Application of Photogrammetric Computer Vision and Deep Learning in High-Resolution Underwater Mapping: A Case Study of Shallow-Water Coral Reefs

Jiageng Zhong[1], Ming Li[1,2,*], Armin Gruen[2], Jianya Gong[1], Deren Li[1], Mingjie Li[3], Jiangying Qin[1]

[1] State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing,
Wuhan University, Wuhan 430079, China

[2] Institute of Geodesy and Photogrammetry, ETH Zurich, 8093 Zurich, Switzerland - mingli39@ethz.ch

[3] South China Sea Development Research Institute, Ministry of Natural Resources, Guangzhou 510310, China

Commission II/WG7

Key Words: Underwater photogrammetry, Computer vision, Deep learning, 3D reconstruction, Coral reefs.

Abstract

Underwater mapping is vital for engineering applications and scientific research in ocean environments, with coral reefs being a primary focus. Unlike more uniform and predictable terrestrial environments, coral reefs present a unique challenge for 3D reconstruction due to their intricate and irregular structures. Traditional 3D reconstruction methods struggle to accurately capture the nuances of coral reefs. This is primarily because coral reefs exhibit a high degree of spatial heterogeneity, featuring diverse shapes, sizes, and textures. Additionally, the dynamic nature of underwater conditions, such as varying light, water clarity, and movement, further complicates the accurate geometrical estimation of these ecosystems. With the rapid advancement of photogrammetric computer vision and deep learning technologies, there are emerging methods that have potential to enhance the quality of its 3D reconstruction. In this context, this study formulates a coral reef reconstruction workflow that incorporates these cutting-edge technologies. This workflow is divided into two core stages: sparse reconstruction and dense reconstruction. We conduct individual summaries of the relevant research efforts in these stages and outline the available methods. To assess the specific capabilities of these methods, we apply them to real-world coral reef images and conduct a comprehensive evaluation. Additionally, we analyze the strengths and weaknesses of different methods and identify areas for improvement. We believe this study offers valuable references for future research in underwater mapping.

1. Introduction

More than 70% of the Earth's surface is covered by water, predominantly oceans, presenting considerable scope for the advancing technologies dedicated to water observation. Affected by climate change and human activities, marine ecosystems, especially coral reef ecosystems, are facing significant challenges (Hughes et al., 2017). Coral reefs represent the most remarkable ecosystems in warm tropical and subtropical oceans. Although they cover less than 0.1% of the ocean floor, their fish communities encompass approximately one-third of the recognized marine species (Bowen et al., 2013). To enhance the understanding, monitoring, and protection of coral reefs, it is essential to use advanced technology to map, monitor and model coral reef habitats.

In coral reef observation, various approaches and platforms are employed, including satellite sensing, aerial remote sensing, vessel-based sonar and LiDAR, underwater vehicle-based imaging, and manual local in-situ underwater surveys (Collin et al., 2018, Price et al., 2019, Rossi et al., 2020, Character et al., 2021). Satellite and aerial remote sensing techniques offer a swift method for acquiring information in large-scale coral-monitoring applications (Casella et al., 2017). However, they fall short in capturing detailed and accurate observations of the intricate structures within coral reefs. In contrast, manual measurements demand substantial time investment, imposing constraints on the spatial and temporal scales. In terms of sensors, while sonar and LiDAR improve the acquisition of geometric information of benthic habitats, they face challenges in acquiring color information. The rise of vision-based underwater imaging enables the collection of higher-resolution data, unaffected by surface refraction of water, facilitating precise 3D reconstruction of real seabed coral reefs at a low cost (Rossi et al., 2020, Zhong et al., 2023). It provides a foundation of high-precision, high-resolution information crucial for subsequent research, and is becoming the centerpiece among various sensors (Zhong et al., 2023).

Over the past decade, thanks to the rapid advancements in computer vision technologies such as Structure-from-Motion (SfM) and Multi-View Stereo (MVS), underwater mapping based on photogrammetric computer vision has been extensively studied for coral reef observation. Utilizing high-resolution images captured by vehicles or divers, these techniques enable 3D observations with precision at the centimeter or even millimeter level (Guo et al., 2016). These approaches provide automated image processing tools, which facilitate the generation of fine 3D models that accurately represent the intricate spatial structural information of coral reefs (Zhong et al., 2023). However, due to the unique characteristics of coral reef environments and the limitations of current algorithms, there is still a need for improvements in the precision, robustness, and efficiency of underwater mapping. This is particularly true in light of the rapid advancements in learning-based image processing.

The data for this paper was collected from the shallow-water coral reefs in the vicinity of Moorea Island in French Polynesia. The island is surrounded by approximately 10 enclosed

https://doi.org/10.5194/isprs-annals-X-2-2024-247-2024 | © Author(s) 2024. CC BY 4.0 License.

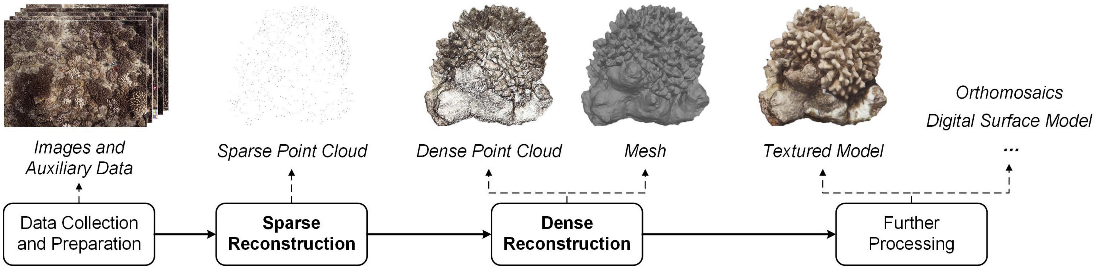


Figure 1. The overall 3D reconstruction workflow.

coral reefs, making it an ideal location for coral monitoring in the South Pacific. This research utilizes seabed images captured by an underwater camera system. To achieve high-resolution underwater mapping, we explore advanced deep learning and computer vision technologies. Specifically, we have established a workflow for coral reef 3D reconstruction based on underwater images, focusing on both sparse and dense reconstruction. We review and summarize the currently available methods, evaluating them qualitatively and quantitatively through comparative experiments. Additionally, we discuss their suitable application scenarios in light of their characteristics. We believe this study provides essential insights for future research in underwater mapping.

