This is the Pre-Published Version.
The following publication Z. Zhang, L. Fang, Z. Yan, T. Chen, B. Wang and C. -y. Wen, "Spatial–Temporal Diffusion Model for Underwater Scene Reconstruction With Application to AUV Navigation," in IEEE/ASME Transactions on Mechatronics, vol. 30, no. 6, pp. 4142-4153, Dec. 2025 is available at https://doi.org/10.1109/TMECH.2025.3600436.

Spatial-temporal Diffusion Model for Underwater Scene Reconstruction with Application to AUV Navigation

Zhengyan Zhang[1,2,3†], Liang Fang[4†], Zheping Yan[4,5], Tao Chen[4,5], Bing Wang[1,2⋆], Chih-yung Wen[1,2,3]

***Abstract*—Autonomous Underwater Vehicles (AUVs) have been extensively utilized in subsea exploration and surveying. However, accurately perceiving the surrounding environment remains a significant challenge for AUVs due to the complexities of subsea terrains. To address this issue, we propose a novel generative scene reconstruction method to enhance AUVs' perception capabilities. Our method is primarily designed for reconstructing dense subsea terrain from 3D multibeam echosounder (MBES) data. We leverage local diffusion and denoising strategies to reconstruct complete subsea terrain at the scene scale directly, without requiring normalization from point clouds. Considering the motion dynamics of AUVs and the overlap between consecutive sonar frames, we introduce a spatial-temporal attention mechanism to aggregate features from consecutive point clouds and guide the reconstruction process as a condition. Then, the reconstructed point cloud is utilized for probabilistic terrain modeling through Bayesian updating, enabling path planning. Experiments conducted on simulation and real-world datasets demonstrate that our method can generate more accurate and complete terrain maps. Furthermore, path planning based on our reconstruction method achieves the shortest and smoothest motion path, further validating that our reconstruction method can provide more complete perception information for AUV navigation. The code of this work is available at https://github.com/sam-zyzhang/SonarPC-Diff.git.**

***Index Terms*—Scene Reconstruction, Diffusion Model, Subsea Terrain Perception, Unmanned Underwater Vehicle.**

I. Introduction

THE perception system is a crucial part of autonomous underwater vehicles (AUVs), enabling them to understand their surroundings and identify traversable areas within the environment [1], [2]. In most scenarios for deep-sea exploration, due to constraints such as water turbidity, light absorption, and limited field of view, optical sensors like cameras and laser range scanners are unable to provide accurate environmental perception information for AUVs [3], [4]. Thus, acoustic sensors, particularly 3D MBES sonars, are better suited for underwater exploration tasks, including environmental perception, mapping, and localization [5]. 3D MBES sonars generate point clouds with a fixed number of points during each ping, with a detection range of tens of meters. This characteristic results in a sparse point cloud collection, hindering the perception system from inferring comprehensive environmental information. Reconstructing dense point cloud scenes from the sparse data can provide richer information for AUV's perception systems, enhancing the capabilities of downstream tasks such as subsea terrain mapping and navigation.

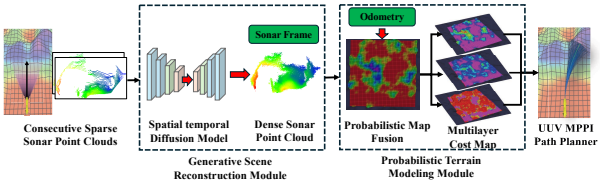

Fig. 1: Overview of the subsea terrain perception framework.

Scene reconstruction aims to obtain the full and dense 3D information of the surrounding environment, given a partial and sparse sensor measurement. Due to the sparsity of 3D MBES data, constructing dense and complete point clouds through scene reconstruction helps augment sparse input point clouds, enhancing the perception capability of AUVs. Previously, this technology was widely applied in the field of autonomous driving [6], [7]. These methods approximate dense scenes by inferring depth maps through the fusion of paired images and light detection and ranging (LiDAR) data or by using signed distance fields for surface representation. However, due to image distortion in the underwater environment and the high noise in sonar data, it is challenging to perform scene reconstruction through multimodal data fusion or surface fitting methods.

To address the challenges of scene reconstruction in underwater exploration for AUVs, this paper adopts the diffusion model to reconstruct a dense and complete point cloud from the sparse 3D MBES data. We propose a scene reconstruction network based on the local denoising diffusion probabilistic models [8], which enables the network to learn scene features directly. Considering the slow sailing speed of AUVs, there is a significant overlap between consecutive sonar point cloud frames. The spatio-temporal attention mechanism is introduced

This work was jointly supported by the Young Scientists Fund of the National Natural Science Foundation of China (42301520), the Major Research Project on Scientific Instrument Development of National Natural Science Foundation of China (42327901), the Research Grants Council of Hong Kong (25206524), the Innovation and Technology Fund (PRP/068/23FX), the Platform Project of Unmanned Autonomous Systems Research Centre (P0049516), Guangdong-Hong Kong Joint Laboratory for Marine Infrastructure (2025B1212150001), the Seed Projects of Smart Cities Research Institute (P0051028, P0054511).

[1,2]Zhengyan Zhang, Bing Wang, and Chih-yung Wen are with the Department of Aeronautical and Aviation Engineering and also with the Research Centre of Unmanned Autonomous Systems, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, 999077, China.

[3]Zhengyan Zhang and Chih-yung Wen are with the Guangdong-Hong Kong Joint Laboratory for Marine Infrastructure, Hong Kong, China.

[4]Liang Fang, Zheping Yan, and Tao Chen are with the School of Intelligent Science and Engineering, Harbin Engineering University, China.

[5]Zheping Yan, and Tao Chen are with the Qingdao Innovation and Development Center of Harbin Engineering University, China.

⋆Corresponding authors: Bing Wang, mail: bingwang@polyu.edu.hk.

†The first two authors contributed equally to this work.

into the scene reconstruction network to aggregate conditional priors and guide the generation process. To address sonar noise and reconstruction distortions affecting the AUV perception system, a continuous multilayer probabilistic terrain map is generated from the dense point cloud and utilized for path planning. Finally, the scene reconstruction and AUV's navigation experiments, conducted on two customized simulation datasets and one real-world dataset, verify that our method can not only reconstruct the comprehensive point cloud scene but also improve the accuracy and safety of AUV's navigation. The whole framework of this paper can be summarized in Fig. 1. The contributions of this work are summarized as follows:

- A novel navigation framework for AUVs is proposed, utilizing generative scene reconstruction and probabilistic terrain modeling to derive complete terrain information from sparse sonar data for path planning.
- A generative scene reconstruction network is proposed based on the local diffusion model to generate the dense point cloud, where the spatial-temporal fusion between consecutive sonar frames is incorporated into the generation process as extended conditions.
- The experiments are conducted on simulation and real-world datasets. The results demonstrate that our method can output accurate, stable, and complete dense point clouds and then guide AUV's path planning.

II. Related Work

A. Scene Reconstruction from Point Cloud

Various methods have been employed to reconstruct dense point clouds from sparse input data. Traditional approaches include nearest neighbor upsampling algorithms [9], such as Midpoint Interpolation (Mid-I). Mid-I calculates the distances between points and their nearest neighbors using Euclidean distance or KD-tree algorithms. Subsequently, it selects the midpoint between two nearest neighbors as an interpolation point to fill gaps between adjacent points. To address the sparsity of sensor data (e.g., LiDAR, sonar), the Gaussian Process (GP) models discrete measurements as continuous probability distributions [10]. This approach predicts attributes of unobserved regions and quantifies uncertainties, thereby generating dense and continuous environmental maps. Compared to interpolation methods, GP eliminates the need for model parameterization, making it suitable for reconstructing complex terrains. However, GP suffers from limitations in computational efficiency and real-time performance due to its high algorithmic complexity. While these methods can reconstruct denser and more complete point clouds from sparse data, they exhibit a high dependence on prior information.

