TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2026.9010009

Multimodal 3D Reconstruction of Tidal Flats Using Neural Implicit Representations and Cross-Attention

Ge Huilin, Zhu Zhiyu*, Wang Biao, Liu Runbang, Yang Denghao, and Qiu Zhiwen

Abstract: Tidal flat environments present unique challenges for 3D reconstruction and multimodal data matching due to their complex geological structures. Traditional methods for 3D modeling struggle with sparse feature distributions, leading to inaccuracies in both geometry and appearance. This paper presents a novel algorithm that combines neural implicit representations with multimodal data fusion to address these challenges. We introduce an attention-based strategy for multimodal feature alignment. It effectively aligns fine-grained optical image features with the structural and intensity information derived from Light Detection and Ranging (LIDAR) ambient images. Our method uniquely integrates Self-Distilled Vision Transformer with No Labels v2 (DINOv2)-guided attention for fusing optical and LIDAR data. Additionally, we leverage signed distance field (SDF) priors to explicitly separate geometry and appearance features, enabling precise surface reconstruction. Experimental results demonstrate that our algorithm outperforms state-of-the-art methods, particularly in reconstructing fine details and adapting to dynamic environmental conditions. Future work will explore incorporating point cloud data and temporal modeling to enhance robustness in highly dynamic tidal flat environments.

Key words: tidal flats; 3D reconstruction; multimodal data fusion; neural implicit representations

1 Introduction

Understanding and reconstructing tidal flat [1] environments pose unique challenges due to their dynamic and complex nature, characterized by changing tides, reflective surfaces, and diverse ecological features. Accurate modeling of such environments is crucial for ecological monitoring [2], coastal management, and environmental impact assessments. Traditional 3D reconstruction methods [3] typically rely on a single data type and often struggle with the intricacies of tidal flats, such as sparse feature distributions and low-texture or specular regions. Consequently, these methods are often insufficient for high-precision applications.

Recent advancements in neural implicit representations, such as Neural Radiance Fields (NeRF), have revolutionized 3D reconstruction by leveraging deep learning to represent complex scenes with high fidelity. However, most existing approaches are still designed for a single input modality and do not fully exploit complementary geometric information from other sensors [4]. This single-modality assumption makes it difficult to robustly reconstruct tidal flats, where purely appearance-based cues are often ambiguous or incomplete. In this study, we address these limitations by introducing a novel algorithm that combines NeRF with geometric priors from signed distance fields (SDF) and integrates multimodal data, including optical images from cameras and ambient images from light detection and ranging (LiDAR). To address the ambiguity of data utilization, we

• Huilin Ge and Biao Wang are with Ocean College, Jiangsu University of Science and Technology, Zhenjiang 212003, China. E-mail: ghl1989@just.edu.cn; wangbiao@just.edu.cn.
• Zhiyu Zhu is with Automation College, Jiangsu University of Science and Technology, Zhenjiang 212003, China. E-mail: zhuzy@just.edu.cn.
• Runbang Liu, Denghao Yang, and Zhiwen Qiu are with Automation College, Jiangsu University of Science and Technology, Zhenjiang 212003, China. E-mail: liurunbang212@just.edu.cn; 211210301120@stu.just.edu.cn; 231110302115@stu.just.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2025-11-22; revised: 2025-11-29; accepted: 2025-12-22

explicitly clarify that our method leverages the co-registered ambient images generated by the LiDAR sensor for cross-modal fusion, while the direct utilization of sparse 3D point cloud data remains out of the scope of this work. Our approach is structured around two key innovations. First, we employ implicit neural representations for multiview fusion, allowing for precise geometric surface modeling through enhanced geometric constraints. Second, we propose a cross-attention mechanism guided by DINOv2 features, achieving robust multimodal data matching through selective focus on relevant keypoints while reducing noise from unrelated features. This hybrid strategy allows us to harness the complementary strengths of optical and LiDAR data, effectively overcoming the limitations of single-modality approaches.

We validate our method on a comprehensive tidal flat dataset acquired via Unmanned Aerial Vehicles (UAVs) over intertidal zones in Tasmania, Australia. The results demonstrate significant improvements in reconstruction quality, as measured by Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), over baseline algorithms. Our method also excels in multimodal matching accuracy and robustness under challenging geomorphological and environmental conditions.

In summary, this paper has the following contributions:

(1) We propose a novel algorithm that augments neural implicit representations with SDF-based geometric priors and multimodal data fusion, substantially improving geometric accuracy and stabilizing surface reconstruction in tidal flat scenes, especially in sparse-feature, reflective, and low-texture regions.

(2) We introduce an attention-based multimodal feature alignment strategy, guided by DINOv2, which effectively aligns fine-grained optical features with LiDAR-derived geometric cues, improving cross-modal matching robustness.

(3) We validate our method on real-world tidal flat datasets acquired by UAVs, demonstrating superior performance over state-of-the-art approaches in PSNR, SSIM, and LPIPS, as well as improved robustness in challenging tidal-flat scenarios.

The remainder of this paper is organized as follows: Section 2 reviews the related work on multi-view 3D reconstruction, NeRF, and multimodal feature matching, highlighting the existing challenges. Section 3 details our proposed methodology, including the integration of neural implicit representations, cross-attention-based multimodal alignment, and view fusion strategies for tidal flat reconstruction. Section 4 presents the experimental setup, dataset description, evaluation metrics, and a comparative analysis of the results. Section 5 discusses the limitations of the current approach and its applicability to dynamic and complex tidal flat environments. Finally, potential future directions for enhancing the robustness and adaptability of the proposed framework are outlined.

2 Related Work

2.1 Classical multi-view surface and volumetric reconstruction

In previous work, there exist two main ways to realize 3D reconstruction methods using multiple views: a localized single-view point and surface-based reconstruction [5–8], and a macroscopic volumetric reconstruction [9–11]. The core of point-and-surface reconstruction lies in enforcing photometric consistency across multi-view images to estimate depth from pixel correspondences [6]. This process generates a global dense point cloud [12, 13], which subsequently serves as the basis for surface reconstruction. Comparatively, surface reconstruction is often seen in post-processing optimization using depth-based methods such as Poisson surface reconstruction [14]; however, the quality of matching in such a way determines the effectiveness of the reconstruction. However, the quality of matching is closely related to the texture of the matched objects, and for rivers in tidal flats, the water surface portion is often difficult to match, resulting in a large number of artifacts and flocs in the reconstruction results. Volume Reconstruction Methods (VRMs) are techniques that reconstruct 3D models by evaluating occupancy and color in a voxel mesh from images captured from different viewpoints, without relying on explicit correspondence matching. The complexity associated with traditional image alignment or feature matching is avoided, and better reconstruction results are usually obtained, especially in the case of noisy or incomplete data. However, due to their dependence on voxel resolution, volumetric reconstruction methods perform poorly in refining details and may not be able to recover very small geometric features or high-frequency details.

2.2 Neural radiance fields

NeRF [15] is capable of generating visually realistic images, especially in complex lighting and viewpoint-dependent effects, such as reflections and refractions. Through neural volume rendering, NeRF is able to capture fine details, especially in complex lighting environments, and generate very realistic visual effects. NeRF does not rely on traditional 3D meshes or structured data, but generates scenes directly from spatial locations and volume densities through neural networks, which makes it capable of dealing with much more complex and detail-rich scenes. However, NeRF has a disadvantage [16–19], it lacks an explicit geometric representation, which makes it less suitable for surface reconstruction, measurement, or modification of scene geometry. This is because NeRF does not explicitly model the surface geometry of the scene, but renders it based on volume density. This approach presents challenges in defining equivalent surfaces and often relies on heuristic density thresholding. Recent works, like NeuS [20] and VolSDF [21], integrate SDF with NeRF for surface reconstruction, but they are limited to single-modal data and static scenes. In contrast, our framework uniquely combines SDF priors with cross-modal attention to address dynamic tidal flat environments, where geometric stability and multimodal alignment are critical.