2. 3D Reconstruction Workflow with Deep Learning

Our workflow for high-resolution 3D reconstruction of coral reefs is illustrated in Figure 1 and mainly comprises four main stages: data collection and preparation, sparse reconstruction, dense reconstruction, and further processing. The first stage mainly involves the acquisition of high-resolution underwater images and auxiliary data. The images should be clear and exhibit overlap between different perspectives, as this forms the basis for 3D reconstruction. The auxiliary data, while not obligatory, may encompass measurements such as Ground Control Points (GCPs) utilized for georeferencing or camera poses derived from an Inertial Measurement Unit (IMU). Sparse reconstruction applied photogrammetric computer vision techniques to extract a set of sparse 3D points from input images. These 3D points correspond to feature points or keypoints within the scene. Through sparse reconstruction, image poses and the accurate structural information of the scene can be estimated simultaneously, even when dealing with unordered and uninformative images. The 3D points obtained from sparse reconstruction are sparse and insufficient to reflect the detailed structure of the scene. Therefore, it is necessary to perform dense reconstruction to generate denser 3D points. This process ultimately results in a dense point cloud or mesh model, facilitating the dense representation of 3D information within the scene, such as at millimetre-level resolution. Finally, based on specific requirements, different 3D products can be generated. For instance, texture models can be created through texture mapping, or orthomosaics can be generated through orthorectification. In this workflow, sparse and dense reconstruction play a crucial role in the accuracy, robustness, reliability, and visual effectiveness of 3D reconstruction. They constitute the core steps of the entire workflow, and are also the focal areas of photogrammetric computer vision and deep learning. Therefore, the following sections will provide detailed descriptions of these two components separately.

2.1 Sparse Reconstruction

At present, one of the most widely used frameworks for sparse reconstruction is the Structure-from-Motion (SfM) technology, which offers fast, low-cost and easy 3D surveys, particularly applied successfully in high-resolution topography for geoscience applications. There exist various SfM strategies, among which incremental SfM stands out as one of the most popular approaches, demonstrating suitable robustness, accuracy, and efficiency. Incremental SfM initiates processing with two images and gradually incorporates new images while continually optimizing (Schonberger and Frahm, 2016). This paper adopts the incremental SfM framework, combined with advanced photogrammetric computer vision and deep learning technologies, to establish a sparse reconstruction method for coral reefs. As illustrated in Figure 2, the method primarily comprises two stages: correspondence search and incremental reconstruction.

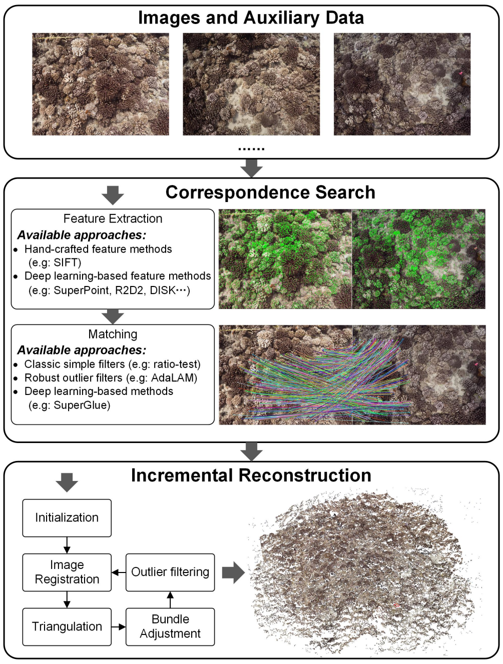

Images and Auxiliary Data
......
Correspondence Search
Feature Extraction
Available approaches:
Hand-crafted feature methods (e.g: SIFT)
Deep learning-based feature methods (e.g: SuperPoint, R2D2, DISK···)
Matching
Available approaches:
Classic simple filters (e.g: ratio-test)
Robust outlier filters (e.g: AdaLAM)
Deep learning-based methods (e.g: SuperGlue)
Incremental Reconstruction
Initialization
Image Registration
Outlier filtering
Triangulation
Bundle Adjustment

Figure 2. Structure-from-Motion with deep learning.

The first stage involves correspondence search, which identifies projections of the same points in overlapping images. For each coral reef image, the first step is to extract local features which are designed to be invariant under radiometric and geometric changes, ensuring their unique recognition across different images. Subsequently, feature matching is applied to discern images capturing the same area and establish feature correspondences across the images. The second stage is incremental reconstruction, utilizing feature correspondences to estimate the 3D relationships between 2D images. Building upon the outcomes of image matching, the inter-image overlap relationships are generated, and two adjacent images are selected for reconstruction initialization. Then, the remaining images are registered to the current model by using feature correspondences to triangulated points from previously registered images. The newly registered images not only encompass the points already observed but also contribute to the addition of new points through triangulation. However, due to the fact that image registration and triangulation are separate procedures, errors inevitably exist, leading to continuous error propagation and accumulation, which could potentially result in drift and failure of SfM reconstruction. Therefore, it is essential to optimize these observed values. Bundle adjustment (Triggs et al., 2000) is consequently used to minimize reprojection errors by jointly refining camera and point parameters through a non-linear optimization process. Through iterative computation based on the above process and outlier filtering, the scene structure and poses of registered images can be estimated accurately.

