

RIFT-Map: Flexible Transect-Tape Arclength Constraints for Metric Monocular Underwater Mapping

First Author¹, Second Author², and Third Author^{1,*}

Running title: Tape Arclength Constraints for Monocular Mapping

Abstract—Metric three-dimensional maps of coral-reef transects are needed to measure colony size and structural complexity over time. However, a single survey camera recovers the reef only up to an unknown scale that can also change along the pass. The flexible tape that divers lay along a transect could supply this scale but measures length along its curve rather than the straight distance between two readings. Therefore, we propose the Robust Integrated Flexible-Tape Mapping method to treat the transect tape as an observed curve of known material length. The method admits a tape interval only when a multiview support test confirms its interior. It rescales each segment of the scene-only reconstruction by the curve lengths of its own tape intervals. A robust consensus of curve lengths then sets the map scale. Bundle adjustment refines cameras, scene points and tape curves jointly with continuous arclength factors. On a straight tape in a pool, 30 of the 31 admitted tape curves stayed within 0.53% of their chord. We then mapped a 50-m coral transect near Koh Tao in Thailand with one action camera. In a leave-one-interval-out validation on 31 tape intervals along the whole pass, the method had a mean absolute length discrepancy of 1.37% and a root-mean-square discrepancy of 2.04%. These were the lowest values of six compared methods. The established three-camera workflow scaled by one scale bar reached 3.51% and 4.78%. In 100 simulated repetitions endpoint constraints enlarged the map of a curved tape by 3.16%. Curve lengths kept the scale error at 0.23%. These results show that a modeled tape curve with checked image support can provide the metric reference of monocular reef mapping.

Index Terms—Bundle adjustment, coral reef monitoring, monocular mapping, transect tape, underwater photogrammetry.

I. INTRODUCTION

CORAL reefs are surveyed along permanent or temporary transects. Divers lay a flexible measuring tape on the reef and record benthic cover or coral colonies along it [1], [2]. Structure-from-motion photogrammetry now turns images of such transects into three-dimensional reef models [3]–[5]. These models give rugosity, colony dimensions and their change over time [6]–[8]. Colony dimensions, areas and volumes depend on the metric scale of the model. Their change over time also depends on a scale that stays consistent along the transect. A single camera cannot observe this scale [9]. Its reconstruction can also drift in scale along a long path [10]. A long sequence can further break into segments

whose relative scale rests on a few shared features. Surveys therefore add scale bars, laser points, surveyed control points, stereo baselines or acoustic ranges [11]–[15]. The transect tape already lies in the images of most surveys. Its printed readings give known lengths at many places along the reef. However, the tape follows the relief of the reef (see Fig. 1). A tape interval that bends by 90° has a material length 11.1% larger than the straight distance between its ends. A pipeline that treats this interval as a rigid scale bar enlarges the map by the same 11.1%. Blur, occlusion and short visibility also leave parts of the tape without reliable image evidence.

Previous studies have built large seafloor reconstructions from the cameras of underwater vehicles [16], [17]. Xu et al. [15] recovered the absolute scale of single-camera odometry from a sonar range to the seabed. Purser et al. [14] sized seafloor features with laser points of known spacing in the images of a towed camera system. Mizuno et al. [18] towed a six-camera array over coral reefs and estimated coral cover from its three-dimensional models. Wang et al. [19] mounted three action cameras on one frame and used tape markings as manually registered control points in Metashape. In this way the tape became a source of metric control for reef photogrammetry. However, the workflow of Wang et al. uses the tape only at isolated marker points. The other workflows fix the scale with additional sensors or a rigid camera baseline. Sünderhauf and Protzel [20] introduced switchable constraints that let a pose-graph optimizer down-weight inconsistent factors. This technique can be adapted to tape factors whose image evidence is unreliable.

Motivated by this, we propose the Robust Integrated Flexible-Tape Mapping (RIFT-Map) method for metric monocular reconstruction of reef transects. RIFT-Map models each visible tape interval as a local three-dimensional curve whose arclength equals the difference of its printed readings. An interval is admitted only when every interior node of its curve passes a multiview support test with at least three consistent views. Join rescaling corrects the relative scale of the segments of the scene-only reconstruction from the curve lengths of their own intervals. A robust consensus of curve lengths then sets the map scale. Continuous arclength factors with reliability switches enter bundle adjustment. To our knowledge, no published method constrains underwater bundle adjustment with the material arclength of a flexible tape. We first validate the processing chain of RIFT-Map in a pool where a tape lies straight on the floor. We then evaluate its metric accuracy along a 50-m coral transect recorded near

¹Affiliation of the first and third authors, City, Country.