2.3 Feature matching in changing environments

Currently, there are several approaches to deal with outdoor luminance transformations [22–31]. Some of the methods [24,29] are focused on local processing, correcting the lighting to gather objects into the scene, while others directly macro-coordinate the whole scene [23,25–28,30–32]. The algorithm proposed in this paper aims to solve the problem of illumination variations in aerial tidal flat time series. To achieve this, it significantly enhances the generalization capability of existing learning-based matchers by introducing two key innovations: guidance from the underlying scene model and advanced positional encoding mechanisms. Traditionally, when deep learning had not yet emerged, researchers worked on developing local feature models that can be widely applied, such as Scale-Invariant Feature Transform (SIFT) [22], Speeded-Up Robust Features (SURF) [33], and Oriented FAST and Rotated BRIEF (ORB) [34], and these algorithms have been widely used in various image matching tasks. To this day, many computer vision systems still rely on these hand-designed feature extraction methods, especially in downstream applications, such as pose estimation [35–38]. A key reason for the continued use of such traditional approaches is that learning-based methods [39–42], despite excelling in data-rich domains, often exhibit limited cross-domain generalization capability. Recently, the research focus has shifted to developing image matching algorithms that can learn automatically, with some approaches leveraging existing feature extraction tools [43], while others learning feature descriptions and matching relationships through joint learning [44]. While these learning-based matching algorithms have advantages over traditional methods, they tend to limit the generalization of the image matching process and remain biased towards domain-specific applications.

3 Method

3.1 General description of the algorithm

As illustrated in Fig. 1, the algorithmic pipeline follows a sequential workflow. First, we employ Implicit Neural Representations (INR) to achieve multi-view fusion within each modality. In the view fusion stage, we model the camera images and LiDAR ambient-light images separately, and introduce continuity- and smoothness-based geometric priors to enable more accurate surface reconstruction of tidal-flat scenes. Subsequently, these fused representations are utilized in the cross-modal alignment stage to establish correspondences between optical and ambient intensity data, as shown in Fig. 1. In this stage, we align the fine-grained texture features of the optical image with the structured features from the LiDAR ambient-light image, which is achieved through an improved multimodal fusion strategy. Specifically, we employ a multimodal feature alignment module based on an attention mechanism to capture complementary information between the two modalities more effectively and mitigate mismatches caused by modality differences. In the view fusion stage, we use Implicit Neural Representations (INR) to model the camera images and LIDAR ambient light images separately. By introducing continuity and smoothness geometric priors, our algorithm is able to reconstruct the geometric surface of the tidal flat scene more accurately. After modal matching, the fusion result combines the optical appearance with LIDAR-derived geometry, yielding a more accurate and robust 3D reconstruction.

3.2 NeRF

Compared to traditional 3D reconstruction methods, NeRF does not require precise geometric modeling and can directly generate high-quality 3D reconstruction results from a small number of 2D images. NeRF simplifies data acquisition and processing for large, detailed scenes, such as tidal flats. In the 3D reconstruction of tidal flats, NeRF can handle lighting effects such as reflections, transparent objects, and shadows, thus generating more realistic and detailed images, which are especially suitable for variable natural lighting conditions. NeRF, first proposed by Mildenhall et al. [15] in 2020, It is a deep learning-based 3D scene reconstruction method. NeRF is able to generate high-quality images by utilizing neural networks to learn and render complex 3D scenes, and is particularly suitable for scenes with rich details and varying illumination. NeRF's approach is revolutionary in that it compresses the representation of 3D space into an efficient neural network model without relying on complex representations of traditional 3D geometry or multi-view geometry. The core idea of NeRF is to learn the radiance and volume density of the scene through a fully connected neural network, thus enabling modeling of lighting and color at each point in the scene.

The goal of NeRF is to generate an image of the 3D scene. To achieve this, it is necessary to sample along rays from the camera's point of view. Each ray travels from the point of view through varying depths of the scene. For each ray, NeRF samples a number of points along the ray and calculates the volume density and radiance of each point.

The starting point and direction of each ray are known. Let the camera position be denoted as $\boldsymbol{o}$, the origin of the ray, and let $\boldsymbol{d}$ denote its unit direction vector. Sampling is performed along the ray direction, where t represents the depth of the ray. The sampled points along the ray can be expressed as $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$.

NeRF approximates the volume density and radiance of each sampled point using a fully connected neural network. Given a 3D coordinate $\boldsymbol{x}$ and a viewing direction $\boldsymbol{d}$, the network $\mathcal{F}_\theta$ outputs the volume density $\sigma(\boldsymbol{x})$ and the radiance $\boldsymbol{c}(\boldsymbol{x}, \boldsymbol{d})$, can be represented as $\mathcal{F}_\theta(\boldsymbol{x}, \boldsymbol{d}) = \left[\, \sigma(\boldsymbol{x}),\ \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{d}) \,\right]$. Here, θ denotes the learnable parameters of the network, which are optimized through backpropagation.

NeRF calculates the pixel values in the image using a volume rendering formula. Along each ray, originating from the camera's point of view, it passes through different volume elements. Each volume element absorbs and scatters light. By accumulating the effects of these elements, the color along the ray can eventually be computed. The color of the ray is computed as follows: Here, $\boldsymbol{C}(\boldsymbol{r})$ denotes the RGB color vector of the pixel corresponding to ray $\boldsymbol{r}$. The symbol $\boldsymbol{r}$ denotes this camera ray parameterized by $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$. The integration bounds t_n and t_f denote the near and far depth limits of the ray.

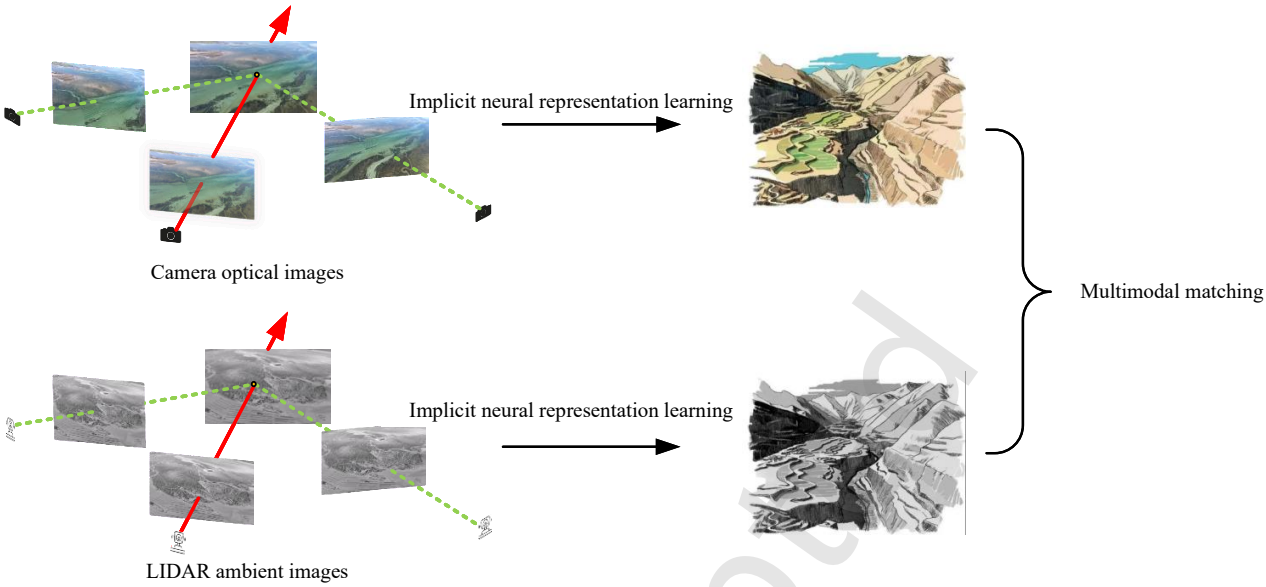

Fig. 1 Overview of the proposed framework. The pipeline proceeds in two logical steps: (1) Camera optical images and LiDAR ambient images are processed via implicit neural representation learning; (2) The learned representations are then utilized in the multimodal matching module to align the two modalities.

$$\boldsymbol{C}(\boldsymbol{r}) = \int_{t_n}^{t_f} T(t)\,\sigma(\boldsymbol{r}(t))\,\boldsymbol{c}(\boldsymbol{r}(t),\boldsymbol{d})\,\mathrm{d}t, \tag{1}$$

Here, $T(t)$ denotes the accumulated transmittance and is defined in Eq. (2).