Benefiting from the data-driven and trainable capabilities of deep learning, researchers have increasingly focused on learning-based scene reconstruction methods. Grad-PU [11] employed midpoint interpolation and a refinement network to address the challenge of fixed point cloud upsampling ratios, formulating the refinement network as an optimization problem minimizing a point-to-point distance function. PoinTr [12] utilized a transformer encoder-decoder architecture to learn the structural information of local interactions and global correlations for point cloud generation. However, this auto-regressive method typically outputs a single, deterministic output, limiting its capacity to model the uncertainty inherent in complex geometries. DPCG-Net [13] proposed using conditional GANs for object-level dense point cloud completion. Nevertheless, GANs often face the challenge of mode collapse when generating high-dimensional complex data, resulting in a lack of sample diversity. Inspired by the achievements of diffusion models in point cloud generation [14], [15], Qu *et al.* [16] proposed a point cloud upsampling method (PUDM) by using denoising diffusion probabilistic models (DDPM) [17] to generate the dense point cloud from the sparse point cloud. Compared to convolutional neural networks (CNN) and transformer networks, diffusion models generate point clouds through a progressive denoising process, effectively reconstructing complex terrain geometries. Additionally, diffusion models directly model the data distribution, offering mathematical stability and interoperability. In contrast to auto-regressive and GAN-based methods, diffusion models leverage an iterative denoising process to capture complex point cloud geometries effectively. This approach facilitates the generation of more diverse samples and richer conditional distributions, thereby significantly mitigating mode collapse and enhancing uncertainty modeling capabilities. However, PUDM [16] targets object-level reconstruction, which requires normalizing the point cloud to generate data following a standard Gaussian distribution. This normalization process may lead to deformation or structural loss in the scene-scale point cloud. To solve this problem, Nunes *et al.* proposed diffusion models for LiDAR scene completion (LiDiff) [8], where the diffusion process can be formulated as a random noise offset added locally to each point in the scene. Therefore, the diffusion process operates directly on the original scale. However, given the slow movement of AUVs and the inherent low density and high noise of 3D MBES data, LiDiff [8] struggles to generate dense point clouds when conditioned solely on the current frame's data. Therefore, this paper proposes to employ a temporal attention mechanism to aggregate features from overlapping regions of consecutive point clouds, which are used to condition the diffusion process.

B. Underwater Probabilistic Terrain Modeling

Recent advancements in probabilistic modeling have yielded various approaches for handling uncertainties in terrain mapping for autonomous driving [18]. The integration of probabilistic frameworks allows for the development of terrain models that explicitly account for the inherent uncertainties in data. Such models are critical for enhancing the dynamic perception and autonomous decision-making capabilities of mobile robots operating in complex environments [19].

Similarly, AUVs also encounter sparse data and challenging underwater environments. A learning-based monocular image scene range estimator was proposed [20] to construct a probabilistic elevation map. However, due to its reliance on optical images, this method requires a series of image preprocessing operations and imposes high demands on the underwater environment. Based on Bayes' theorem, a probabilistic roadmap

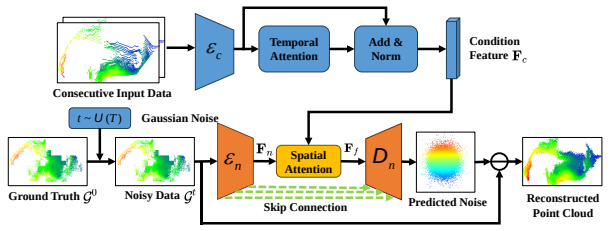

Fig. 2: Overview of the proposed reconstruction method.

was constructed using sidescan sonar data [21], with a naive assumption of mutual independence among three seabed features. A large-scale probabilistic mapping framework based on stochastic variational Gaussian processes was developed by Torroba *et al.* [22] using MBES data. Although sonar was employed in both studies, the raw data's low resolution still significantly limits the accuracy of probabilistic terrain maps.

This paper combines conditional DDPM-based scene reconstruction with probabilistic terrain modeling to generate continuous, multilayer probabilistic maps that accurately represent seafloor topography, thereby enhancing terrain fidelity and AUV perception across diverse scenarios.

III. Nomenclature

i	The index of timestamps.
$r = 10$	The upsampling ratio.
$\mathbf{X}_i = \{\mathbf{x}_{i_n}\}_{n=1}^{N} \in \mathbb{R}^{N\times 3}$	The input point cloud frame.
$\mathbf{Y}_i = \{\mathbf{y}_{i_m}\}_{m=1}^{rN} \in \mathbb{R}^{rN\times 3}$	The reconstructed point cloud frame.
$\boldsymbol{\mathcal{G}}_i = \{\boldsymbol{g}_{i_m}\}_{m=1}^{rN} \in \mathbb{R}^{rN\times 3}$	The ground truth point cloud frame.
$0 \le t \le T$	The diffuison step.
$\boldsymbol{\epsilon}_{\mathrm{g}} \in \mathbb{R}^{rN\times 3}$	The random noise in the global diffusion model.
$\boldsymbol{\epsilon}_{\mathrm{l}} \in \mathbb{R}^{3}$	The random noise in the local diffusion model.
$\boldsymbol{c}$	The genration condition.
$\boldsymbol{\epsilon}_\theta(\boldsymbol{g}_{i_m}^t, \boldsymbol{c}, t)$	The noise prediction network.
$\mathbf{F}_i, \mathbf{F}_n \in \mathbb{R}^{N\times d}$	The feature representation of $\mathbf{X}_i$, $\boldsymbol{\mathcal{G}}_i^t$.
$\mathbf{F}_c, \mathbf{F}_f \in \mathbb{R}^{N\times d}$	The temporal and spatial fusion feature.
d	The feature dimension of $\mathbf{F}_i$.
$Q_t, K_t, V_t \in \mathbb{R}^{N\times d}$	The query, key, and value matrices of temporal attention fusion.
$Q_s, K_s, V_s \in \mathbb{R}^{N\times d}$	The query, key, and value matrices of spatial attention fusion.

IV. Underwater Generative Scene Reconstruction with Diffusion Model

This section details the proposed reconstruction framework, which encompasses the formulation of the local DDPM for scene reconstruction, a spatial-temporal attention-guided diffusion model, and the associated training and inference procedures. The overall architecture is shown in Fig. 2.

A. Local DDPM for Scene Reconstruction

Given an input point cloud frame $\mathbf{X}_i$, we use the conditional DDPM to reconstruct a dense and complete point cloud frame $\mathbf{Y}_i$. Based on a sequence of consecutive point clouds and AUV's odometry, we can build the global map and generate the ground truth $\boldsymbol{\mathcal{G}}_i$ for each input $\mathbf{X}_i$. The objective of our reconstruction model is to make $\mathbf{Y}_i$ as close as possible to $\boldsymbol{\mathcal{G}}_i$.

The DDPM [17] consists of a forward diffusion process and a reverse denoising process. The forward diffusion process progressively adds Gaussian noise to $\boldsymbol{\mathcal{G}}_i$, transforming it into a randomly distributed noisy point cloud. In the reverse denoising process, a neural network learns to gradually remove this noise by predicting the added noise at each step of the forward diffusion. This iterative removal ultimately restores the noisy point cloud to a state closely approximating $\boldsymbol{\mathcal{G}}_i$.

1) The Local Diffusion Process: Given a pair of input point cloud $\mathbf{X}_i$ and the ground truth $\boldsymbol{\mathcal{G}}_i$, the diffusion process aims to add the Gaussian noise over T steps. This process is parameterized by a sequence of predefined values, $\beta_1, \ldots, \beta_T$. At each step t, Gaussian noise is sampled and added to $\boldsymbol{\mathcal{G}}_i$ based on the corresponding β_t. Following DDPM [17], the noisy point cloud at t step $\boldsymbol{\mathcal{G}}_i^t$ can be written as:

$$\boldsymbol{\mathcal{G}}_i{}^t = \sqrt{\bar{\alpha}_t}\boldsymbol{\mathcal{G}}_i + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\mathrm{g}},\ \boldsymbol{\epsilon}_{\mathrm{g}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha} = \prod_{k=1}^{t} \alpha_k$. This equation is typically employed in object-scale point cloud reconstruction methods [16], [23]. Within these methods, point clouds are either normalized to the space of $[-1, 1]$ or possess a small scale close to a standard Gaussian distribution. However, 3D MBES data typically exhibit a larger scale and are not normalized. Normalizing 3D MBES data to the $[-1, 1]$ space can lead to the loss or distortion of scene information. To solve this problem, a local denoising strategy is adopted, following [8]. Rather than directly applying the noise $\boldsymbol{\epsilon}_{\mathrm{g}}$ to $\boldsymbol{\mathcal{G}}_i$, the local diffusion process adds a noise offset to each point $\boldsymbol{g}_{i_m}^t$. Therefore, Eq. (1) can be rewritten as:

$$\begin{aligned} \boldsymbol{g}_{i_m}^t &= \boldsymbol{g}_{i_m} + (\sqrt{\bar{\alpha}_t}\mathbf{0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\mathrm{l}}) \\ &= \boldsymbol{g}_{i_m} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\mathrm{l}}, \end{aligned} \tag{2}$$

By iteratively applying the operation in Eq. 2, we can obtain a noisy point $\boldsymbol{g}_{i_m}^T$ that fully conforms to a random distribution. These points collectively form the noisy point cloud $\boldsymbol{\mathcal{G}}_i{}^t$.