In the above sparse reconstruction process, the procedures of pose estimation, triangulation, and bundle adjustment have matured in research. The current major challenge lies in image matching, specifically in obtaining a sufficient quantity of accurate and reliable corresponding features. Due to the intricate and complex structure of coral reefs, the texture in captured images is more disorderly compared to typical images, which undoubtedly presents a considerable challenge for image matching. The process can be further divided into feature extraction and matching, as shown in Figure 2. For feature extraction, traditional hand-crafted local feature methods are based on a two-stage pipeline, first detecting keypoints and then generating local descriptors for each keypoint. Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) is the most representative and widely applied method, capable of extracting keypoints with scale and rotation invariance from images. There are also methods such as SURF (Bay et al., 2006) and KAZE (Alcantarilla et al., 2012). These methods utilize predefined criteria to extract points with certain characteristics from images. However, as these criteria may not be applicable to different scenarios, adjustments to algorithm parameters are often required. With the advancement of deep learning technology, many studies have attempted to overcome these limitations using learning-based approaches. Early methods such as LIFT (Yi et al., 2016) utilize keypoint labels obtained by existing hand-crafted methods to enhance the repeatability of keypoints by optimizing the objective functions. Convolutional neural network-based approaches emerged next, such as SuperPoint (DeTone et al., 2018). It uses a fully convolutional model to extract pixel-level interest point locations and associated descriptors from the input image, and also applies self-supervised learning to improve the generalizability of the model. R2D2 (Revaud et al., 2019) uses a Siamese decoding structure to generate repeatable and reliable features. DISK (Tyszkiewicz et al., 2020) employs reinforcement learning to optimize the model for more correct matches. ALIKED (Zhao et al., 2023) adopts a deformable descriptor head that learns the deformable positions of supporting features for each keypoint, thereby outputting robust and accurate descriptors. There are also detector-free local feature matching methods like LoFTR (Sun et al., 2021), but they cannot be directly used in modern SfM systems because they do not explicitly extract keypoints and descriptors.

For feature matching, the difficulty lies in how to accurately match the features and minimize the number of mismatches. The classic ratio-test (Lowe, 2004) matches features based on the similarity between descriptors. While this method is simple and effective, it often results in a large number of outliers, leading to registration failure. To achieve robust feature matching, researchers have studied various strategies. On the one hand, robust outlier filters are designed to eliminate outliers while obtaining more correct matches. For example, AdaLAM (Cavalli et al., 2020) takes the keypoint positions and corresponding descriptors as input and achieves robust matching using an adaptive strategy. On the other hand, methods based on deep learning have been utilized, such as SuperGlue (Sarlin et al., 2020), which takes images and features as input, employing graph neural networks and attention mechanisms to obtain accurate matches.

2.2 Dense Reconstruction

Based on the camera poses estimated by sparse reconstruction, dense reconstruction techniques can be applied to generate a dense point cloud model or mesh model of the scene, as shown in Figure 3. For coral reefs, the primary significance of dense reconstruction is to recover the fine structure of coral reefs. Due to the presence of structures like tentacles in coral reefs, images often encounter issues such as occlusion and texture repetition, imposing high demands on dense reconstruction. Over the past two decades, many excellent algorithms have emerged, ranging from traditional multi-view stereo (MVS) to deep learning-based MVS, and more recently, rapidly developing methods based on Neural Radiance Fields (NeRF). These methods vary in terms of accuracy, robustness and efficiency. The ongoing advancements in this field exemplify the continuous evolution and innovation in 3D reconstruction techniques, thereby opening up new possibilities for underwater mapping.

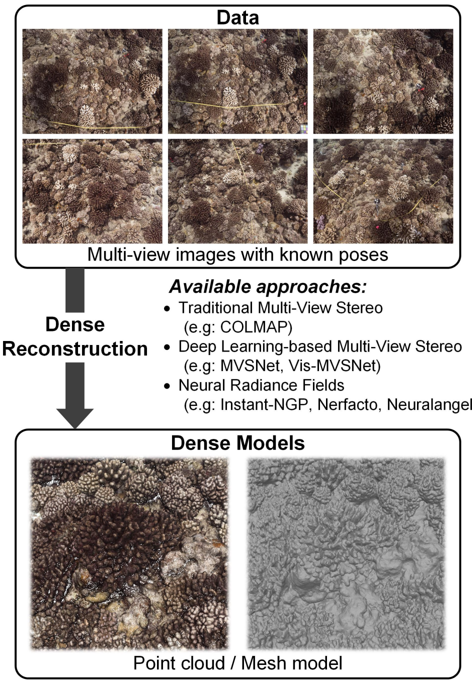

Data
Multi-view images with known poses
Dense Reconstruction
Available approaches:
• Traditional Multi-View Stereo (e.g: COLMAP)
• Deep Learning-based Multi-View Stereo (e.g: MVSNet, Vis-MVSNet)
• Neural Radiance Fields (e.g: Instant-NGP, Nerfacto, Neuralangelo)
Dense Models
Point cloud / Mesh model

Figure 3. Dense reconstruction with deep learning.

The commonly used traditional MVS adopts a depth-map-based strategy, obtaining depth maps corresponding to images through multi-view matching. Subsequently, it fuses all the depth maps to ultimately generate a dense 3D point cloud. This approach is flexible, concise, and suitable for reconstructing the 3D structure of large-scale scenes. It is a mature study and has been widely applied. Taking COLMAP (Schonberger and Frahm, 2016) as an example, its key steps involve the cost computation for multi-view matching. It uses bilateral weighted Normalized Cross Correlation (NCC) to calculate the similarity between image patches, thus obtaining matching costs. The optimization methods are then applied to minimize the matching cost for depth map generation.