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(u))\,\mathrm{d}u\right). \tag{2}$$

where u is the dummy integration variable. $\sigma(\boldsymbol{r}(u))$,$\sigma(\boldsymbol{r}(t))$ is the density along the ray.

The above equations can be discretized as follows: Let N_{ray} denote the number of sampled points with indices $i, j \in \{1,\ldots,N_{\text{ray}}\}$, and let $\{t_i\}_{i=1}^{N_{\text{ray}}}$ be the sampled depths along the ray in ascending order. The spacing between consecutive samples is denoted by δ_j; the density at the j-th point is $\sigma_j = \sigma(\boldsymbol{r}(t_j))$; the radiance at the i-th point is $\boldsymbol{c}_i = \boldsymbol{c}(\boldsymbol{r}(t_i),\boldsymbol{d})$. We define the opacity at the j-th point as $\alpha_j = 1-\exp(-\sigma_j\delta_j)$, and T_i represents the cumulative transmittance up to the i-th point.

$$T_i = \exp\Big(-\sum_{j=1}^{i-1}\sigma_j\delta_j\Big) = \prod_{j=1}^{i-1}(1-\alpha_j), \tag{3}$$

$$\widehat{\boldsymbol{C}}(\boldsymbol{r}) = \sum_{i=1}^{N_{\text{ray}}} T_i\,\alpha_i\,\boldsymbol{c}_i. \tag{4}$$

The notation $\widehat{\boldsymbol{C}}$ denotes the discrete approximation of the continuous volume rendering integral given in Eq. (1).

The final rendered color is obtained by integrating the effects of density and radiance along the ray as defined by the above volume rendering formulation.

3.3 Implicit neural representation

Generating neural scene representations with optimized robustness through volume rendering techniques can effectively handle highly complex objects. Despite this, the challenge of extracting high-quality surfaces from implicitly learned representations arises from the insufficient surface constraints that the representation offers, making it difficult to accurately recover surface details and geometry. Specifically, the implicit representation learns the volumetric information of an object primarily by learning its volumetric properties rather than its explicit surface geometry. This makes it difficult to extract clear surfaces directly, especially when high-precision surface modeling is required.

In the process of tidal flat reconstruction, there are challenges posed by water surfaces and wet sand areas. Variations in illumination and reflections significantly impact the accuracy of 3D reconstruction. The brightness of the same location can vary considerably across viewpoints, leading to flocculent artifacts in images generated from new viewpoints.

NeRF incorporates an implicit neural representation for 3D reconstruction, where a neural network learns an implicit geometric representation of the scene or object to achieve high-quality 3D reconstruction. Unlike traditional mesh or point cloud representations, NeRF, combined with an implicit neural representation, does not rely on explicit geometric structures. Instead, it learns the surface and volume information of the object directly through the neural network. By modeling the geometric surface of an object through a neural network and combining it with volumetric rendering techniques, NeRF produces 3D images with both detail and realism.

The core idea of integrating NeRF with implicit neural representation is to model 3D object surfaces through an implicit function, replacing explicit representations like triangular meshes or point clouds. In this paradigm, a neural network encodes the object's geometry by mapping 3D coordinates $\boldsymbol{x}$ to a scalar value

$\varphi(\boldsymbol{x})$. This function defines the surface via its zero level set $\{\boldsymbol{x} : \varphi(\boldsymbol{x}) = 0\}$, serves as a signed-distance estimate so that $|\varphi(\boldsymbol{x})|$ indicates proximity to the surface, and encodes spatial occupancy by sign (negative inside, positive outside, and zero on the surface).

To this end, occlusion-aware and unbiased volumetric rendering techniques are employed. The probability density $\rho(t)$ in the volumetric rendering process is defined as

$$\rho(t) = \max\left(\frac{\frac{\mathrm{d}}{\mathrm{d}t}\,\Phi_s\big(\varphi(\boldsymbol{r}(t))\big)}{\Phi_s\big(\varphi(\boldsymbol{r}(t))\big)},\, 0\right), \tag{5}$$

with the sigmoid function

$$\Phi_s(z) = \frac{1}{1+\mathrm{e}^{-sz}}. \tag{6}$$

Here, $s > 0$ denotes the sharpness or gain parameter of the sigmoid Φ_s, It controls the steepness of the transition. $\rho(t)$ is the probability density along the ray, $\Phi_s(x)$ is the sigmoid function with sharpness parameter s. With the SDF-induced transmittance, the discrete volume rendering takes the form: Following Eqs. (3) and (4), the discrete opacity α_i for $i = 1, \ldots, N_{\text{ray}} - 1$ is given by

$$\alpha_i = \max\left(\frac{\Phi_s(\varphi(\boldsymbol{r}(t_i))) - \Phi_s(\varphi(\boldsymbol{r}(t_{i+1})))}{\Phi_s(\varphi(\boldsymbol{r}(t_i)))},\, 0\right), \tag{7}$$

t_i, t_{i+1} are the depth values of the i-th and $(i+1)$-th sampling points along the ray.

To train the network, we minimize a weighted combination of the RGB color reconstruction loss and the Eikonal regularization term. The total loss function $\mathcal{L}_{\text{total}}$ is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{color}} + \lambda_{\text{eik}}\mathcal{L}_{\text{eik}}, \tag{8}$$

where $\mathcal{L}_{\text{color}} = \sum_{\mathbf{r}\in\mathcal{R}} \|C(\mathbf{r}) - C_{\text{gt}}(\mathbf{r})\|_1$ ensures photometric consistency between rendered and ground truth pixels. The Eikonal term $\mathcal{L}_{\text{eik}} = \sum_{\mathbf{x}} \big(\|\nabla\varphi(\mathbf{x})\|_2 - 1\big)^2$ enforces the SDF property, ensuring smooth and valid geometric surfaces, and λ_{eik} is a hyperparameter balancing the regularization strength.

3.4 Image matching

The input consists of two images with shared content, denoted as I_A and I_B. We refer to the sets of SuperPoint keypoints for these images as $A = \{A_1, A_2, \ldots, A_N\}$ and $B = \{B_1, B_2, \ldots, B_M\}$, where N and M represent the number of keypoints identified in I_A and I_B, respectively. Every keypoint is connected to its SuperPoint local descriptor f. Moreover, we use positional embeddings to encode the normalized positions of the keypoints and refine these using a multilayer perceptron (MLP). The positional features associated with the keypoints are denoted as p.

For each keypoint, we associate a SuperPoint descriptor f and a positional embedding p, which we obtain from normalized image coordinates and then refine with an MLP. In addition, we interpolate the dense DINOv2 feature maps at the SuperPoint locations to obtain a per-keypoint DINOv2 descriptor g. Thus, for image I_A, each keypoint A_i is associated with three features: the SuperPoint descriptor f_i^A, the positional embedding p_i^A, and the DINOv2 descriptor g_i^A. Analogously, for image I_B, each keypoint $B_j \in B$ is associated with the three features f_j^B, p_j^B, and g_j^B.

The dense DINOv2 feature maps induce four keypoint-association graphs: the inter-image graphs $G_{A\to B}$ and $G_{B\to A}$, and the same-image graphs G_A and G_B. As shown in Fig. 2, the inter-image graphs connect keypoints across I_A and I_B. The same-image graphs G_A and G_B are undirected and enable bidirectional message passing within I_A and I_B.

To bootstrap using DINOv2 features, for each keypoint A_i in the set A, the similarity between its DINOv2 descriptor g_i^A and the DINOv2 descriptor g_j^B of all keypoints B_j in the set B is calculated. Before computing similarities, we L2-normalize g_i^A and g_j^B along the channel dimension. The keypoints with the highest similarity are selected to be connected, and only the most likely matching keypoint pairs are kept.

Repeating the above for all keypoints in A constructs $G_{A\to B}$. Symmetrically, repeating over B constructs $G_{B\to A}$.

This multi-level feature extraction and graph construction method, which combines local details (SuperPoint) and global information (DINOv2), makes the matching process more reliable and efficient.