2) The Local Denoising Process: Started from $\boldsymbol{g}_{i_m}^T$, the denoising process aims to reverse the diffusion process to obtain $\boldsymbol{g}_{i_m}$. Specifically, the noise prediction network $\boldsymbol{\epsilon}_\theta$ predicts the noise $\boldsymbol{\epsilon}_{\mathrm{l}}$ added at each step of the diffusion process. The noisy point $\boldsymbol{g}_{i_m}^t$, the step t, and the generation condition $\boldsymbol{c}$ are input to $\boldsymbol{\epsilon}_\theta$. Building upon the local diffusion process defined in Eq. (2), the local denoising process can be formulated as:

$$\boldsymbol{g}_{i_m}^{t-1} = \boldsymbol{g}_{i_m}^t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{g}_{i_m}^t, \boldsymbol{c}, t) + \sqrt{1-\bar{\alpha}_{t-1}}\mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

By Eq. (3), $\boldsymbol{g}_{i_m}$ is recovered through an iterative process that removes the predicted noise $\boldsymbol{\epsilon}_\theta$ from the noisy point $\boldsymbol{g}_{i_m}^T$.

B. Spatial-Temporal Attention Guided Diffusion

The point cloud frames collected during AUV locomotion are often highly sparse and noisy, making it challenging

to guide the point cloud generation process effectively. To address this, our proposed method leverages a spatial-temporal attention mechanism across consecutive point clouds to aggregate conditional features. Specifically, for consecutive point clouds $\mathbf{X}_{i-1}$ and $\mathbf{X}_i$, the encoder ε_c of MinkUNet [24] is first employed to extract their features, $\mathbf{F}_{i-1}$ and $\mathbf{F}_i$. Next, Multilayer Perceptrons (MLPs) transform $\mathbf{F}_i$ into the temporal query matrix Q_t, and $\mathbf{F}_{i-1}$ into the temporal key and value matrices (K_t, V_t). Subsequently, these conditional features are aggregated using the temporal attention mechanism:

$$\mathbf{F}_c = \text{MLP}(\text{softmax}(\frac{Q_t K_t^\top}{\sqrt{d}} V_t)) + \mathbf{F}_i \ , \tag{4}$$

where softmax denotes the activation function. The noise prediction network $\boldsymbol{\epsilon}_\theta$ is employed to predict the noise $\boldsymbol{\epsilon}_\text{l}$ at each diffusion step t, leveraging the aggregated condition feature $\mathbf{F}_c$ and the ground truth $\boldsymbol{\mathcal{G}}_i$. MinkUNet [24], which adopts a standard U-Net framework, serves as the backbone. Within each layer of the encoder ε_n, sinusoidal positional encoding, as adopted in PVD [14], is utilized to encode the temporal information $\tau \in \mathbb{R}^d$ associated with denoising step t. The output of each layer, $\mathbf{F}_l$, is then obtained via an element-wise multiplication $\mathbf{F}_l \odot \tau$, serving as the input for the subsequent layer. After four encoder layers of ε_n, the feature representation $\mathbf{F}_n$ of the noisy point cloud $\boldsymbol{\mathcal{G}}_i^t$ is obtained. Subsequently, $\mathbf{F}_n$ and the condition feature $\mathbf{F}_c$ are aggregated using a spatial attention mechanism. Specifically, $\mathbf{F}_n$ is transformed into the spatial query matrix Q_s, while $\mathbf{F}_c$ is transformed into the spatial key and value matrices (K_s, V_s), both via MLPs. The spatial fusion feature $\mathbf{F}_f$ is expressed as:

$$\mathbf{F}_f = \text{MLP}(\text{softmax}(\frac{Q_s K_s^\top}{\sqrt{d}} V_s)) + \mathbf{F}_n \ . \tag{5}$$

Subsequently, $\mathbf{F}_f$ is fed into the decoder of the MinkUNet to estimate the noise added into the ground truth $\boldsymbol{\mathcal{G}}_i$.

C. Training and Inference

1) Training: The training process aims to optimize the noise prediction network $\boldsymbol{\epsilon}_\theta$ to output the noise $\boldsymbol{\epsilon}_\text{l}$ added at t step with the given ground truth $\boldsymbol{\mathcal{G}}_i$ and the condition information $\mathbf{c}$. To avoid training an encoder separately, we adopt the strategy of classifier-free guidance [25] to train our diffusion model. The generation model is trained to learn the conditional and unconditional noise distributions. During each training step, the model uses a predetermined probability p to decide whether to learn the conditional noise distribution or the unconditional noise distribution. If the random number is less than the predetermined probability p, the model learns the unconditional noise distribution, where the conditional information $\mathbf{c} = \varnothing$; otherwise, the model learns the conditional noise distribution, where the conditional information $\mathbf{c} = \mathbf{X}_i$. Given the ground truth $\boldsymbol{\mathcal{G}}_i$ and the random step t, the noisy data $\boldsymbol{\mathcal{G}}_i^t$ is sampled from Eq. (2) by local noise $\boldsymbol{\epsilon}_\text{l}$. Then, the predicted noise can be written as follows:

$$\boldsymbol{\epsilon}_\theta(\boldsymbol{\mathcal{G}}_i^t, \mathbf{c}, t) = \boldsymbol{\epsilon}_\theta(\boldsymbol{\mathcal{G}}_i^t, \varnothing, t) + w[\boldsymbol{\epsilon}_\theta(\boldsymbol{\mathcal{G}}_i^t, \mathbf{X}_i, t) - \boldsymbol{\epsilon}_\theta(\boldsymbol{\mathcal{G}}_i^t, \varnothing, t)], \tag{6}$$

The model training objective can be written as $L_\theta(\boldsymbol{\mathcal{G}}_i^t, \mathbf{c}, t) = \left\| \boldsymbol{\epsilon}_\text{l} - \boldsymbol{\epsilon}_\theta(\boldsymbol{\mathcal{G}}_i^t, \mathbf{c}, t) \right\|$, where $\mathbf{c} = \{\mathbf{X}_i, \varnothing\}$ and $\boldsymbol{\epsilon}_\theta$ denote the predicted noise. We also add the noise prediction regularization to guide the predicted noise distribution to approximate the expected Gaussian distribution. Therefore, the final training loss can be written as $L = L_\theta(\boldsymbol{\mathcal{G}}_i^t, \mathbf{c}, t) + \gamma(\bar{\boldsymbol{\epsilon}}_\theta^2 + (\hat{\boldsymbol{\epsilon}}_\theta - 1)^2)$, where γ is a weighting factor, and $\bar{\boldsymbol{\epsilon}}_\theta$ and $\hat{\boldsymbol{\epsilon}}_\theta$ denote the mean and standard deviation of the predicted noise distribution.

2) Inference: During the inference process, the points of $\mathbf{X}_i$ are first replicated r times. Local noise is then sampled and added to each point, generating the noisy point cloud $\boldsymbol{\mathcal{G}}_i^T$. Subsequently, the predicted noise is iteratively removed from $\boldsymbol{\mathcal{G}}_i^T$ based on Eq. (3), ultimately generating $\mathbf{Y}_i$.

V. Underwater Probabilistic Terrain Modeling

In complex near-seabed environments, the probabilistic terrain map (PTM) provides a robust framework for autonomous AUV navigation by integrating spatial and temporal data to represent underwater terrain with uncertainty. Leveraging 3D MBES sonar data, the PTM constructs a probabilistic point cloud that accounts for sensor noise and environmental dynamics, enabling the identification of navigable regions and obstacles. This approach can enhance AUV navigation in cluttered and uneven terrains while supporting real-time path planning and decision-making.

The construction of the PTM for AUV navigation employs an incremental updating method based on 3D probabilistic occupancy grids [19], [26], [27]. By partitioning the environment into 3D voxel grids, each assigned a probability of occupancy, the PTM enables real-time updates as new 3D MBES sonar sensor data is acquired. Based on the probabilistic point cloud representation method of OctoMap [26] and the probabilistic terrain mapping framework developed by [27], the process of updating the terrain map involves two key components: local sonar sensor-based probabilistic map update and global AUV odometry-based map update. In this process, the occupancy grid map takes the AUV's odometry $\mathbf{p}_i \in \mathbb{R}^3$ and a reconstructed 3D MBES sonar point cloud $\mathbf{Y}_i$ as input. By applying incremental Bayesian updates, it fuses the incoming sonar measurements with existing map data to construct an octree-based probabilistic map. The incrementally reconstructed probabilistic terrain point cloud enables the generation of a continuous and accurate multilayer probabilistic terrain map, shown in Fig. 3. This methodology not only accounts for sonar sensor uncertainty and environmental dynamics but also ensures computational efficiency, critical for onboard AUV systems with limited processing power.

The resulting multilayer probabilistic terrain map, spanning 250 m × 250 m with a resolution of 0.5 m, is supplied to the path planner, which is tasked with obstacle avoidance. A multilayer probabilistic terrain map encompasses layers representing depth, flatness, and slope. The depth layer provides elevation constraints in the vertical plane for the AUV's path planner, while the slope layer penalizes large pitch angle maneuvers over steep terrains. The flatness layer guides the AUV towards non-flat areas, offering significant terrain feature constraints for terrain-matching navigation algorithms. These three terrain characteristics are combined as a multilayer probabilistic terrain map to support the path planning algorithm to achieve the optimal motion path in near-seabed environments.