Driven by deep learning technology, deep learning-based MVS have emerged. In this category of methods, MVSNet (Yao et al., 2018) is a pioneering approach that takes multi-view images and corresponding camera parameters as input to predict depth maps for the corresponding images. Specifically, it divides the images into one reference image and others as source images. A differentiable homography warping operation is employed to build 3D cost volumes from 2D feature maps, enabling the integration of camera parameters in network training. Ultimately, this results in the generation of a high-quality dense depth map for the reference image. Finally, a high-quality dense depth map of the reference image can be generated. There are several derivative algorithms based on MVSNet, among which Vis-MVSNet (Zhang et al., 2023) is one of the best in terms of overall performance, using a coarse-to-fine strategy to achieve multi view depth map estimation. An innovative aspect lies in its consideration of pixel visibility. To mitigate the impact of unmatched pixels, it generates an uncertainty map to estimate per-pixel visibility. The uncertainty is used as a weighting guidance, fusing the latent volume which is further regularized into a probability volume and regresses to the final depth estimation.

Another category of methods that has emerged in recent years is based on Neural Radiance Fields (NeRF) (Mildenhall et al., 2021). Unlike traditional 3D reconstruction methods, NeRF methods can represent real-world complex geometry and appearance using a neural network, storing 3D scene information in the parameters of the network. The typical input comprises images with known poses, and the output is the weights of the network. These methods are theoretically able to realize a finer representation of the continuous scene. To address the slow speed of NeRF, Instant-NGP (Müller et al., 2022) applied multiresolution hash encoding to reduce computational complexity while maintaining accuracy. This approach also facilitates parallel implementation on GPUs, thereby significantly improving efficiency. Nerfacto (Tancik et al., 2023) integrates improvements from multiple previous methods, allowing the model to balance accuracy and efficiency. With a modular design, it facilitates easy improvements in subsequent developments. Additionally, it introduces the Python framework Nerfstudio, supporting the output of results in the form of point clouds or mesh models. There are methods specifically designed for multi-view 3D reconstruction, such as Neuralangelo (Li et al., 2023), which utilizes Instant-NGP as a neural Signed Distance Function (SDF) representation of the underlying 3D scene and is optimized from multi-view image observations via neural surface rendering. To enhance the effectiveness of multi-resolution hash encoding, it uses numerical gradients to compute higher-order derivatives, and a progressive optimization schedule is adopted to recover structures at different levels of detail, ultimately achieving high-quality surface reconstruction.

3. Experiments and Discussion

3.1 Research Data

The data used in this research is supported by the Moorea Island Digital Ecosystem Avatar (IDEA) project, consisting of high-resolution underwater coral reef images captured in the same area in August 2018 and August 2019. Specifically, in 2018, 523 images were captured, and in 2019, 323 images were captured. The images were acquired along pre-planned routes, with overlap rates between adjacent images mostly ranging from 70% to 85%, enabling multi-view 3D reconstruction. The camera system includes a PANASONIC LUMIX GH5S camera body (resolution of 3680×2760 pixels) and a wide-angle lens Lumix G 14 mm f/2.5.

3.2 Image Matching and SfM Reconstruction

This section focuses on the impact of different feature extraction and matching methods on image matching and SfM reconstruction. In the comparative experiments, feature extraction methods include SIFT (Lowe, 2004), SuperPoint (DeTone et al., 2018), R2D2 (Revaud et al., 2019), DISK (Tyszkiewicz et al., 2020), and ALIKED (Zhao et al., 2023), while feature matching methods include classic ratio-test (Lowe, 2004), AdaLAM (Cavalli et al., 2020), and SuperGlue (Sarlin et al., 2020).

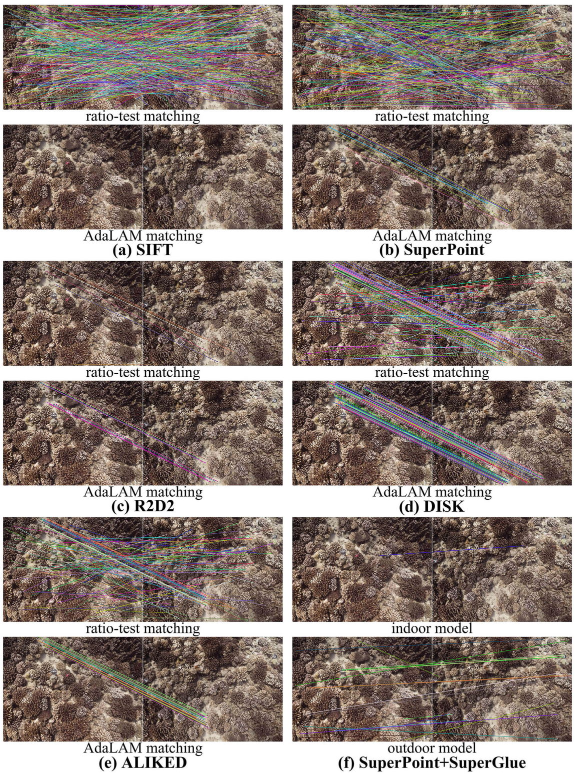
For image matching, each feature extraction method is applied to extract 8000 features from coral reef images. Subsequently, ratio-test and AdaLAM are employed for feature matching. Additionally, considering SuperGlue works particularly well with SuperPoint (Sarlin et al., 2020), SuperPoint features are also matched using SuperGlue. SuperGlue offers two pre-trained weight models, with one tailored for indoor environments (referred to as SG (in)) and the other designed for outdoor settings (referred to as SG (out)). The ratio-test is applied with a mutual nearest neighbor check, and the ratio is set to 0.9. After preliminary experimental analysis, we found that when there is substantial overlap and only minor translation or rotation is present between a pair of images, various methods can generally obtain a sufficient number of correct matches. However, significant differences arise among different methods when the overlap is low or there is a large rotation. Figures 4 and 5 illustrate the results of image matching under two challenging scenarios.