Guided by DINOv2, we suppress distractors by selecting, for each keypoint A_i, those keypoints in B that are most relevant to A_i. Concretely, we compute cosine similarities between the L2-normalized descriptors of A_i and all keypoints in B, and then form a candidate set $S \subseteq B$ for aggregation. Subsequent cross-attention aggregates only over S rather than all of B. Here, S denotes the DINOv2-bootstrapped candidate set, and K denotes the number of selected candidates.

We keep positional and descriptor features separate during message passing: positional features provide spatial context in queries and keys, but are not injected into values.

As shown in Fig. 3, the local descriptor f for each keypoint, such as A_i, is updated in the following way: Let h_i^A denote the current descriptor state initialized from the SuperPoint descriptor f_i^A. The update is

$$h_i^A \leftarrow h_i^A + \mathrm{MLP}\Big(\big[\, h_i^A \mid \Delta h_i^A \,\big]\Big), \tag{8}$$

The query, key, and value used for cross-attention are defined as

$$q_i^A = W^q\Big(h_i^A + p_i^A\Big) + b^q, \tag{9}$$

$$k^S = W^k\Big(h^S + p^S\Big) + b^k, \tag{10}$$

$$v^S = W^v\Big(h^S\Big) + b^v, \tag{11}$$

where h^S and p^S denote the row-stacked descriptor and position matrices of the candidates in S (with $|S| = K$). Here $q_i^A \in \mathbb{R}^D$, $k^S, v^S \in \mathbb{R}^{K\times D}$, and $W^q, W^k, W^v \in \mathbb{R}^{D\times D}$ with biases b^q, b^k, b^v are learned end-to-end (Xavier uniform initialization). The per-head feature dimension is D (default $D = 64$).

Given these, the attention-based update increment is

$$\Delta h_i^A = \mathrm{Softmax}\left(\frac{q_i^A (k^S)^\top}{\sqrt{D}}\right) v^S, \tag{12}$$

where the Softmax is applied row-wise over the K candidates in S (i.e., along the second dimension).

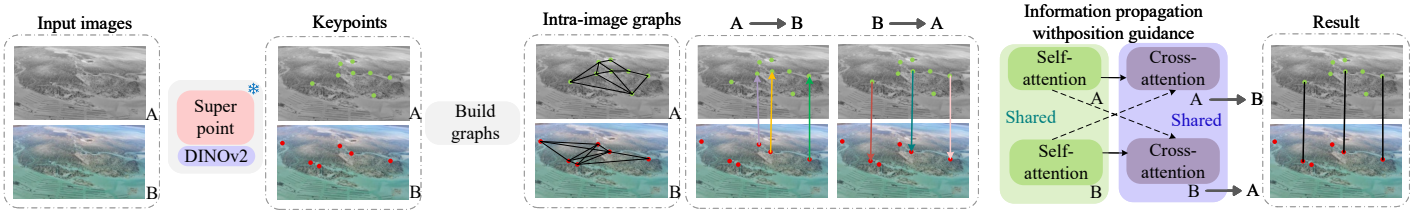

Fig. 2 Overview of the proposed graph based keypoint matching pipeline. SuperPoint keypoints in the two images are enriched with positional embeddings and interpolated DINOv2 descriptors to build graphs within each image and graphs across images, on which position guided self attention and cross attention propagate information to obtain refined, reliable keypoint correspondences.

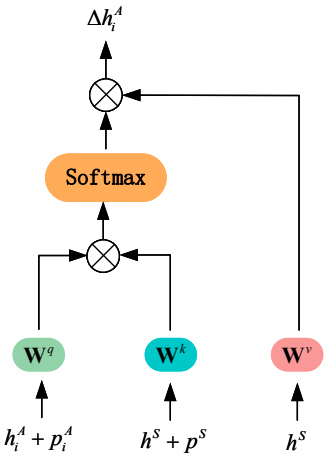

Fig. 3 Schematic of the DINOv2-bootstrapped cross-attention descriptor update with positional separation.

4 Experiments

4.1 Datasets

In this study, we utilized UAV imagery collected from a tidal flat environment through a series of flyovers along a vast stretch of the Australian coastline. The imagery was captured at an altitude of 1000 m over the intertidal zones between Smithton and Woolnorth in northwestern Tasmania. This region is renowned for its rich and diverse intertidal ecosystems, encompassing mudflats, sandflats, salt marshes, and mangrove forests, which provide an ideal setting for assessing the robustness and adaptability of our algorithms in a dynamic, heterogeneous tidal flat environment.

To minimize tidal variability while still capturing realistic illumination changes, images were acquired within a 2-hour window and frames with sudden illumination shifts were excluded. All sensors and observation devices were time-synchronized to a unified reference (UTC) to avoid temporal discrepancies, and repeated flyovers ensured that the same geographic areas were consistently covered. The flight plan included both outbound and return legs so that variations in sun angle and surface reflectance could be systematically incorporated as an additional source of complexity for evaluating algorithm robustness to non-uniform illumination.

The dataset spans a wide range of intertidal ecosystem characteristics, including "tidal trees", estuaries, sediment textures, vegetation zones, and submerged areas. Each scene consists of 30–90 carefully selected images that were pre-processed to standardize resolution and maintain color fidelity. Image calibration was used to correct color deviations and geometric distortions, followed by segmentation and quality filtering to remove unsuitable images and retain only high-quality, representative samples of tidal flat characteristics. As a result, the dataset reflects the diversity of tidal flat types and ecosystems while maintaining a balanced representation across the intertidal zone and sufficient spatial coverage for rigorous validation of our methods.

While the dataset includes common tidal flat types, it does not yet cover more extreme environments such as coral reefs. Future work will extend data collection to these scenarios to further assess the generalizability of our approach.

4.2 Methodology of the evaluation

We evaluate our method along two complementary dimensions: reconstruction quality of rendered views and cross-modal image registration accuracy. Reconstruction quality is assessed using three widely adopted metrics, PSNR, SSIM, and LPIPS, while registration accuracy between aerial RGB images and LiDAR ambient images is measured by the Percentage of Correct Keypoints (PCK). For reconstruction, all metrics are computed per image and then averaged over the test set; higher values indicate better PSNR/SSIM, whereas lower values indicate better LPIPS.

Peak Signal-to-Noise Ratio quantifies the fidelity of a reconstructed image relative to a reference by comparing the maximum possible signal level with the average reconstruction error. The error is summarized by the mean squared error—the spatial average of the squared pixel-wise differences between the reconstructed and reference images. PSNR is reported in decibels (dB); larger values indicate better fidelity. For images with a given bit depth, the maximum possible pixel value associated with that bit depth is used in the PSNR computation.

The Structural Similarity Index measures similarity by jointly assessing luminance, contrast, and structural consistency between a pair of images. Means, variances, and the covariance are computed over local sliding windows, and small positive constants are included to stabilize the ratios when denominators

are close to zero. SSIM ranges from -1 to 1 and is commonly reported between 0 and 1; larger values indicate higher similarity.

The Learned Perceptual Image Patch Similarity compares images in a deep feature space rather than raw pixel space. Features are extracted from multiple layers of a pretrained network; at each layer, channel-wise normalized feature differences are reweighted by learned perceptual weights and averaged spatially. The layer-wise distances are then aggregated to yield a single score. Smaller LPIPS values correspond to higher perceptual similarity.

To quantitatively evaluate image registration between aerial RGB images captured by the onboard camera and ambient-ambient images recorded by the LiDAR sensor, we adopt the PCK as our main metric. In our dataset, we consider pairs of RGB camera images and LiDAR ambient images that exhibit sufficient overlap and noticeable viewpoint changes. For each selected pair, we treat the camera RGB image as the source and the corresponding LiDAR ambient image as the target, and manually annotate a set of visually salient cross-modal correspondences. These annotations are then used to estimate a ground-truth homography with a robust fitting procedure. For evaluation, we place a regular grid of keypoints over the source RGB image so that the points cover the interior of the image rather than being concentrated in a small region, and we do not use these keypoints during registration itself but only for measuring accuracy. After a registration method predicts a homography aligning the RGB image to the LiDAR ambient image, each grid keypoint is warped by both the ground-truth and predicted transforms, and the Euclidean distance between the two warped positions is computed. We then report PCK@1%/3%/5%, counting a keypoint as correct if this distance is below 1%, 3%, or 5% of the image size.