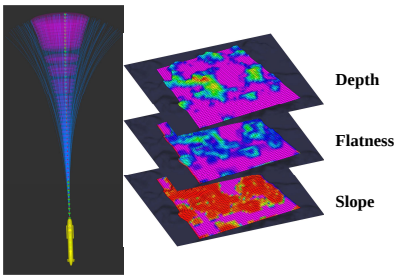

Fig. 3: Sampling-based MPPI planner for AUV and the multilayer probabilistic terrain map as the planner constraints, including depth, flatness, and slope.

VI. Application to AUV Navigation

This section focuses on exploring the AUV's path planning task based on these probabilistic terrain maps and formulating the underactuated AUV's dynamic model for navigation.

A. Integration with Path Planner

Our path planner is implemented by the Model Predictive Path Integral (MPPI) method, leveraging the dense probabilistic terrain map derived from our diffusion model as an obstacle constraint in the near-seabed environment. MPPI is a sampling-based optimal control method that solves nonlinear control systems with complex dynamics and uncertainties [28]. It generates numerous sampled trajectories by perturbing control inputs with noise, evaluates their cumulative cost using the AUV's dynamic model, and then selects the optimal trajectory with optimal control command based on a path integral formulation. The detailed underactuated AUV's dynamics model can be found in VI-B. As shown in Fig. 3, MPPI utilizes parallel sampling and forward dynamic trajectory prediction to compute the optimal planned path in complex underwater terrain environments, while the multilayer probabilistic terrain map serves as an obstacle constraint for MPPI. The MPPI planner, with a 10-second horizon and 2 Hz evaluation frequency, leverages the AUV's dynamics model to optimize trajectories that minimize time-to-goal while balancing soft risk constraints and path smoothness. The planner enforces strict compliance with physical feasibility constraints, including pitch angle limits, collision avoidance protocols, and velocity bounds, ensuring safe navigation and efficient motion.

B. Underactuated AUV Modeling

To study the motion planner of underactuated AUV, the mathematical modeling of the AUV can be developed using an earth-fixed frame W and a body-fixed frame B. Our underactuated AUV, propelled by a single thruster, utilizes cruciform control surfaces to govern its pitch motion in the vertical plane and yaw motion in the horizontal plane. According to [29], the dynamic model is simplified as a cylindrical rotating body with a single rear propeller and cruciform control surfaces. Six independent coordinates of the AUV are employed to determine the vehicle's position and orientation. The AUV's dynamic modeling can be expressed as follows:

$$\boldsymbol{M}\dot{\boldsymbol{\nu}} + \boldsymbol{C}(\boldsymbol{\upsilon})\boldsymbol{\upsilon} + (\boldsymbol{D}(\boldsymbol{\upsilon}) + \boldsymbol{D_n}(\boldsymbol{\upsilon}))\boldsymbol{\upsilon} = \boldsymbol{\tau} + \boldsymbol{\omega}, \quad (7)$$

$$\dot{\boldsymbol{\eta}} = \boldsymbol{J}(\boldsymbol{\eta})\boldsymbol{\nu}, \quad (8)$$

where $\boldsymbol{\eta} = [x\ y\ z\ \phi\ \theta\ \psi]^\top \in \mathbb{R}^6$, $\boldsymbol{\nu} = [u\ v\ w\ p\ q\ r]^\top \in \mathbb{R}^6$ are the vectors of position/Euler angles in the earth-fixed frame and velocities in the body-fixed frame, respectively. The matrix $\boldsymbol{J}(\boldsymbol{\eta}) \in \mathbb{R}^{6\times6}$ is the transformation matrix between the earth-fixed frame and the body-fixed frame. $\boldsymbol{M} \in \mathbb{R}^{6\times6}$ denote the inertial matrix. The vector $\boldsymbol{\tau} \in \mathbb{R}^6$ represents the vector of external forces and moments acting on the vehicle. $\boldsymbol{\omega} \in \mathbb{R}^6$ is the dynamic time-varying disturbance vector. The transform matrix $\boldsymbol{J}(\boldsymbol{\eta})$ can be expressed as $\boldsymbol{J}(\boldsymbol{\eta}) = diag\,[\boldsymbol{J_1}(\boldsymbol{\eta})\ \boldsymbol{J_2}(\boldsymbol{\eta})]$, where the matrices $\boldsymbol{J_1}(\boldsymbol{\eta})$ and $\boldsymbol{J_2}(\boldsymbol{\eta})$ are given by:

$$\boldsymbol{J_1}(\boldsymbol{\eta}) = \begin{bmatrix} \cos(\psi)\cos(\theta) & -\sin(\psi)\cos(\phi) + \sin(\phi)\sin(\theta)\cos(\psi) & \sin(\psi)\sin(\phi) + \sin(\theta)\cos(\psi)\cos(\phi) \\ \sin(\psi)\cos(\theta) & \cos(\psi)\cos(\phi) + \sin(\phi)\sin(\theta)\sin(\psi) & -\cos(\psi)\sin(\phi) + \sin(\theta)\sin(\psi)\cos(\phi) \\ -\sin(\theta) & \sin(\phi)\cos(\theta) & \cos(\phi)\cos(\theta) \end{bmatrix}, \quad (9)$$

$$\boldsymbol{J_2}(\boldsymbol{\eta}) = \begin{bmatrix} 1 & \sin(\phi)\tan(\theta) & \cos(\phi)\tan(\theta) \\ 0 & \cos(\phi) & -\sin(\phi) \\ 0 & \sin(\phi)/\cos(\theta) & \cos(\phi)/\cos(\theta) \end{bmatrix}. \quad (10)$$

Since the symmetry assumption allows the off-diagonal terms of the matrix $\boldsymbol{M}$ to be neglected, the inertia matrix $\boldsymbol{M}$ can be expressed as $\boldsymbol{M} = diag[m_{11}\ m_{22}\ m_{33}\ m_{44}\ m_{55}\ m_{66}]$. $\boldsymbol{C}(\boldsymbol{\upsilon})$ is the state-dependent matrix of Coriolis and centripetal terms. $\boldsymbol{D}(\boldsymbol{\upsilon})$ and $\boldsymbol{D_n}(\boldsymbol{\upsilon})$ represent the linear and nonlinear hydrodynamic damping matrices respectively. The detailed definition and derivation process can be found in [29].

VII. Experiments and Results

A. Dataset and Implementation Details

Dataset: The experiments were conducted on two simulation datasets and a real-world dataset. In two simulation datasets, the 3D MBES sonar operates at a central frequency of 1200 Hz with a configurable update rate of 1 Hz. The MBES was mounted on the front bottom section of the AUV to optimize downward-facing coverage. The sonar had a horizontal aperture of 120°, a vertical aperture of 40°, and an angular resolution of 0.5 degrees. For each frame of the sparse point cloud in the simulation dataset, the detection range was set to range from 10 m to 100 m. In the simulation, the AUV's odometry is obtained by propagating its six-degree-of-freedom (6-DOF) kinematic and dynamic model. In the URTerrain sonar dataset, the AUV traveled 100 m at a height of 50 m from the seabed, collecting 420 frames of sonar and odometry data. In the RMTerrain sonar dataset, the AUV's height from the seabed was reduced to 35 m, including a total of 220 frames of sonar and odometry data. We divided the data into training (70%), validation (20%), and test (10%)

sets based on the chronological order of AUV data collection. To ensure spatial non-overlap among the training, validation, and test sets, we meticulously planned the AUV's trajectory to ensure it followed a unidirectional and non-overlapping path. We also filtered out timestamp data that could potentially lead to overlapping scanned regions. Each frame of the point cloud contains point numbers between 5,000 and 8,000. For the real-world dataset, the MBES data were collected by the University of Gothenburg's Kongsberg Hugin AUV in the eastern nearshore waters of the Dotson Ice Shelf in West Antarctica [30]. The AUV was equipped with a Kongsberg EM2040 multibeam echosounder. During data collection, the AUV maintained a constant-altitude, constant-velocity cruising mode at approximately 2 m/s, 100 m above the seabed. This collection spanned approximately 19 hours, covering a total distance of 138 km. The sonar features 400 beams, operates at a frequency of 400 kHz, and possesses a horizontal scanning angle of 120°. The AUV's trajectory was recorded by the onboard inertial navigation system (INS).