As shown in Figure 4, when there is a low overlap between two images, image matching becomes a challenging task. When using the ratio-test for feature matching, some correct matches can be obtained, but there are also many mismatches. While a limited number of mismatches can be filtered out during reconstruction, an excessive amount is problematic. Particularly, the matching results of SIFT and SuperPoint are not suitable for reconstruction, whereas the performance of the other methods is slightly better. When using AdaLAM, there is almost no existence of mismatches. SIFT features, due to a high error rate, struggle to obtain correct matches, while the other four deep learning-based local features, with the support of AdaLAM, successfully achieve correspondence search, indicating the effectiveness of the AdaLAM algorithm in outlier filtering.

The scenario depicted in Figure 5 is markedly different, involving approximately a 90-degree rotation between the two images. It is evident that the use of deep learning-based features fails to yield correct matches, primarily due to the regular CNNs lacking equivariance to rotation. Despite ALIKED's specific design to enhance rotational invariance, it proves insufficient for handling the current data. In contrast, SIFT demonstrates significantly better results, generating a majority of correct matches and a small number of mismatches when using the ratio-test. When employing AdaLAM, there are nearly no mismatches, and a lot of correct matches are obtained. Therefore, in such cases, using SIFT is a more suitable choice. The SIFT descriptor achieves rotation invariance by assigning a consistent orientation to each keypoint based on local image properties. This is a capability lacking in current deep learning-based

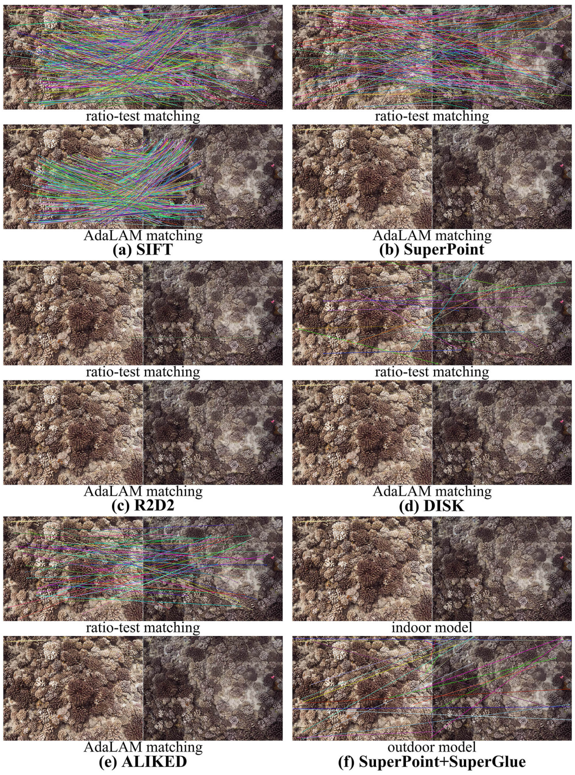
Data	Feature	SIFT		SuperPoint		R2D2		DISK		ALIKED		SuperPoint	
	Match	RT	Ada	RT	Ada	RT	Ada	RT	Ada	RT	Ada	SG(in)	SG(out)
2018	$nImage$	523	523	495	523	192	275	518	520	523	523	303	523
	$Feat$	3739	3368	3088	2832	993	1953	3638	4112	2721	3084	3480	4111
	$nPoint$	536k	364k	457k	369k	49k	128k	448k	515k	354k	401k	286k	552k
	$Track$	3.65	4.83	3.34	4.01	3.87	4.2	4.20	4.15	4.02	4.02	3.68	3.89
	$Error$	0.62	0.79	0.68	0.83	1.09	1.20	1.06	1.10	0.59	0.67	0.89	0.98
2019	$nImage$	323	323	312	316	190	286	312	315	317	317	290	321
	$Feat$	3958	3625	2898	2735	1093	2210	3850	4293	2795	3217	3384	3275
	$nPoint$	332k	234k	260k	208k	52k	145k	273k	310k	212k	246k	309k	334k
	$Track$	3.85	4.99	3.47	4.14	3.99	4.35	4.4	4.36	4.16	4.13	3.17	3.14
	$Error$	0.55	0.70	0.60	0.73	1.03	1.18	1.03	1.08	0.51	0.58	1.24	1.27

Table 1. Various metrics of the reconstruction results of different methods. "RT" represents ratio-test, and "Ada" represents AdaLAM.



(a) SIFT (b) SuperPoint (c) R2D2 (d) DISK (e) ALIKED (f) SuperPoint+SuperGlue

Figure 4. Qualitative visual inspection of underwater image matching with low overlap.



(a) SIFT (b) SuperPoint (c) R2D2 (d) DISK (e) ALIKED (f) SuperPoint+SuperGlue

Figure 5. Qualitative visual inspection of underwater image matching with a large rotation.

features and represents a crucial area for future research. As for SuperGlue, in both sets of data, the indoor model is almost unable to generate any matches, and the outdoor model only produces mismatches. This is likely due to SuperGlue being a supervised learning method, and its training data lacks scenes similar to those in the given scenarios, and these scenarios are inherently challenging, so their performance is poor.