4.3 Experimental Results

We evaluate the proposed method on the collected tidal flats dataset. The experiments consist of two parts: (i) multi-view neural rendering and 3D reconstruction using camera-captured tidal flat images and LiDAR ambient-light images, and (ii) cross-modal matching between camera images and LiDAR ambient-light images. For reconstruction quality, we report PSNR, SSIM, and LPIPS; for cross-modal matching, we adopt PCK@1%, PCK@3%, and PCK@5% as the main metrics.

Nerfacto is a NeRF-based optimization method that integrates camera pose optimization, per-image appearance conditioning, proposal sampling, scene contraction, and hash encoding, and has shown strong performance in multi-view rendering and 3D reconstruction of static scenes. Our algorithm is built on Nerfacto but specifically enhances multimodal data alignment and the ability to capture fine details, targeting the challenges of cross-modal fusion and dynamic environmental variations in complex tidal flat scenes.

Figure 4 compares the reconstruction results of our method and Nerfacto on representative tidal flat views. The zoomed-in regions show that our method reconstructs sharper textures, especially on tidal tree branches and mudflat cracks, where Nerfacto tends to over-smooth structures and lose high-frequency details. By contrast, our model benefits from SDF priors and cross-modal alignment, which help preserve clear edges and subtle geometric cues. From the visual results, we observe that the baseline often fakes reflections by creating floating shapes. In contrast, our method enforces a strict rule that surfaces must be solid. This forces the network to model reflections by changing the color on the surface, rather than changing the shape of the surface itself.

Quantitative reconstruction results are summarized in Table 1. Our method improves PSNR in four out of five scenarios and increases the average PSNR from 22.55 dB to 24.08 dB. Gains are particularly notable for Ground Textures and Deep-Water Areas, where the combination of SDF priors and LiDAR fusion facilitates recovery of high-frequency details under low-texture or strongly reflective conditions. The average SSIM also increases slightly, with small drops for Tidal Trees and River Mouths, but consistent improvements on the other three scenarios. LPIPS decreases from 0.334 to 0.227 on average, indicating that our reconstructions are perceptually closer to the reference views. The only scenario where Nerfacto still clearly dominates is River Mouths, where it achieves higher PSNR, higher SSIM, and lower LPIPS, likely due to pronounced cross-modal discrepancies between RGB appearance and LiDAR ambient intensity in this region.

We further evaluate cross-modal matching between LiDAR ambient-light images and camera images. First, we perform multi-view fusion and alignment across the two modalities; an example of the fused and matched views is shown in Fig. 5. Since visual differences in stitched mosaics are not always obvious, we additionally visualize feature correspondences in Fig. 6. Green lines denote correct matches between LiDAR (left) and camera (right) images under large viewpoint gaps. Candidate matches are filtered using DINOv2-guided attention weights and RANSAC-based geometric verification.

PCK@1%, PCK@3%, and PCK@5% measure the fraction of correspondences whose reprojection error lies within 1%, 3%, and 5% of the image size, respectively. As reported in Table 2, when the error threshold is very small, SFD2 [45], XoFTR [46], DeDoDe [39], LoFTR [47], Alike [48], and our method all achieve comparable matching accuracy. As the threshold becomes looser, our method maintains consistently higher PCK values and can still establish a large number of correct correspondences, indicating stronger robustness under large viewpoint changes and cross-modal discrepancies.

Regarding computational efficiency, our method incurs a moderate increase in cost compared to Nerfacto due to the additional overhead from the cross-attention mechanism and SDF-based geometric regularization. However, considering the significant improvements in geometric accuracy and multimodal alignment shown in Fig. 4 and Table 1, this trade-off is acceptable for high-precision tidal flat monitoring where reconstruction quality is prioritized over real-time performance. Overall, our method achieves comparable or slightly higher average PSNR and SSIM than Nerfacto and significantly lower average LPIPS, while also delivering more robust cross-modal matching performance than existing feature-matching baselines. These results collectively demonstrate the practicality and reliability of our approach for multimodal fusion and 3D

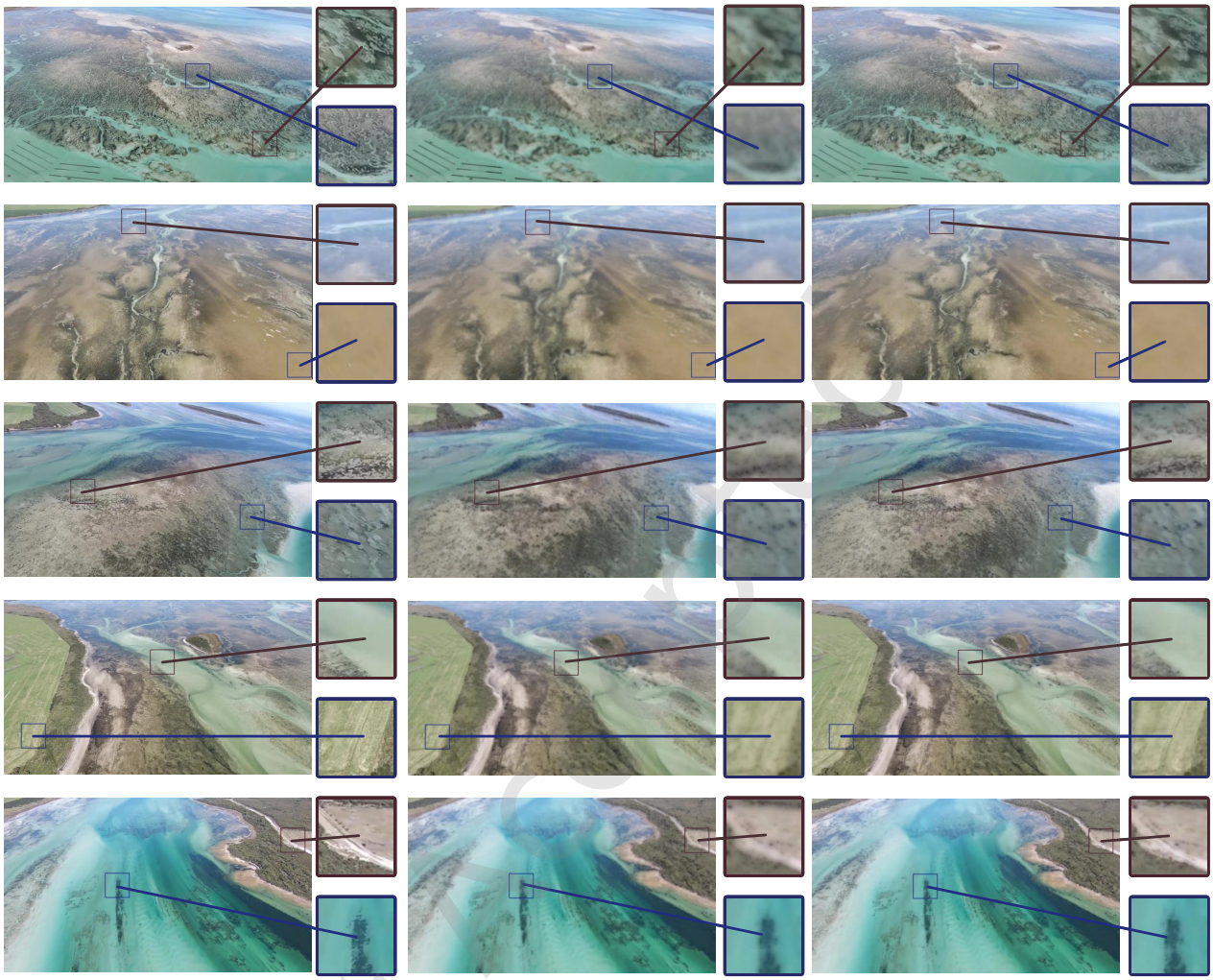

Fig. 4 Comparison of reconstruction results between Nerfacto and our method on tidal flat scenes. Zoomed-in regions highlight differences in texture sharpness and fine details.

reconstruction in challenging tidal flat environments.

4.4 Ablation studies

To better understand the contribution of each component in our framework, we conduct ablation studies from two perspectives: (i) the effect of SDF priors and multimodal fusion on reconstruction quality, and (ii) the effect of the DINOv2-guided graph design on cross-modal image matching.