Implementation Details: To train our Diffusion-based model, following the map generation method in [8], we concatenated each frame of the point cloud into a complete map based on its corresponding pose information. We integrated the AUV's pose information from a high-precision INS with sparse point clouds from a 3D MBES using the point cloud stitching technique to generate a global and complete point cloud map. Based on this information, in two simulation datasets, we extracted the corresponding region of the global map that represents the dense point cloud ground truth for the given timestamp and then randomly sampled 40,000 points as the ground truth for the reconstructed point cloud. For the real-world dataset, we configured each frame of the point cloud to contain 4,000 points, with the reconstructed point cloud comprising 20,000 points. For the simulation dataset, through synchronized sonar-pose data streams via timestamp alignment, we compute the global point cloud maps with a 2 cm mean Chamfer Distance error. For field datasets, the GPS/INS drift correction with time synchronization enabled centimeter-level absolute positioning in pose-fused dense maps. This method enabled the generation of pose-fused dense point cloud maps with absolute positional errors constrained.

During training, we selected the cosine noise schedule introduced in the improved DDPM [31], as it has been shown to provide smoother transitions during the denoising process. The diffusion step $T = 1000$ was chosen to balance computational efficiency and reconstruction quality. Following LiDiff [8], the noise regularization coefficient was set to 5.0 to prevent overfitting to noisy underwater point clouds. The classifier-free guidance probability and weights were set to 0.2 and 6.0, respectively, to achieve a balance between diversity and fidelity in the generated point clouds. The quantization resolution of 0.2 m was determined based on the density characteristics of underwater point clouds. This resolution ensures that geometric details essential for AUV navigation, such as obstacles and terrain features, are preserved while avoiding excessive computational overhead. The Adam optimizer [32] with the initial learning rate of 10^{-3} was chosen for its stable convergence during training. The learning rate decay schedule (halved every 5 epochs with a decay of 10^{-4}) was employed to refine the model weights during training, avoiding overfitting and ensuring generalization. The batch size was set to 1. During testing, we utilized DPMSolver [33] to reduce the denoising steps from 1000 to 50, improving inference speed while maintaining comparable reconstruction quality, making it meet the real-time requirement of AUV navigation.

TABLE I: Mean CD and FS evaluation on the test set of two simulation datasets. The best results are displayed in bold, while the second-best results are underlined.

Method	URTerrain		RMTerrain	
	CD (m)	FS-0.5	CD (m)	FS-0.5
Mid-I [34]	<u>0.8232</u>	–	**1.1033**	–
GP [10]	7.2018	0.6576	4.8034	0.5283
Grad-PU [11]	21.5288	0.3231	34.3028	0.2614
PoinTr [12]	15.2360	0.5229	30.8864	0.3025
PUDM [16]	1.0091	0.8629	1.3196	<u>0.8887</u>
LiDiff [8]	0.8451	<u>0.8709</u>	1.3751	0.8493
Ours	**0.7180**	**0.9002**	<u>1.2808</u>	**0.897**

Baselines: We compared our method with different scene completion and point cloud upsampling methods including traditional optimization-based approaches (Mid-I [34] and GP [10]), CNN-based methods (Grad-PU [11]), Transformer-based methods (PoinTr [12]), and diffusion model-based methods (PUDM [16] and LiDiff [8]). Mid-I [34] and GP [10] have been widely applied in robotic navigation scenarios, making them relevant baselines for underwater scene reconstruction. Grad-PU [11] and PoinTr [12] are included to evaluate the advantages of diffusion models in scene reconstruction. PUDM [16] and LiDiff [8] are selected to validate the effectiveness of the two key innovations in our approach: the local diffusion and denoising processes, and the spatial-temporal attention mechanism for underwater scene reconstruction. For all compared methods, we used official codes to train the model on three datasets. All reconstruction methods were implemented in Pytorch and trained on an Intel(R) Core(TM) I9-13900KF CPU and an NVIDIA GeForce RTX 4090 GPU.

B. Scene Reconstruction Results

1) Evaluation Metrics: The Chamfer Distance (CD) and F-score (FS) are chosen as the evaluation metrics to compare the scene reconstruction results. CD is a common metric to evaluate the average nearest distance between the reconstructed point cloud $\mathbf{Y}_i$ and the ground truth point cloud $\boldsymbol{\mathcal{G}}_i$, which can be defined as follows:

$$CD = \frac{1}{|\mathbf{Y}_i|}\sum_{\mathbf{y}\in\mathbf{Y}_i}\min_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_i}\|\mathbf{y}-\boldsymbol{g}\|_2^2 + \frac{1}{|\boldsymbol{\mathcal{G}}_i|}\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_i}\min_{\mathbf{y}\in\mathbf{Y}_i}\|\boldsymbol{g}-\mathbf{y}\|_2^2, \tag{11}$$

where $\mathbf{y}$ and $\boldsymbol{g}$ denote points in $\mathbf{Y}_i$ and $\boldsymbol{\mathcal{G}}_i$. However, CD may sometimes be misleading because of its sensitivity to outliers [35]. Therefore, we also utilize the F-score to evaluate the precision and recall of scene reconstruction simultaneously: $F = (1+\beta^2)\frac{PR}{\beta^2 P+R}$, where P and R denote the precision and recall value of the scene reconstruction. β represents a parameter that adjusts the weight between precision and recall. When precision is more important, the β value is adjusted to

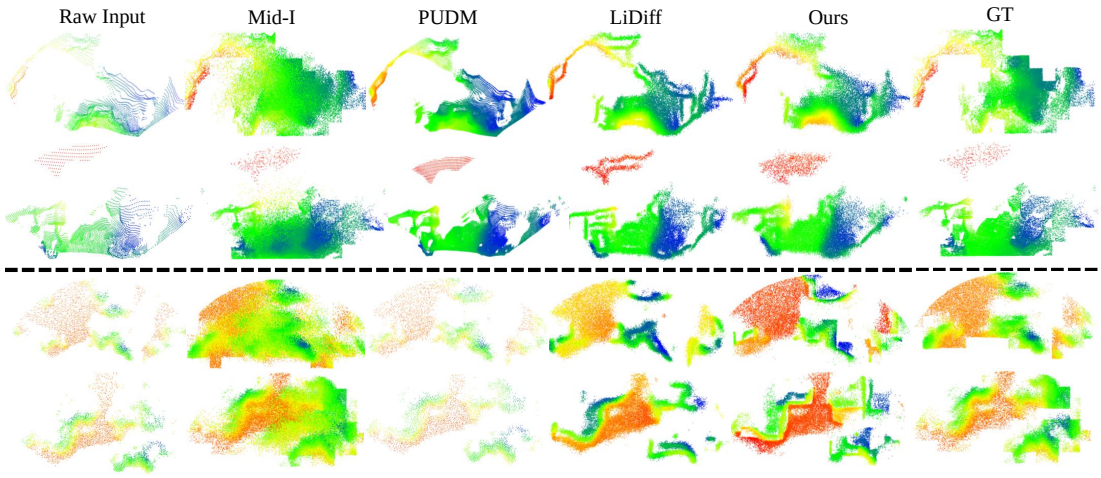

Fig. 4: Scene reconstruction results on four raw input point clouds from two simulation datasets. The right column shows ground truth (GT) point clouds, which are constructed by stitching each frame of sonar point clouds together based on the trajectory of the AUV. The top two rows show results on the URTerrain dataset, while the bottom two rows show results on the RMTerrain dataset. Colors represent the height of points, normalized based on the height range of each point cloud.

be less than 1. In this paper, we place greater emphasis on the precision of scene reconstruction, so we set $\beta = 0.5$.

2) Scene Reconstruction Evaluation: Table I shows the comparison results between our method and other methods. As shown in Table I, our method achieves the best performance in both metrics. Compared with the results of point cloud reconstruction methods with normalization (i.e., PUDM [16], Grad-PU [11], and PoinTr [12]), our method and LiDiff [8] achieve lower CD values and higher FS values. This shows that even with the point cloud normalization for scene scale, the current object-scale point cloud reconstruction method cannot effectively address the issue of scene reconstruction. This demonstrates the effectiveness of the local diffusion and denoising strategy adopted to reconstruct the large underwater scene in our method. On the other hand, the superior performance of our method over LiDiff [8] demonstrates the significance of aggregating the consecutive point cloud features under the condition by utilizing the spatial-temporal attention mechanism. GP [10], as a traditional method, is often used in navigation tasks of mobile robots to reconstruct dense scenes. However, its reconstruction error shown in Table I remains significant, and the reconstructed area is limited to the local range of the current frame point cloud. Compared to scene reconstruction methods based on generative models (i.e., our method and LiDiff [8]), GP [10] cannot provide richer and more accurate terrain information.

Fig. 4 compares the scene reconstruction results of different methods qualitatively. Although Mid-I [34] achieved a smaller error in CD error in Table I, when calculating the F-score, the results exceeded the valid range of $[0, 1]$. According to the reconstruction results shown in Fig. 4, the point cloud in the reconstruction results of Mid-I [34] appears in some non-continuous terrain areas. This means that the Mid-I method reconstructs areas that were originally obstacles into unobstructed seabed areas, which subsequently leads to the creation of erroneous probabilistic terrain maps for AUV navigation. Therefore, Mid-I [34] is not considered to be more advantageous. Compared to the results of LiDiff [8], the point clouds generated by our method have the advantage of clearer boundaries and more complete geometric structures.