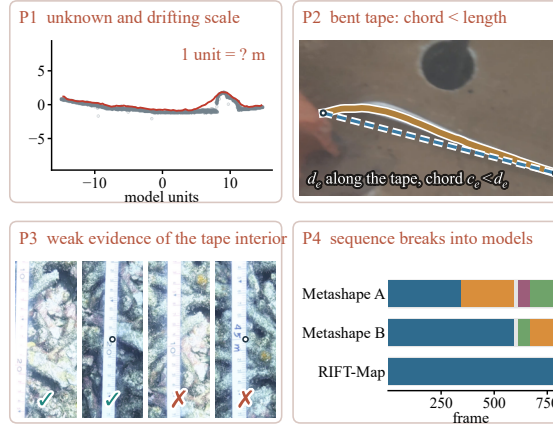
²Affiliation of the second author, City, Country.

*Corresponding author: Third Author (e-mail: name@institution.edu).

(a) Monocular survey along a transect tape



(b) Problems



(c) RIFT-Map output

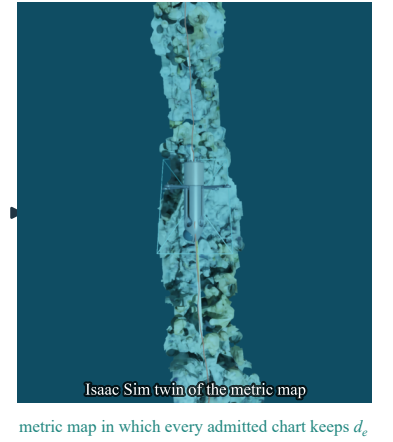


Fig. 1. Workflow and problems of metric monocular transect mapping. (a) A single camera records the transect tape in a pool and on a coral reef. (b) Four problems prevent a direct use of the tape as a metric reference. P1 is the unknown and drifting scale of a monocular reconstruction. P2 is the bend of the tape over the relief, which makes the chord shorter than the material length. P3 is the weak image evidence of the tape interior. P4 is the break of a long sequence into several models. (c) RIFT-Map resolves them in one metric map whose tape intervals keep their material lengths.

Koh Tao in Thailand.

Our contributions are summarized below.¹

- We formulate the transect tape as a flexible material curve. We derive the scale bias that an endpoint constraint introduces on a bent interval. We propose a tape chart graph that recovers interior tape nodes from semantic tape masks and verifies them with a multiview support test.
- We propose a metric backend that couples the tape curves to the reconstruction. Join rescaling corrects the relative scale of segments that the scene-only reconstruction joined through few shared features. A robust consensus of curve lengths sets the global scale. Continuous Catmull-Rom arclength factors with reliability switches and an evidence-scaled trajectory prior then enter bundle adjustment.
- We validate RIFT-Map in a pool test, a field test and a controlled simulation. On the straight pool tape 30 of the 31 admitted curves stayed within 0.53% of their chord. On the coral transect a leave-one-interval-out validation along the whole pass gave the lowest length discrepancy of six methods. In 100 simulated repetitions the curve-length model removed the 3.16% scale bias of endpoint constraints on a curved tape.

II. RELATED WORKS AND EXISTING GAPS

A. Optical Reef and Seafloor Mapping

Pizarro et al. [16] reconstructed seafloor submaps with navigation-aided structure from motion and refined them by bundle adjustment. Johnson-Roberson et al. [17] turned the stereo imagery of an autonomous underwater vehicle into textured seafloor meshes. Mizuno et al. [18] combined a towed camera array with deep-learning segmentation to estimate coral cover at sea. Menacho et al. [5] improved reef

reconstruction by merging several cameras and correcting the image colors. Sauder et al. [21] mapped a 100-m transect from consumer-camera video with learning-based structure from motion. On shallow reefs, Burns et al. [3] measured the three-dimensional complexity of coral colonies from structure-from-motion models. Leon et al. [4] compared roughness metrics along a 250-m reef-flat transect. Storlazzi et al. [6] derived rugosity from existing reef video as an alternative to the chain-and-tape method. Bayley and Mogg [22] gave a protocol for the large-scale analysis of reefs with structure from motion.

The accuracy of these models depends on their resolution and on their metric reference. Figueira et al. [7] found that photogrammetric models of coral colonies deviated from laser references by only 1 to 3 mm. Their rugosity was nevertheless 8% to 37% lower than that of the references. Bryson et al. [11] quantified the bias of complexity measures on artificial reefs of known structure. Nocerino et al. [13] anchored diver photogrammetry to a permanent underwater geodetic network. Neyer et al. [23] and Rossi et al. [8] showed that the quality of this network limits the detection of coral growth. These studies fix the scale with scale bars, laser points, geodetic control points or calibrated camera baselines. None of them used the transect tape as a distributed metric constraint.