For SfM reconstruction, we utilized the open-source software COLMAP (Schonberger and Frahm, 2016) to implement incremental reconstruction. The configuration for feature extraction and matching methods remains the same, and the number of features of each image is still limited to 8000. We perform reconstruction using the images collected from Plot18 in 2018 and 2019, respectively. Figure 6 shows the SfM reconstruction results using SIFT features with AdaLAM, including the point cloud of the scene and camera poses. To compare and evaluate the effectiveness of SfM reconstruction, we calculated 5 metrics: $nImage$, $Feat$, $nPoint$, $Track$, and $Error$. $nImage$ represents the number of aligned images, $Feat$ is the average number of features successfully used for triangulation per image, $nPoint$ is the number of reconstructed 3D points (1k=1000), $Track$ is the mean repeat observation number of 3D point, and $Error$ denotes the average reprojection error of keypoints. The quantitative results are shown in Table 1.

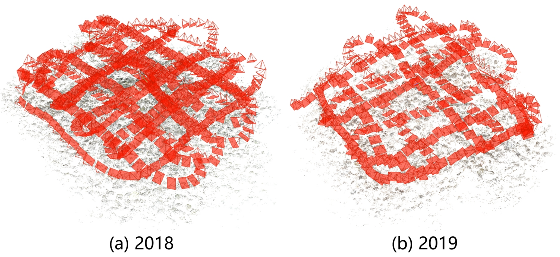


Figure 6. Visualization of our SfM reconstruction results.

Overall, the comprehensive performance of SIFT is excellent. Regardless of the matching method used, all images can be successfully aligned, and the reprojection error is only slightly larger than ALIKED. When using AdaLAM, both $Feat$ and

$nPoint$ decrease, but $Track$ increases, becoming the highest among the methods. This indicates that AdaLAM can connect more features of different images, making the reconstruction more stable. On the contrary, R2D2 performs the worst. It has the fewest successfully aligned images in both datasets, with the lowest values for $Feat$ and $nPoint$. Even when using AdaLAM, only about 2000 features are successfully matched, and the reprojection error is also high. Among SuperPoint, DISK, and ALIKED, ALIKED has a relatively good overall performance. It not only has the highest number of successfully aligned images but also the smallest reprojection error. This suggests that among deep learning-based feature methods, ALIKED is more suitable for high-precision SfM reconstruction. DISK and SuperPoint are close in terms of $nImage$, but DISK features have a higher $Feat$ and $Track$. However, the reprojection error of SuperPoint is slightly lower. As for the results of SuperGlue, there is a notable distinction between the indoor and outdoor models. Compared to non-learning methods, the indoor model performs poorly and struggles to align all images. The outdoor model shows improvement, but the reprojection error is relatively high, indicating less accurate feature matching. It is noteworthy that, compared to the ratio-test, AdaLAM is relatively less accurate, while generally improving matching robustness and increasing the number of repeated observations. It may match keypoints that are not the most accurate correspondences but rather nearby keypoints. This is related to the geometric assumptions within the algorithm, sacrificing a bit of accuracy for better stability.

In summary of the above experimental results, SIFT stands out as the most practical local feature method, showcasing excellent precision and reliability. On the other hand, deep learning-based methods currently face challenges in effectively handling coral reef image matching. Despite this, the progression from early methods like SuperPoint to the recent ALIKED indicates a continuous enhancement in reliability and accuracy. Therefore, it is reasonable to anticipate the development of even more outstanding methods in the future. As for feature matching, AdaLAM exhibits remarkable robustness, significantly addressing issues related to excessive outliers and enhancing the stability of SfM reconstruction.

3.3 Dense Reconstruction

Based on the accurately reconstructed camera poses derived from sparse reconstruction, the intricate fine structure of coral reefs can be estimated through dense reconstruction. This section conducts comparative experiments on the three categories of dense reconstruction methods mentioned in Section 2.2. The first is traditional MVS, and we use the dense reconstruction functionality in COLMAP (Schonberger and Frahm, 2016). The second is deep learning-based MVS, and we employ Vis-MVSNet (Zhang et al., 2023). The third is the recently popular method based on NeRF, and we use Instant-NGP (Müller et al., 2022), Nerfacto (Tancik et al., 2023) and Neuralangelo (Li et al., 2023). It should be noted that the direct outputs of COLMAP and Vis-MVSNet are dense point clouds. As for Instant-NGP and Nerfacto, they are not designed for generating point clouds or mesh models, but it is still possible to get point clouds by some means (Tancik et al., 2023). Neuralangelo, on the other hand, can output mesh models directly. For the purpose of visualization, here we use Poisson surface reconstruction (Kazhdan et al., 2006) to transform point clouds into mesh models. Since the three NeRF-based methods applied here are not suitable for large-scale scenes, we select a subset of images (42 images) from a specific region for experiments. We adjust the parameters for each method to obtain the best possible results, and the visualization of partial dense reconstruction results is shown in Figure 7.

The results of COLMAP and Vis-MVSNet are similar, but Vis-MVSNet produces relatively denser results, albeit with slightly more noise, showcasing its commendable generalization. Meanwhile, COLMAP's result appears overly smoothed. The results of Instant-NGP and Neuralangelo are the least satisfactory in general. The point cloud obtained by Instant-NGP contains many outliers, making it difficult to correctly reconstruct a mesh model. On the other hand, the results of Neuralangelo are excessively smooth, failing to capture the fine structure of coral reefs, and the contrast of its mesh texture is abnormally high. In comparison, Nerfacto's result is significantly better, generally able to reconstruct the intricate details of coral reefs, especially the tentacles of the corals. However, there are still quite a few outliers in the point cloud, leading to surface irregularities in the mesh model. While NeRF-based methods often exhibit noise or over-smoothness in their results, indicating a current limitation in effectively handling data noise and suggesting a need for enhanced reliability in subsequent improvements, they have already achieved acceptable results. With ongoing enhancements, it is anticipated that satisfying outcomes will be achieved in the near future. Additionally, a major factor contributing to poor performance is that the viewpoints of the images are mainly downward-looking and lack side-view data. NeRF captures 3D scene information by modeling the volumetric scene as a continuous function that predicts the color and opacity of any given 3D point. Lower image overlap implies fewer corresponding 2D projections of 3D points across different images. This results in fewer constraints in the NeRF optimization process, and inadequate information for NeRF to accurately model the geometry and appearance of the underlying scene. In practice, improving the quality of 3D reconstruction can be achieved by increasing image overlap and coverage.