4.4.1 Effect of SDF priors and multimodal fusion

We first ablate the reconstruction pipeline by progressively enabling the SDF prior and multimodal LiDAR fusion, while keeping the underlying Nerfacto backbone fixed. We consider four variants: (a) Nerfacto, which is the original Nerfacto implementation without any SDF prior or multimodal fusion; (b) Nerfacto and SDF, which introduces the SDF-based implicit geometric prior on top of Nerfacto, but still trains on monocular RGB images only, without using LiDAR information; (c) Nerfacto and LiDAR, which fuses LiDAR ambient images with Nerfacto in the radiance field, but does not enforce the SDF prior, thereby isolating the effect of multimodal cues without explicit geometric regularization; and (d) Full, which is our full reconstruction model and combines SDF-based implicit geometry with cross-modal fusion of RGB and LiDAR ambient images. We report PSNR, SSIM, and LPIPS, averaged over all test views and scenes, using the same experimental protocol as in Table 1. The results are summarized in Table 3.

Overall, the ablation results in Table 3 confirm the complementary effects of SDF priors and multimodal fusion. Compared with the Nerfacto baseline, introducing the SDF prior consistently improves reconstruction quality, yielding higher PSNR and SSIM and a markedly lower LPIPS, indicating more accurate and perceptually plausible geometry. Incorporating LiDAR alone brings further gains in PSNR and LPIPS, but slightly degrades SSIM, suggesting that multimodal cues

Table 1 Comparison of PSNR, SSIM, and LPIPS between Nerfacto and our method on tidal flat scenarios.

Scenario	PSNR		SSIM		LPIPS	
	Nerfacto	Ours	Nerfacto	Ours	Nerfacto	Ours
Tidal Trees	22.13	24.96	0.593	0.585	0.328	0.193
River Mouths	24.08	23.92	0.567	0.555	0.211	0.322
Ground Textures	21.92	23.81	0.506	0.539	0.238	0.158
Vegetation	23.76	24.47	0.554	0.571	0.322	0.149
Deep-Water Areas	20.87	23.22	0.525	0.542	0.569	0.246
Average	22.55	24.08	0.549	0.558	0.334	0.227

Fig. 5 Multi-view fusion and matching results for a camera image (left) and a LiDAR ambient-light image (right) in a tidal flats scene.

Table 2 Relative attitude estimation performance on the tidal flats dataset. PCK@τ is reported in %.

Method	PCK@1%	PCK@3%	PCK@5%
SFD2	14.5	19.1	23.3
XoFTR	17.0	25.6	33.0
DeDoDe	17.2	25.0	32.6
LoFTR	12.8	19.2	21.5
Alike	15.0	27.4	28.8
Ours	17.6	30.6	34.2

Table 3 Ablation on SDF priors and multimodal fusion. PSNR (dB), SSIM, and LPIPS are averaged over all tidal-flat scenes. Higher PSNR/SSIM and lower LPIPS are better.

Model	PSNR	SSIM	LPIPS
(a) Nerfacto (Baseline)	22.55	0.549	0.334
(b) Nerfacto + SDF	23.26	0.632	0.243
(c) Nerfacto + LiDAR	23.65	0.525	0.278
(d) Full	24.08	0.558	0.227

without explicit geometric regularization can introduce structural inconsistencies. Our full model, which jointly leverages SDF-based geometry and RGB–LiDAR fusion, achieves the best overall performance across all three metrics, demonstrating that explicit geometric priors and multimodal information are both crucial for high-fidelity tidal-flat reconstruction.

4.4.2 Effect of semantic backbone on cross-modal matching

We further ablate the choice of semantic backbone that provides dense semantic descriptors at keypoint locations. In this experiment, the detector, positional embeddings, graph structure, and all training settings are kept fixed, and only the backbone used to extract semantic features is changed.

The variant None disables the semantic backbone entirely and relies solely on SuperPoint descriptors and positional embeddings for each keypoint. ResNet replaces DINOv2 with a conventional convolutional backbone; intermediate feature maps from a ResNet are sampled at the keypoint positions. SupViT instead uses a supervised vision transformer, where dense transformer features are interpolated at the keypoints. Finally, DINOv2 (ours) corresponds to our default setting, in which dense DINOv2 features guide the graph-based matcher.

For all variants, we evaluate relative pose estimation and registration quality on the tidal-flat dataset using PCK@1%, PCK@3%, and PCK@5%, following the protocol in Table 2. The results are reported in Table 4.

The ablation in Table 4 highlights the importance of a strong semantic backbone for cross-modal matching. Removing semantic descriptors (None) yields the lowest PCK across all thresholds, indicating that SuperPoint and positional cues alone

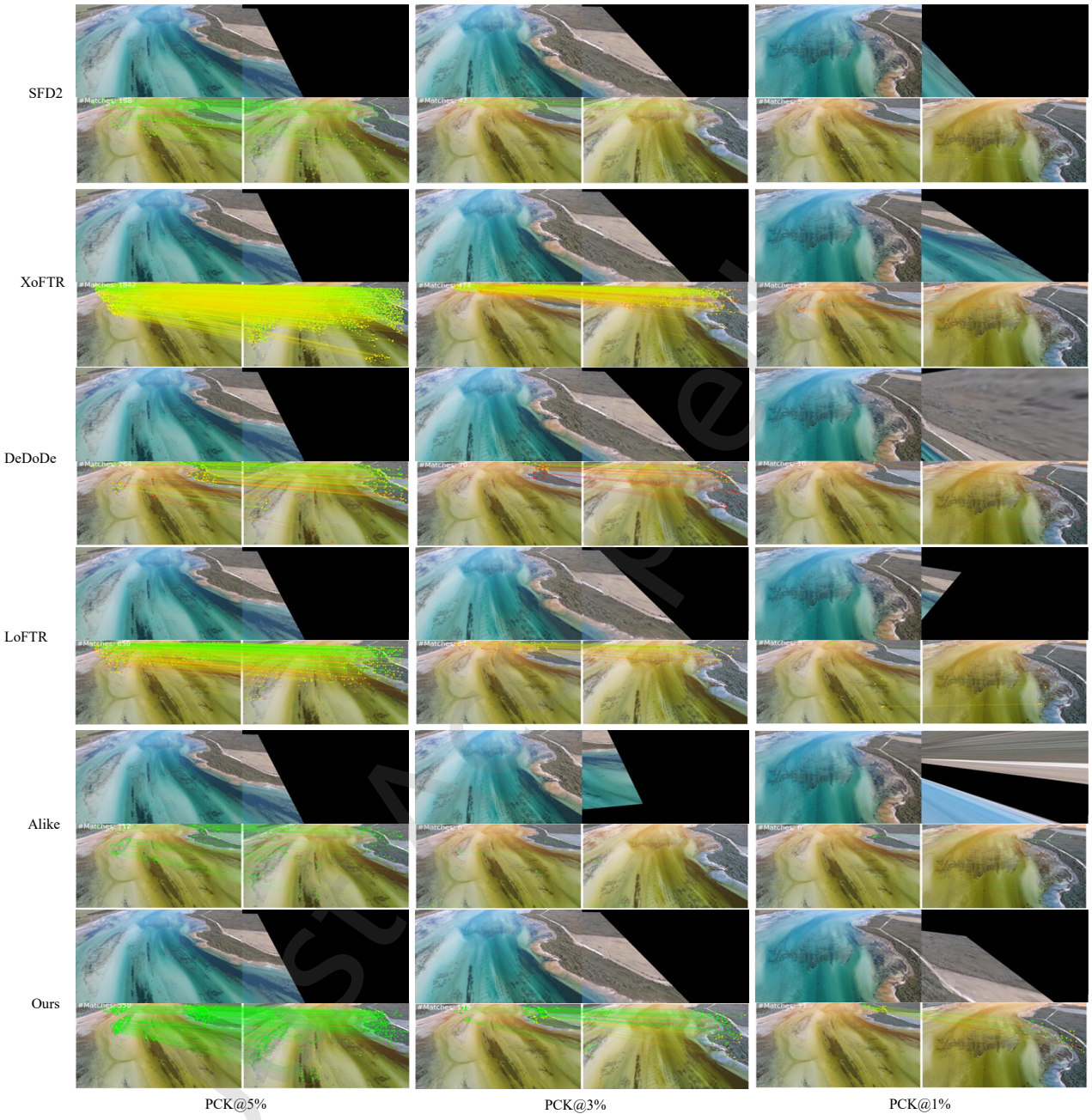

Fig. 6 Viewpoint alignment results between LiDAR ambient-light images and camera images, with matching accuracy evaluated at PCK@5% (left), PCK@3% (middle), and PCK@1% (right). Green lines indicate correct cross-modal matches.

