YY09 and Grant 2025XJKY007.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.Y., H.W. or S.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.