In 3D reconstruction, dense reconstruction has always been the most time-consuming task. Therefore, we also test the execution time of different methods, where COLMAP is implemented in C++, and the other methods are implemented in Python. Instant-NGP and Nerfacto are implemented using Nerfstudio (Tancik et al., 2023). All experiments are conducted using an NVIDIA Geforce RTX 3090 GPU. We test the time taken by different dense reconstruction methods used in Figure 7, and the results are shown in Table 2. Vis-MVSNet exhibits the fastest speed, significantly outperforming other methods. Instant-NGP and Nerfacto have similar runtime, while Neuralangelo is considerably slower. The runtime of NeRF-based methods is directly correlated with the number of training iterations set. Typically, a larger number of iterations result in better network fitting, although the improvement becomes less pronounced over time. As for how to better balance between effectiveness and efficiency, further research is needed.

Method	Time (second)
COLMAP	991
Vis-MVSNet	224
Instant-NGP	982
Nerfacto	921
Neuralangelo	34708

Table 2. The runtime in seconds for small-scale scenes.

In addition, we also test the runtime of COLMAP and Vis-MVSNet in reconstructing large-scale scenes. We apply these

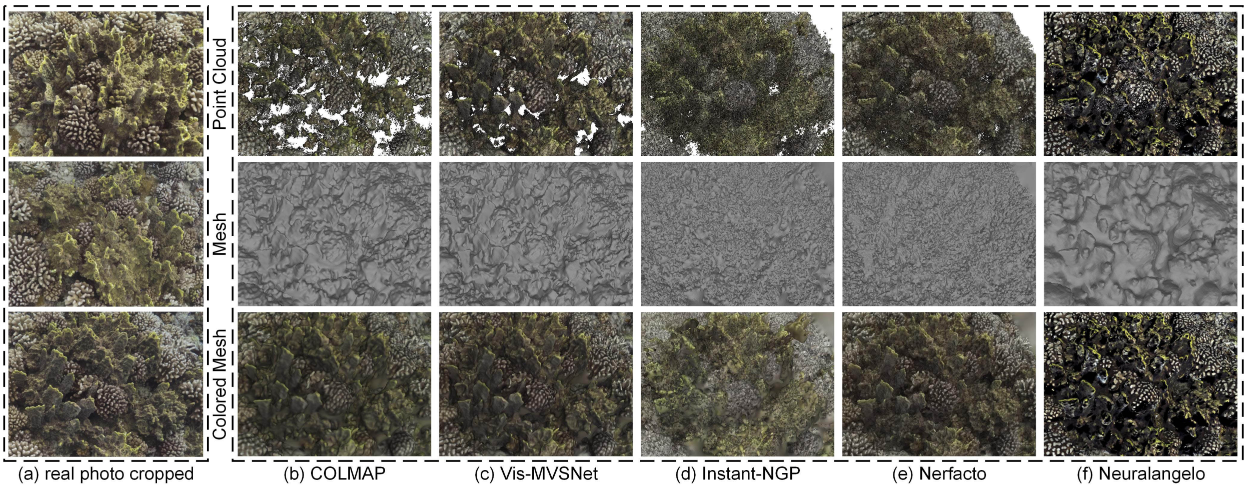


Figure 7. Comparison of dense reconstruction results.

two methods for the dense reconstruction of the entire area. As they are both depth map-based methods, the entire process can be divided into three steps: I. Data preprocessing, II. Depth map generation, and III. Fusion of depth maps to obtain point clouds. The specific execution times are presented in Table 3. The total time taken by Vis-MVSNet is significantly less than that of COLMAP, mainly due to the time-consuming tasks involved in the MVS process within COLMAP, like iterative computations. The most time-consuming step in Vis-MVSNet is also depth map generation. Nevertheless, its preprocessing stage is also resource-intensive, mainly due to the conversion of sparse reconstruction results from COLMAP into a format suitable for the network input. Overall, Vis-MVSNet achieves appropriate results with excellent operational efficiency, while NeRF-based methods demonstrate outstanding potential in fineness but come with a longer processing time.

Data	Step	COLMAP	Vis-MVSNet
2018	I	30	724
	II	10364	1601
	III	148	232
	Total	10542	2557
2019	I	21	215
	II	5131	983
	III	108	70
	Total	5260	1268

Table 3. The runtime in seconds for large-scale scenes.

In future practical applications, the reconstruction approach may not necessarily rely on a specific method alone. Instead, it may be beneficial to integrate the strengths and weaknesses of various methods in accordance with the specific requirements of the task, achieving a balance between effectiveness and efficiency. For example, a coarse-to-fine strategy can be adopted for large-scale underwater mapping. Specifically, after sparse reconstruction, initial dense reconstruction of the terrain could be rapidly achieved using deep learning-based MVS methods, resulting in a preliminary dense model. Subsequently, based on this model and task requirements, more densely sampled and higher-resolution data could be collected in areas of interest, such as coral reefs. Finally, fine-grained dense reconstruction could be carried out using NeRF-based methods.

In summary, deep learning-based dense reconstruction methods are not inferior to, and in some aspects even surpass, traditional MVS in fineness and efficiency. This indicates that emerging computer vision and deep learning technologies have achieved remarkable advancements, with substantial room for improvement. Looking ahead, these advancements hold the potential to significantly advance the field of high-resolution underwater mapping, leading to more in-depth and comprehensive outcomes.