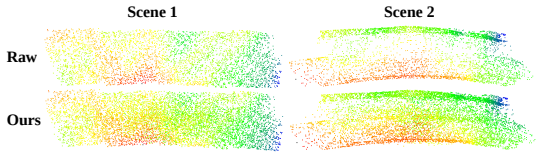

Fig. 5: Reconstruction results on two raw point clouds from the real-world dataset.

By comparing the quantitative and qualitative reconstruction results, it is obvious that our method achieves the best reconstruction results. Specifically, compared to traditional optimization-based methods (Mid-I [34] and GP [10]), our approach leverages deep neural networks for stronger data generation capabilities, resulting in lower reconstruction errors. Compared to CNN- and Transformer-based architectures (Grad-PU [11] and PoinTr [12]), our iterative denoising process better preserves complex geometric features of underwater terrains. Furthermore, compared to normalization-dependent methods like PUDM [16], our locally guided diffusion and denoising process avoids geometric distortions and loss caused by normalization, enabling more accurate terrain detail recovery. Finally, compared to LiDiff [8], our spatial-temporal attention mechanism provides better guidance for noise prediction, enhancing reconstruction accuracy.

3) Real-world Evaluation: To further validate the reconstruction performance of our method in practical underwater environments, we conducted scene reconstruction experiments using MBES data collected in the real-world environment. As shown in Fig. 5, reconstruction results demonstrate that our method enhances the resolution of seabed topography while preserving the fundamental geometric structure of the

TABLE II: Results of the spatial-temporal attention mechanism ablation experiment.

Method	CD-Mean (m)	CD-STD (m)	FS-0.5
LiDiff [8]	0.8451	0.1937	0.8709
Ours-S	0.7238	0.1398	0.8770
Ours-T	0.7291	0.1447	0.8747
Ours	**0.7180**	**0.1294**	**0.9002**

Fig. 6: Sensitivity curves of three model parameters.

terrain point cloud, thereby providing precise geometric point cloud data for seabed terrain modeling. Following the method we used to generate ground truth, we also utilized the AUV trajectory in the real dataset to stitch sparse point clouds and compute the ground truth as a reference. The quantitative results show that the average CD error between our reconstructed and the ground truth point cloud is 0.8094 m.

4) Ablation Study: To better demonstrate the significance of the proposed spatial-temporal attention mechanism incorporated in the conditional DDPM-based scene reconstruction method [8], we also evaluated the respective influence of these two modules on the scene reconstruction task. Table II shows quantitative results on the URTerrain dataset for different methods. The first row shows the results of the original LiDiff [8]. In the Ours-S experiment, we replaced the original closest-point linear fusion condition module in LiDiff [8] with our proposed spatial attention mechanism to guide the generation network. The improved performance of Ours-S over LiDiff [8] and our method over Ours-T validates the effectiveness of the spatial attention mechanism. Compared with the closest point linear fusion module in LiDiff [8], the spatial attention mechanism can fuse abundant and detailed geometric information, which provides a more efficient condition to guide the generation network. In the Ours-T experiment, the proposed temporal attention mechanism was added to LiDiff [8] to aggregate the consecutive point cloud features. The marked improvement of Ours-T relative to LiDiff [8] and our method over LiDiff-S confirms the importance of incorporating the consecutive point cloud features by the temporal attention mechanism. Fusing consecutive frame point clouds can aggregate more significant geometric features and boundary information in overlapping regions, thereby generating more effective conditional features. Combining the spatial and temporal attention mechanisms, our proposed method achieves the best results in this experiment. This highlights the effectiveness of employing both the temporal attention mechanism in fusing consecutive frame point clouds to generate augmented conditional features and the spatial attention mechanism in fusing conditional features with noise features to better guide the generation task.

To validate the rationality of our parameter selection, we conducted ablation studies on three key hyperparameters: attention depth, voxel grid size, and classifier-free guidance weight. The results are presented in Fig. 6, where the blue, green, and red curves correspond to the sensitivity analyses of these three parameters, respectively. As described in Section VII-A, our baseline model sets the attention depth to 1, the Voxel grid size to 0.2, and the classifier-free guidance weight to 6.0. It is observed from the figure that increasing the attention depth or decreasing the Voxel grid size does not yield significant performance improvements but rather introduces additional computational overhead. Although a larger Voxel grid size can enhance computational efficiency, it comes at the cost of reconstruction accuracy. Therefore, setting the attention depth to 1 and the voxel grid size to 0.2 achieves an effective trade-off between computational cost and model performance. Furthermore, the reconstruction accuracy peaks when the classifier-free guidance weight is set to 6.0. In summary, this series of ablation studies thoroughly validates the rationality and effectiveness of our chosen model parameters.

C. Probabilistic Terrain Mapping Performance

Due to the inherent noise and uncertainties of MBES data and those introduced by reconstruction models, achieving accurate probabilistic terrain predictions is crucial for the AUV's safe motion planner. Fig. 7 illustrates the point cloud maps reconstructed from different methods alongside the raw terrain point cloud. Building upon the raw accumulated point cloud map, our reconstruction method predicts a denser and more uniformly distributed point cloud map compared to other methods. It achieves a consistent estimation of the planar structure of the seafloor, as shown by the orange point cloud regions in our point cloud map, which is not achievable by other methods. Our method also effectively maintains consistency with the original terrain map, without introducing significant distortions to the geometric structure of the terrain.

Fig. 8 shows the estimated error map by comparing the probabilistic terrain depth maps generated by the raw and reconstructed point cloud scenes from three methods. This error represents the difference in seabed depth maps between two complete terrain structures due to reconstruction inaccuracies and is visualized using a color bar, where colors closer to deep blue indicate smaller errors. It is evident that our method's reconstruction results exhibit the largest deep blue regions in the error map, demonstrating that our method produces reconstructions most closely aligned with the GT when converted into probabilistic terrain depth maps. However, we observed that in the edge regions of the global map, our method's reconstructions exhibit relatively larger errors. As shown in Fig. 8, the red dashed box illustrates the comparison between the GT and reconstructed point clouds, highlighting the corresponding CD error. This could be attributed to the reduced overlap between consecutive point clouds caused by the AUV's roll angle during turning manoeuvres, which affects the consistency of the reconstruction. We believe that optimizing the AUV's motion trajectory and angular control could further improve the reconstruction results.

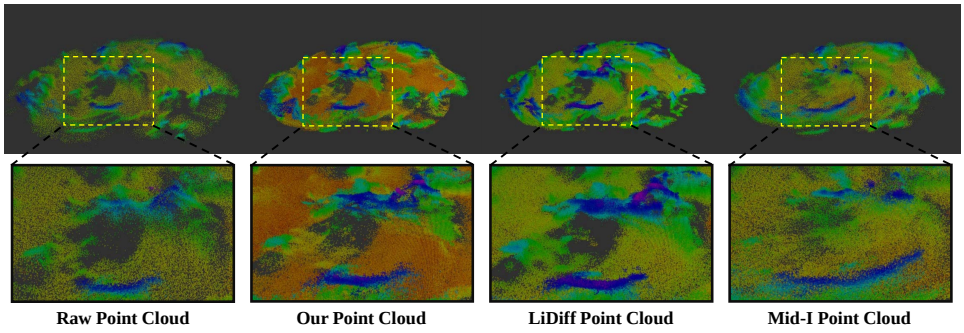

Fig. 7: Global probabilistic terrain point cloud maps generated by the GT and reconstructed point clouds. Colors represent the height of points, normalized based on the height range of each point cloud. The global probabilistic terrain map is generated by stitching single-frame point clouds along the AUV trajectory.

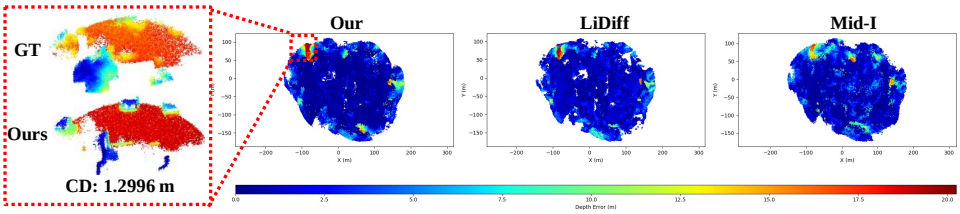

Fig. 8: Global probabilistic terrain depth maps generated by the GT and reconstructed probabilistic point clouds. Colors represent the depth error between the GT and reconstructed point clouds.