Table 4 Ablation on the semantic backbone used to provide dense descriptors. We report the PCK at different error thresholds.

Variant	PCK@1%	PCK@3%	PCK@5%
None	10.4	18.7	23.9
ResNet	12.9	24.8	26.7
SupViT	15.3	27.9	32.1
DINOv2 (ours)	17.6	30.6	34.2

are insufficient for reliable registration in tidal-flat scenes. Introducing convolutional semantics with ResNet leads to noticeable gains, particularly at PCK@3%, while replacing it with a supervised transformer (SupViT) further improves matching accuracy. Our full setting with DINOv2 achieves the highest PCK at all thresholds, demonstrating that self-supervised transformer features provide more discriminative and robust dense descriptors for cross-modal matching than both convolutional and supervised transformer baselines.

5 Conclusions

Similar to many neural implicit reconstruction methods, our approach is limited by the number of available images. Dense observations from multiple viewpoints are essential for accurate reconstruction of objects. This reliance on extensive multi-view data can be a significant limitation in cases where such comprehensive coverage is not feasible. In addition, our method, while effective in detailed surface reconstruction, may encounter difficulties in environments with sparse or inhomogeneous image distributions, potentially resulting in less accurate reconstruction of these regions. Tidal flats environments are dynamic in nature, e.g., tidal variations, erratic lighting conditions, and reflections from the water column. Our method currently assumes static scenes within short time windows. Handling large-scale tidal dynamics such as hourly water level changes requires temporal modeling beyond the scope of this work. Unlike NeRF++ [19] and Mip-NeRF [16], which focus on anti-aliasing and multi-scale representations but lack explicit mechanisms for multimodal alignment, our method uniquely integrates DINOv2-guided cross-attention to fuse optical and LIDAR data. Furthermore, we leverage SDF priors to disentangle geometry and appearance, a capability absent in existing NeRF frameworks. These dynamics may degrade the accuracy of our multimodal matching. Our method's generalization capability has not been explicitly optimized for such temporal and environmental variations. Tidal flats environments have a large number of complex terrain structures such as dunes, mudflats, vegetation, and these terrain features may have subtle texture variations and complex geometric relationships. Although our dataset covers mudflats, sandflats, and mangroves in Tasmania, future work will validate the method on tidal flats with extreme terrains to test generalizability. The attention module, while solving part of the positional entanglement problem, is not fully optimized for the unique geometric characteristics of tidal flats and may not be able to capture these complex features. In tidal flats regions with homogeneous textures, the attention module assigns uniform weights to ambiguous features, leading to suboptimal matches. Integrating elevation gradients or sediment type priors could mitigate this issue. For example, in mudflat regions with homogeneous textures, the attention module may assign uniform weights to ambiguous features. Future work will integrate elevation gradients as terrain priors to resolve this.

6 Future work

For the common multimodal data in tidal flats scenes, we mainly utilize the ambient light images and camera images from LIDAR, and the multimodal combining methods for point cloud data will be further explored in the subsequent work, although we use neural implicit reconstruction methods, which have geometric a priori with respect to traditional NeRF, and combining with NeRF can provide the models with provide stronger geometric constraints. A clear separation between geometry and appearance can be made. Further combining LIDAR data can provide highly accurate depth information and 3D geometry, which is particularly critical in tidal flats scenarios due to the complexity of the tidal flats terrain with a large number of subtle texture variations and geometric features. By combining LIDAR point cloud data with optical imagery, the limitations of single-modal data in certain situations, such as illumination variations or occlusion issues, can be effectively compensated. After combining LIDAR data, the geometric a priori of the point cloud can further enhance the modeling capability of the SDF and provide a more accurate geometric representation of the model. Meanwhile, the rich texture and color information in the optical image is used for appearance modeling through NeRF, and the combination of the two can achieve highly accurate multimodal reconstruction. Specifically, the LIDAR data can guide the generation of neural implicit surfaces through the sparse point cloud to complement the geometric details in the sparse region of the point cloud, while the optical image can provide more delicate appearance information, thus realizing a high degree of unity between geometry and appearance. Combining the topographic characteristics of tidal flats, geometrically constrained models or 3D feature enhancement modules are introduced to improve the understanding and matching ability of the complex topography of tidal flats. For dynamic scenarios such as tides and light changes, explore time series models or dynamic attention mechanisms to enhance the model's ability to adapt to the dynamic environment of tidal flats. For low-contrast or sparse texture regions, develop specific feature enhancement modules, such as combining contrast learning or Generative Adversarial Networks to generate richer feature representations, to improve the matching ability in sparse scenes.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 62576157), the National Natural Science Foundation of China (Grant No. U25A20442), the National Natural Science Foundation of China (Grant No. 62476113), the Natural Science Foundation of Jiangsu Province of China (Grant No. BK20253020).

References

[1] N. J. Murray, S. R. Phinn, M. DeWitt, R. Ferrari, R. Johnston, M. B. Lyons, N. Clinton, D. Thau, and R. A. Fuller, The global distribution and trajectory of tidal flats, Nature, vol. 565, no. 7738, pp. 222–225, 2019.

[2] D. B. Lindenmayer and G. E. Likens, The science and application of ecological monitoring, Biol. Conserv., vol. 143, no. 6, pp. 1317–1328, 2010.

[3] Z. Ma and S. Liu, A review of 3D reconstruction techniques in civil engineering and their applications, Adv. Eng. Inf., vol. 37, pp. 163–174, 2018.

[4] D. Lahat, T. Adali, and C. Jutten, Multimodal data fusion: An overview of methods, challenges, and prospects, Proc. IEEE, vol. 103, no. 9, pp. 1449–1477, 2015.

[5] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, PatchMatch: A randomized correspondence algorithm for structural image editing, ACM Trans. Graph., vol. 28, no. 3, p. 24, 2009.

[6] Y. Furukawa and J. Ponce, Accurate, dense, and robust multiview stereopsis, IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 8, pp. 1362–1376, 2010.

[7] S. Galliani, K. Lasinger, and K. Schindler, Gipuma: Massively parallel multi-view stereo reconstruction, Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V, vol. 25, pp. 361–369, 2016.

[8] J. L. Schönberger, E. Zheng, J. M. Frahm, and M. Pollefeys, Pixelwise view selection for unstructured multi-view stereo, in Proc. 14th European Conf. Computer Vision, Amsterdam, The Netherlands, 2016, pp. 501–518.

[9] J. S. De Bonet and P. Viola, Poxels: Probabilistic voxelized volume reconstruction, in Proc. 7th Int. Conf. Computer Vision (ICCV), Corfu, Greece, 1999, pp. 418–425.

[10] A. Broadhurst, T. W. Drummond, and R. Cipolla, A probabilistic framework for space carving, in Proc. 8th IEEE Int. Conf. Computer Vision, Vancouver, Canada, 2001, pp. 388–393.

[11] S. M. Seitz and C. R. Dyer, Photorealistic scene reconstruction by voxel coloring, Int. J. Comput. Vis., vol. 35, no. 2, pp. 151–173, 1999.

[12] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J. M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, Real-time visibility-based fusion of depth maps, in Proc. 11th IEEE Int. Conf. Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[13] C. Zach, T. Pock, and H. Bischof, A globally optimal algorithm for robust TV-L1 range image integration, in Proc. 11th IEEE Int. Conf. Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[14] M. Kazhdan and H. Hoppe, Screened Poisson surface reconstruction, ACM Trans. Graph., vol. 32, no. 3, p. 29, 2013.

[15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, NeRF: Representing scenes as neural radiance fields for view synthesis, Commun. ACM, vol. 65, no. 1, pp. 99–106, 2021.

[16] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields, in Proc. 2021 IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, Canada, 2021, pp. 5835–5844.