4. Conclusions

In this paper, we take the coral reefs of Moorea Island as an example and elaborate in detail on how current emerging photogrammetric computer vision and deep learning technologies can be applied in high-resolution underwater mapping, in response to the limitations of traditional methods. Combining the current research, we establish an improved workflow for 3D reconstruction of coral reefs. Delving into both sparse and dense reconstruction, this paper conducts an analysis and summary of classical and state-of-the-art methods, elucidating how to apply them concretely. Through experiments on actual coral reef images, qualitative and quantitative evaluations of these methods are performed in terms of accuracy, reliability, efficiency, etc. Building upon this foundation, we analyze their strengths and limitations, confirm the promising prospects of cutting-edge methods, and propose feasible directions for improvement, providing outlooks for future research and applications.

Acknowledgements

The data for this study was obtained from the Moorea IDEA project. This work was supported by the U.S. National Science Foundation (grant number OCE 2224354 and earlier awards) for the Moorea Coral Reef LTER; the Gordon and Betty Moore Foundation in support of the MCR LTER; the Italian Minister of University and Research (grant number PNRA18_00263-B2); the National Natural Science Foundation of China (NSFC-41901407); the Institute of Theoretical Physics, ETH Zurich. We especially would like to thank Prof. Matthias Troyer for his financial and scientific support provided. Special thanks are extended to the ETH Zurich Photogrammetry and Remote Sensing Group and the LIESMARS of Wuhan University.

References

Alcantarilla, P. F., Bartoli, A., Davison, A. J., 2012. Kaze features. *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, Springer, 214–227.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, Springer, 404–417.

Bowen, B. W., Rocha, L. A., Toonen, R. J., Karl, S. A., 2013. The origins of tropical marine biodiversity. *Trends in ecology & evolution*, 28(6), 359–366.

Casella, E., Collin, A., Harris, D., Ferse, S., Bejarano, S., Parravicini, V., Hench, J. L., Rovere, A., 2017. Mapping coral reefs using consumer-grade drones and structure from motion photogrammetry techniques. *Coral Reefs*, 36, 269–275.

Cavalli, L., Larsson, V., Oswald, M. R., Sattler, T., Pollefeys, M., 2020. Adalam: Revisiting handcrafted outlier detection. *arXiv preprint arXiv:2006.04250*.

Character, L., Ortiz Jr, A., Beach, T., Luzzadder-Beach, S., 2021. Archaeologic machine learning for shipwreck detection using lidar and sonar. *Remote Sensing*, 13(9), 1759.

Collin, A., Ramambason, C., Pastol, Y., Casella, E., Rovere, A., Thiault, L., Espiau, B., Siu, G., Lerouvreur, F., Nakamura, N. et al., 2018. Very high resolution mapping of coral reef state using airborne bathymetric LiDAR surface-intensity and drone imagery. *International journal of remote sensing*, 39(17), 5676–5688.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Guo, T., Capra, A., Troyer, M., Grün, A., Brooks, A. J., Hench, J. L., Schmitt, R. J., Holbrook, S. J., Dubbini, M., 2016. Accuracy assessment of underwater photogrammetric three dimensional modelling for coral reefs. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41(B5), 821–828.

Hughes, T. P., Kerry, J. T., Álvarez-Noriega, M., Álvarez-Romero, J. G., Anderson, K. D., Baird, A. H., Babcock, R. C., Beger, M., Bellwood, D. R., Berkelmans, R. et al., 2017. Global warming and recurrent mass bleaching of corals. *Nature*, 543(7645), 373–377.

Kazhdan, M., Bolitho, M., Hoppe, H., 2006. Poisson surface reconstruction. *Proceedings of the fourth Eurographics symposium on Geometry processing*, 7, 0.

Li, Z., Müller, T., Evans, A., Taylor, R. H., Unberath, M., Liu, M.-Y., Lin, C.-H., 2023. Neuralangelo: High-fidelity neural surface reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8456–8465.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60, 91–110.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4), 1–15.

Price, D. M., Robert, K., Callaway, A., Lo Lacono, C., Hall, R. A., Huvenne, V. A., 2019. Using 3D photogrammetry from ROV video to quantify cold-water coral reef structural complexity and investigate its influence on biodiversity and community assemblage. *Coral Reefs*, 38, 1007–1021.

Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., Humenberger, M., 2019. R2D2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*.

Rossi, P., Castagnetti, C., Capra, A., Brooks, A. J., Mancini, F., 2020. Detecting change in coral reef 3D structure using underwater photogrammetry: critical issues and performance metrics. *Applied Geomatics*, 12, 3–17.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. Loftr: Detector-free local feature matching with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922–8931.

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A. et al., 2023. Nerfstudio: A modular framework for neural radiance field development. *ACM SIGGRAPH 2023 Conference Proceedings*, 1–12.

Triggs, B., McLauchlan, P. F., Hartley, R. I., Fitzgibbon, A. W., 2000. Bundle adjustment—a modern synthesis. *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, Springer, 298–372.

Tyszkiewicz, M., Fua, P., Trulls, E., 2020. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33, 14254–14265.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo. *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Yi, K. M., Trulls, E., Lepetit, V., Fua, P., 2016. Lift: Learned invariant feature transform. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, Springer, 467–483.

Zhang, J., Li, S., Luo, Z., Fang, T., Yao, Y., 2023. Vismvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1), 199–214.

Zhao, X., Wu, X., Chen, W., Chen, P. C., Xu, Q., Li, Z., 2023. ALIKED: A Lighter Keypoint and Descriptor Extraction Network via Deformable Transformation. *IEEE Transactions on Instrumentation and Measurement*.

Zhong, J., Li, M., Zhang, H., Qin, J., 2023. Fine-Grained 3D Modeling and Semantic Mapping of Coral Reefs Using Photogrammetric Computer Vision and Machine Learning. *Sensors*, 23(15), 6753.