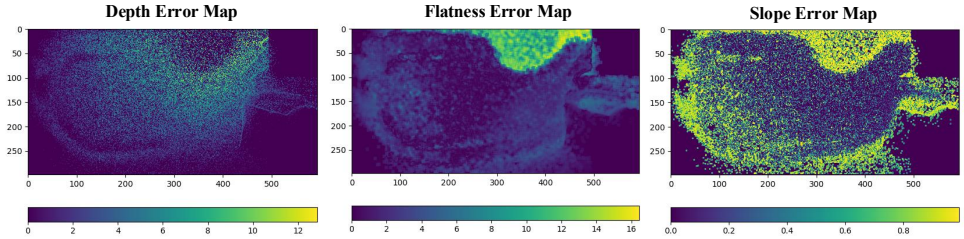

Fig. 9: Error analysis of multilayer probabilistic terrain map to quantify the reconstruction performance. Colors represent the errors between the predicted depth, flatness, and slope within each grid and GT.

To better analyze the error of a single-frame probabilistic terrain, as shown in Fig. 9, we compared the reconstruction results with the raw point cloud in terms of errors in multilayer probabilistic terrain maps, including terrain depth, flatness, and slope. For Fig. 9, we conducted a statistical analysis of errors presented in three subplots. The mean depth error in the depth error map is 1.74m, with 74.23% of the grid cells exhibiting errors smaller than 1m. In the flatness error map, the mean error is 2.31, with 90% of the grid cells having errors below 5. For the slope error map, the mean error is 0.31, with 66% of the grid cells showing errors smaller than 0.25. The results demonstrate that the terrain maps reconstructed by our method, utilizing probabilistic grid modeling, achieve an accuracy sufficient for AUV path planning.

D. Path Planning Performance

Based on the global probabilistic terrain maps generated offline by different methods, the MPPI path planning experiments were conducted under the same initial and terminal conditions, with a single terminal constraint. Different global probabilistic terrain maps are used as terrain obstacle cost constraints in the MPPI predictive sampling process, resulting in different planned trajectories as shown in Fig. 10. The quantitative results for path planning performance are shown in Table III. Smoothness is quantified by the sum of the squared accelerations in the x, y, and z directions along the path, while Depth Variation refers to the sum of absolute variation in the depth direction. Min Distance and Mean Distance represent the minimum and average distances, respectively, from the planned path to the different probabilistic terrain maps. Compared to other baselines, our spatial-temporal diffusion model approach shows the best results across all metrics, with the shortest and smoothest motion path. Furthermore, we also observe improvements in the vertical absolute depth variation and the safety distance from the seafloor, reflecting smoother pitch maneuvers and an enhanced safety margin to the seafloor that can improve sonar detection stability and overall navigation safety for the AUV. As shown in Fig. 10 and Table III, by utilizing the dense probabilistic terrain point clouds predicted by our spatial-temporal diffusion model, the

TABLE III: Path Planner Performance.

Point Cloud	Path Length (m)	Smoothness	Depth Variation (m)	Min Distance (m)	Mean Distance (m)
Mid-I [34]	164.79	19.45	22.92	2.44	10.43
LiDiff [8]	158.67	18.38	22.18	3.13	10.60
Our	**153.65**	**13.97**	**21.92**	**3.22**	**11.58**

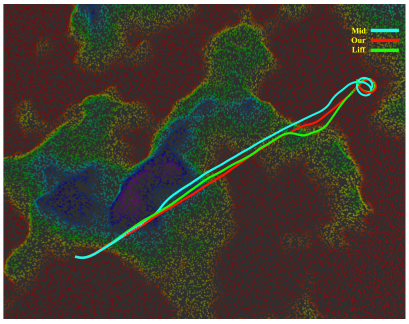

Fig. 10: Comparison of MPPI path planning results.

TABLE IV: The computation cost of our reconstruction method and AUV's path planning.

Module	Runtime (ms)	Memory (MB)
Encoder	5.31	12
Diffusion Sampler	847	840
Terrain Modelling	34	124
Path Planning	42	431

MPPI planner is capable of generating smoother and more efficient motion paths for the underactuated AUV. Compared to the paths planned with other probabilistic point cloud terrains, under the uniformly dense probability point cloud terrain generated by our spatial-temporal diffusion model, a smoother and lower energy consumption planned path can be obtained. In general, we observe that since our spatial-temporal diffusion model is capable of generating dense and uniform terrain structure from raw sonar point cloud, the AUV path planner using the probabilistic terrain generated by our spatial-temporal diffusion model enables optimal near-seabed exploration, resulting in a smooth and efficient motion trajectory. It is noted that although the reconstruction CD errors of Mid-I [34] are relatively low, its reconstruction results exhibit an inherent limitation: Mid-I [34] tends to generate point clouds in regions where no actual structures exist. This inconsistency can have significant implications for downstream tasks like probabilistic terrain modeling and path planning. For instance, Mid-I's reconstruction may introduce obstacles in originally obstacle-free regions, leading to erroneous outputs in the AUV's path planning module and ultimately causing navigation errors. Combining the reconstruction and path planning results, it is obvious that our method can not only achieve competitive quantitative results but also ensure more consistent and physically meaningful reconstructions, critical for real-world underwater navigation tasks.

To better analyze the application of our method in AUV navigation systems, we summarized the computation costs of each module in our navigation system during inference. According to Table IV, the total runtime of our reconstruction method is approximately 850 ms. Given that the sonar data update frequency in the dataset is approximately 1 Hz, the system still has a remaining time margin of roughly 150-250 ms, which is sufficient to cope with delays potentially caused by sensor synchronization or environmental interference. Furthermore, the GPU memory of the reconstruction network is less than 1GB. Considering both runtime and memory, our reconstruction method can be deployed on practical AUV platforms with limited GPU memory and satisfies the real-time requirements of AUV navigation.

E. Discussion

This paper proposes a novel and effective underwater scene reconstruction method to generate a dense and complete point cloud from a sparse and partial input. The local diffusion and denoising processes are formulated for the scene generation task. The spatial-temporal attention mechanism is introduced into the conditional Diffusion model to aggregate the consecutive point cloud features and fuse the condition feature and the generation feature. The reconstruction experiment results on three datasets demonstrate that our method outperforms other methods in terms of lower reconstruction errors and higher F-scores. To better illustrate the superiority of our method over other Diffusion-based scene reconstruction methods (i.e., PUDM [16] and LiDiff [8]), the comparisons of the generation networks and strategies are given as follows: (a) Compared with PUDM [16], we assume each point as the origin of sampled Gaussian noise rather than sampling the original point cloud as a mixed distribution. This generation strategy is more applicable for large-scale scene reconstruction to avoid point cloud normalization. (b) Compared with LiDiff [8], we introduce the temporal attention mechanism to fuse the consecutive point cloud features to better aggregate the conditional feature. We also adopt the spatial attention mechanism to fuse the condition feature and the generation feature to provide more detailed information to guide the generation process.

To explore the application of our method for AUVs' navigation, we also built a probabilistic terrain map and used path planning to validate the accuracy of the reconstructed scene point cloud. The path planning results demonstrate that our proposed method can provide the complete point cloud for the AUV to generate the shortest and smoothest motion path. To summarize, our scene reconstruction method brings great benefits to underwater terrain mapping and AUV navigation.

Limitations and future works: We look forward to motivating more scene reconstruction methods for AUVs' navigation. There are still several limitations in our proposed reconstruction methods: (a) Although our reconstruction method meets real-time requirements and improves reconstruction accuracy on existing hardware platforms, the introduction of a lightweight coarse-to-fine refinement network can further

enhance reconstruction accuracy. Therefore, future work will focus on end-to-end deployment technologies, such as knowledge distillation, model quantization, and hardware-software collaborative acceleration. This represents a research direction in the field of scene reconstruction that is both challenging and promising. (b) Currently, due to the lack of underwater point cloud datasets, it is challenging to perform extensive generalization training and testing. To evaluate the generalization ability of our model, we conducted cross-dataset tests using the URTerrain and RMTerrain datasets. Specifically, a model trained on the URTerrain dataset was tested on the RMTerrain dataset, yielding a CD of 1.9081 m. In comparison, a model trained and tested on the RMTerrain dataset produced a reconstruction error of 1.2808 m. These results indicate that while the performance of our model declines when applied to a different dataset, the reconstruction results still outperform traditional optimization-based methods such as GP. To address this limitation, future work could consider improving the generalization of our method by incorporating more environmental conditions to guide the reconstruction model. (c) Our reconstruction method relies on accurate AUV pose for aligning consecutive point clouds into a common coordinate frame, thus requiring a high-precision AUV positioning system. To mitigate this dependency, future work will focus on robust multi-sensor simultaneous localization and mapping (SLAM) and integrating localization uncertainty directly into the feature fusion network.