[17] C. Sun, M. Sun, and H. T. Chen, Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction, in Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 5449–5459.

[18] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, Plenoxels: Radiance fields without neural networks, in Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022, pp. 5491–5500.

[19] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, NeRF++: Analyzing and improving neural radiance fields, arXiv preprint arXiv: 2010.07492, 2020.

[20] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction, in Proc. 35th Int. Conf. Neural Information Processing Systems, Virtual Event, 2021, p. 2081.

[21] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, Volume rendering of neural implicit surfaces, in Proc. 35th Int. Conf. Neural Information Processing Systems, Virtual Event, 2021, p. 367.

[22] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[23] S. Duchêne, C. Riant, G. Chaurasia, J. L. Moreno, P. Y. Laffont, S. Popov, A. Bousseau, and G. Drettakis, Multiview intrinsic images of outdoors scenes with an application to relighting, ACM Trans. Graph., vol. 34, no. 5, p. 164, 2015.

[24] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem, Rendering synthetic objects into legacy photographs, ACM Trans. Graph., vol. 30, no. 6, pp. 1–12, 2011.

[25] P. Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis, Coherent intrinsic images from photo collections, ACM Trans. Graph., vol. 31, no. 6, p. 202, 2012.

[26] J. F. Lalonde, A. A. Efros, and S. G. Narasimhan, Webcam clip art: Appearance and illuminant transfer from time-lapse sequences, ACM Trans. Graph., vol. 28, no. 5, pp. 1–10, 2009.

[27] J. Philip, M. Gharbi, T. Zhou, A. A. Efros, and G. Drettakis, Multi-view relighting using a geometry-aware network, ACM Trans. Graph., vol. 38, no. 4, p. 78, 2019.

[28] K. Sunkavalli, W. Matusik, H. Pfister, and S. Rusinkiewicz, Factored time-lapse video, ACM Trans. Graph., vol. 26, no. 3, p. 101, 2007.

[29] G. Xing, X. Zhou, Q. Peng, Y. Liu, and X. Qin, Lighting simulation of augmented outdoor scene based on a legacy photograph, Comput. Graphics Forum, vol. 32, no. 7, pp. 101–110, 2013.

[30] Y. Yu, A. Meka, M. Elgharib, H. P. Seidel, C. Theobalt, and W. A. P. Smith, Self-supervised outdoor scene relighting, in Proc. 16th European Conf. Computer Vision, Glasgow, UK, 2020, pp. 84–101.

[31] Y. Yu and W. A. P. Smith, InverseRenderNet: Learning single image inverse rendering, in Proc. 2019 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 3150–3159.

[32] J. T. Barron and J. Malik, Shape, illumination, and reflectance from shading, IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 8, pp. 1670–1687, 2015.

[33] H. Bay, T. Tuytelaars, and L. Van Gool, SURF: Speeded up robust features, in Proc. 9th European Conf. Computer Vision, Graz, Austria, 2006, pp. 404–417.

[34] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, ORB: An efficient alternative to SIFT or SURF, in Proc. 2011 Int. Conf. Computer Vision (ICCV), Barcelona, Spain, 2011, pp. 2564–2571.

[35] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, Zip-NeRF: Anti-aliased grid-based neural radiance fields, in Proc. 2023 IEEE/CVF Int. Conf. Computer Vision (ICCV), Paris, France, 2023, pp. 19640–19648.

[36] M. Rünz, K. Li, M. Tang, L. Ma, C. Kong, T. Schmidt, I. Reid, L. Agapito, J. Straub, S. Lovegrove, et al., FroDO: From detections to 3D objects, in Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 14708–14717.

[37] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, NeRF: Representing scenes as neural radiance fields for view synthesis, in Proc. 16th European Conf. Computer Vision, Glasgow, UK, 2020, pp. 405–421.

[38] M. J. Tyszkiewicz, K. K. Maninis, S. Popov, and V. Ferrari, RayTran: 3D pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers, in Proc. 17th European Conf. Computer Vision, Tel Aviv, Israel, 2022, pp. 211–228.

[39] J. Edstedt, G. Bökman, M. Wadenbäck, and M. Felsberg, DeDoDe: Detect, don't describe—describe, don't detect for local feature matching, in Proc. 2024 Int. Conf. 3D Vision (3DV), Davos, Switzerland, 2024, pp. 148–157.

[40] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, Large-scale image retrieval with attentive deep local features, in Proc. 2017 IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 3476–3485.

[41] Y. Ono, E. Trulls, P. Fua, and K. M. Yi, LF-Net: Learning local features from images, in Proc. 32nd Int. Conf. Neural Information Processing Systems, Montréal, Canada, 2018, pp. 6237–6247.

[42] J. Revaud, P. Weinzaepfel, C. De Souza, and M. Humenberger, R2D2: Repeatable and reliable detector and descriptor, in Proc. 33rd Int. Conf. Neural Information Processing Systems, Vancouver, Canada, 2019, p. 1113.

[43] M. J. Tyszkiewicz, P. Fua, and E. Trulls, Disk: Learning local features with policy gradient, in Proc. 34th Int. Conf. Neural Information Processing Systems, Vancouver, BC, Canada, 2020, p. 1195.

[44] D. DeTone, T. Malisiewicz, and A. Rabinovich, SuperPoint: Self-supervised interest point detection and description, in Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018, pp. 224–236.

[45] F. Xue, I. Budvytis, and R. Cipolla, SFD2: Semantic-guided feature detection and description, in Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023, pp. 5206–5216.

[46] Ö. Tuzcuoğlu, A. Köksal, B. Sofu, S. Kalkan, and A. A. Alatan, XoFTR: Cross-modal feature matching transformer, in Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPR), Seattle, WA, USA, 2024, pp. 4275–4286.

[47] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, LoFTR: Detector-free local feature matching with transformers, in Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 8918–8927.

[48] X. Zhao, X. Wu, J. Miao, W. Chen, P. C. Y. Chen, and Z. Li, ALIKE: Accurate and lightweight keypoint detection and descriptor extraction, IEEE Trans. Multimedia, vol. 25, pp. 3101–3112, 2023.

Huilin Ge Dr. Huilin Ge received his Ph.D. degree in Naval and Ocean Engineering from Jiangsu University of Science and Technology in 2023. He is currently a dedicated researcher and educator affiliated with Jiangsu University of Science and Technology in Zhenjiang, China. His research interests are centered on the fields of Deep Learning, Underwater Information Perception, and Artificial Intelligence.

Zhiyu Zhu received the M.S. degree in control theory and control engineering from the Automation School of Nanjing University of Science and Technology (NUST), Nanjing, China, in 2001, and the Ph.D. degree in control theory and control engineering from the Automation School of Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2006. He has been teaching at Jiangsu University of Science and Technology since his graduation, where he now serves as a doctoral advisor. Main research interests include intelligent control systems, radar signal processing, and ship power system control.

Biao Wang received a Ph.D. degree in signal and information processing from the Institute of Acoustics, Chinese Academy of Sciences (IACAS), Beijing, China, in 2009. Since 2019, he has been a professor with the Ocean College, Jiangsu University of Science and Technology, Zhenjiang, China. His research interests are underwater acoustic array signal processing, underwater acoustic communication, and underwater acoustic sensor networks.

Runbang Liu received his M.S. degree in Control Science and Engineering from Jiangsu University of Science and Technology, Jiangsu, China, in 2018. He joined Jiangsu University of Science and Technology in 2018. He is now a Ph.D student of Naval and Ocean Engineering at Jiangsu University of Science and Technology. His research interests include control algorithms, image processing, artificial intelligence, target detection, and tracking.

Denghao Yang obtained his bachelor's degree in electrical engineering and Automation from Qingdao Agricultural University in Shandong, China, in 2021. He is currently pursuing a Ph.D. in Systems Science at Jiangsu University of Science and Technology. His research interests include remote sensing image processing, tidal flat feature identification, and biomass prediction.

Zhiwen Qiu was born on May 14th, 2000, in Zhejiang. He received the Bachelor's degree in Automation from Chengdu University of Information Technology. He is currently pursuing a master's degree in Control Science and Engineering at Jiangsu University of Science and Technology. His main research direction is Artificial Intelligence.