VIII. Conclusion

This paper introduces a novel DDPM for generating dense point clouds from sparse 3D sonar data. Our approach formulates the diffusion and denoising processes locally, treating each sparse input point as a distinct origin for Gaussian noise. Consequently, the noise prediction network learns to estimate per-point offsets rather than a single noise distribution for the entire scene. We introduce a spatial-temporal attention mechanism that aggregates features across consecutive frames. Then these aggregated features are fused with the noise prediction output to guide the denoising process more effectively. We evaluate our method on the downstream tasks of probabilistic terrain mapping and AUV path planning. Our method is benchmarked against both traditional and data-driven approaches on three datasets. Experimental results demonstrate that our method produces more accurate, stable, and complete point clouds than baseline approaches. These high-fidelity reconstructions provide the precise perception required for AUV navigation and path planning modules. Future work will focus on improving the real-time performance of our method, enabling its deployment in practical AUV navigation systems.

References

[1] J. Zhang, X. Xiang, W. Li, and Q. Zhang, "Adaptive neural control of flight-style auv for subsea cable tracking under electromagnetic localization guidance," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 5, pp. 2976–2987, 2023.

[2] E. Simetti, R. Campos, D. D. Vito, J. Quintana, G. Antonelli, R. Garcia, and A. Turetta, "Sea mining exploration with an uvms: Experimental validation of the control and perception framework," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 3, pp. 1635–1645, 2021.

[3] T. Lin, A. Hinduja, M. Qadri, and M. Kaess, "Conditional gans for sonar image filtering with applications to underwater occupancy mapping," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1048–1054.

[4] J. Tan, I. Torroba, Y. Xie, and J. Folkesson, "Data-driven loop closure detection in bathymetric point clouds for underwater slam," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3131–3137.

[5] S. Shen, Y. Zeng, C. Lai, S. Jiang, S. Wu, and S. Ma, "Rapid three-dimensional reconstruction of underwater defective pile based on two-dimensional images obtained using mechanically scanned imaging sonar," *Structural Control and Health Monitoring*, vol. 2023, no. 1, p. 3647434, 2023.

[6] C. Fu, C. Dong, C. Mertz, and J. M. Dolan, "Depth completion via inductive fusion of planar lidar and monocular camera," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 843–10 848.

[7] P. Li, R. Zhao, Y. Shi, H. Zhao, J. Yuan, G. Zhou, and Y.-Q. Zhang, "Lode: Locally conditioned eikonal implicit scene completion from sparse lidar," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8269–8276.

[8] L. Nunes, R. Marcuzzi, B. Mersch, J. Behley, and C. Stachniss, "Scaling Diffusion Models to Real-World 3D LiDAR Scene Completion," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[9] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation (ICRA)*. Shanghai, China: IEEE, May 9-13 2011.

[10] S. Vasudevan, F. Ramos, E. Nettleton, and H. Durrant-Whyte, "Gaussian process modeling of large-scale terrain," *Journal of Field Robotics*, vol. 26, no. 10, pp. 812–840, 2009.

[11] Y. He, D. Tang, Y. Zhang, X. Xue, and Y. Fu, "Grad-pu: Arbitrary-scale point cloud upsampling via gradient descent with learned distance functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "Pointr: Diverse point cloud completion with geometry-aware transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 478–12 487.

[13] M. Cheng, G. Li, Y. Chen, J. Chen, C. Wang, and J. Li, "Dense point cloud completion based on generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.

[14] L. Zhou, Y. Du, and J. Wu, "3d shape generation and completion through point-voxel diffusion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5826–5835.

[15] Z. Lyu, Z. Kong, X. Xu, L. Pan, and D. Lin, "A conditional point diffusion-refinement paradigm for 3d point cloud completion," *arXiv preprint arXiv:2112.03530*, 2021.

[16] W. Qu, Y. Shao, L. Meng, X. Huang, and L. Xiao, "A conditional denoising diffusion probabilistic model for point cloud upsampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 20 786–20 795.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] B. Forkel, J. Kallwies, and H.-J. Wuensche, "Probabilistic terrain estimation for autonomous off-road driving," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 864–13 870.

[19] Y. Ren, Y. Cai, F. Zhu, S. Liang, and F. Zhang, "Rog-map: An efficient robocentric occupancy grid map for large-scene and high-resolution lidar-based motion planning," 2023.

[20] B. Arain, F. Dayoub, P. Rigby, and M. Dunbabin, "Close-proximity underwater terrain mapping using learning-based coarse range estimation," *arXiv preprint arXiv:2001.00330*, 2020.

[21] S.-W. Huang, E. Chen, and J. Guo, "Efficient seafloor classification and submarine cable route design using an autonomous underwater vehicle," *IEEE Journal of Oceanic Engineering*, vol. 43, no. 1, pp. 7–18, 2018.

[22] I. Torroba, C. I. Sprague, and J. Folkesson, "Fully-probabilistic terrain modelling and localization with stochastic variational gaussian process maps," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8729–8736, 2022.

[23] X. Zheng, X. Huang, G. Mei, Y. Hou, Z. Lyu, B. Dai, W. Ouyang, and Y. Gong, "Point cloud pre-training with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 22 935–22 945.

[24] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.
[25] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
[26] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.
[27] P. Fankhauser, M. Bloesch, and M. Hutter, "Probabilistic terrain mapping for mobile robots with uncertain localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3019–3026, 2018.
[28] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Information-theoretic model predictive control: Theory and applications to autonomous driving," *IEEE Transactions on Robotics*, vol. 34, no. 6, pp. 1603–1622, 2018.
[29] T. I. Fossen, "Marine control systems : guidance, navigation and control of ships, rigs and underwater vehicles," *Rigs and Underwater Vehicles, Marine Cybernetics AS*, 2002.
[30] L. Ling, J. Zhang, N. Bore, J. Folkesson, and A. Wåhlin, "Benchmarking classical and learning-based multibeam point cloud registration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6118–6125.
[31] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.
[32] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[33] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.
[34] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.
[35] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3d reconstruction networks learn?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3405–3414.

Zhengyan Zhang (Graduate Student Member, IEEE) received the bachelor's degree in Measurement and Control Technology and Instrument from Nanjing Tech University, Nanjing, China, in 2020, and the M.Eng. degree in Control Engineering from the Harbin Institute of Technology, Shenzhen, China, in 2022, under the supervision of Prof. Max Q.-H. Meng. He is currently pursuing the Ph.D. degree in Flight Mechanics and Control with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong.

Liang Fang (Graduate Student Member, IEEE) received the B.E. degree in Detection, Guidance, and Control from the School of Intelligent Science and Engineering, Harbin Engineering University, China, in 2021. He is currently pursuing the Ph.D. degree in Control Science and Engineering with the School of Intelligent Science and Engineering, Harbin Engineering University, China. His research interests include autonomous navigation and motion control, active SLAM, and intelligent perception of unmanned systems.

Zheping Yan (Member, IEEE) received the B.E. degree in nuclear power plants, the M.E. degree in special auxiliary devices and systems for marine engineering, and the Ph.D. degree in control theory and control engineering from Harbin Engineering University, in 1994, 1997, and 2001, respectively. He is currently a Professor at Harbin Engineering University, China. His research interests include marine equipment automation and intelligent technology, modeling and autonomous motion control of AUVs.

Tao Chen (Member, IEEE) received the Ph.D. degree in control science and engineering from Harbin Engineering University, Harbin, China, in 2011. He is currently working as a Professor at Harbin Engineering University, China. His research interests include autonomous control, motion control, and cooperative control of AUVs.

Bing Wang (Member, IEEE) received the D.Phil./Ph.D. degree from the Department of Computer Science, University of Oxford, Oxford, U.K., in 2022. He is an Assistant Professor with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong. His research interests broadly lie in the design of intelligent perception solutions for autonomous systems and the development of reliable 3-D scene understanding algorithms on mobile robotics operating in the real world.

Chih-Yung Wen (Member, IEEE) received the B.Sc. degree in mechanical engineering from the Department of Mechanical Engineering, National Taiwan University, Taipei, Taiwan, in 1986, and the M.Sc. and Ph.D. degrees in fluid mechanics from the Department of Aeronautics, California Institute of Technology (Caltech), Pasadena, CA, USA, in 1989 and 1994, respectively. He is currently the Chair Professor of Aeronautical Engineering in the Department of Aeronautical and Aviation Engineering, the Director of COMAC-PolyU Research Institute for Large Aircraft, the Associate Director of Research Institute for Sports Science and Technology, and the Director of Research Centre of Unmanned Autonomous Systems. Professor Wen has authored and co-authored more than 300 scientific papers, conference papers, and book chapters. He was also awarded 14 patents. Professor Wen, currently a Fellow of ASME, RAeS, HKIE, and an AIAA Associate Fellow, actively engages in professional academic activities related to mechanical and aerospace engineering at both domestic and international levels. In addition, he serves as a member of various key professional boards and bodies related to the Aerospace Engineering.