B. Metric Scale for Monocular Underwater Reconstruction

A monocular reconstruction is defined only up to a similarity transform [9]. Strasdat et al. [10] corrected the scale drift of monocular simultaneous localization and mapping with similarity constraints at loop closures. ORB-SLAM3 [24] adds inertial, stereo and depth inputs that make the scale observable. Ferrera et al. [25] recorded monocular, inertial and pressure data from remotely operated vehicles. Rahman et al. [26] fused sonar, visual, inertial and depth measurements. Their system initialized the scale from the depth. Istenić et al. [12] evaluated monocular reconstruction strategies with laser scalars. Weakly

¹The project page is available at <https://panzerjagerwang.github.io/RIFT-Map/>.

connected sequential surveys bent into a broad dome unless navigation data entered the optimization. Their best strategy had scale errors of about 1% in the center of the models and below 5% at their edges.

Refraction at the housing port changes the effective camera geometry [27]. Kahmen et al. [28] showed that absorbing flat-port refraction into a standard camera model causes systematic errors in bundle adjustment. She et al. [29] modeled refraction explicitly in structure from motion. Commercial photogrammetry software fixes the scale with markers and scale bars of known length. Shape-from-template methods recover an inextensible surface from one image of it and its known template [30]. They reconstruct the deforming object itself rather than the rigid scene around it. None of these approaches used the known length of a deformable curve to fix the scale of a rigid scene.

C. Robust Factors in Bundle Adjustment

Bundle adjustment refines cameras and structure jointly by nonlinear least squares [31]. COLMAP implements incremental structure from motion with this machinery [32]. Robust kernels such as the Huber loss limit the influence of outlying residuals [33]. Sünderhauf and Protzel [20] attached a continuous switch variable to each loop-closure factor. Agarwal et al. [34] proposed dynamic covariance scaling, which rejects outlier constraints without additional variables.

These methods weight point, pose or loop-closure factors. None of them applied a reliability switch to the arclength of an observed curve.

III. PROBLEM FORMULATION AND PRELIMINARIES

A. Problem Definition

We estimate a metric reconstruction of a reef transect from the image sequence of one camera. Frame i has the image I_i and the world-to-camera pose $T_i = (R_i, t_i)$. The scene consists of points $X_j \in \mathbb{R}^3$. A central camera model maps a scene point to the observation

$$u_{ij} = \pi_\theta(R_i X_j + t_i) + \epsilon_{ij}, \quad (1)$$

where u_{ij} denotes the pixel of point j in image i , π_θ denotes the projection with intrinsic parameters θ , and ϵ_{ij} denotes the measurement noise.

The transect tape provides anchors. An anchor a is a tape location with a printed reading s_a and pixel observations in several frames. For an interval $e = (a, b)$ between two anchors, the material length and the endpoint chord are

$$d_e = |s_b - s_a|, \quad c_e = \|G_b - G_a\|, \quad (2)$$

where G_a and G_b denote the three-dimensional anchor positions (see Fig. 2(a)). An inextensible tape gives $c_e \leq L_e = d_e$, where L_e denotes the length of the tape curve between the two anchors.

Consider a reconstruction whose shape is correct up to the scale factor α_0 . An endpoint constraint chooses the scale factor α that satisfies $\alpha \hat{c}_e = d_e$ for the reconstructed chord $\hat{c}_e$. It

TABLE I
NOTATION

Symbol	Meaning
$I_i, T_i = (R_i, t_i)$	Image and world-to-camera pose of frame i
X_j, u_{ij}	Scene point and its pixel observation in image i
π_θ	Projection with intrinsic parameters θ
s_a, G_a	Printed reading and position of anchor a
$e = (a, b)$	Tape interval between adjacent anchors a and b
d_e, c_e, L_e	Material length, endpoint chord and curve arclength
$G_{e\ell}, \gamma_k$	Chart node ℓ and Catmull-Rom segment k
σ_e, r_e	Length uncertainty and normalized arclength residual
v_e, a^*	Scale vote of a chart and consensus scale
S_g, m_g	Segment of the scene-only reconstruction and its vote
w_e, λ	Reliability switch and switch prior weight
τ, η	Lag and relative strain scale of the trajectory prior
$\mathcal{F}, \mathcal{C}, \mathcal{I}$	Fit anchors, check anchors and consensus inliers
ε_k	Relative length discrepancy of withheld interval k

therefore gives $\alpha/\alpha_0 = d_e/c_e$. For a circular arc with the bend angle ϕ and the radius ϱ in Fig. 2(b), this ratio is

$$\frac{\alpha}{\alpha_0} = \frac{\varrho\phi}{2\varrho\sin(\phi/2)} = \frac{\phi}{2\sin(\phi/2)}. \quad (3)$$

A bend of 90° gives a ratio of 1.111. An endpoint constraint therefore enlarges the map whenever the tape bends. The inequality $c_e \leq d_e$ avoids this bias but admits every smaller scale. We instead recover the tape curve between the anchors and constrain its integrated length.

We split the anchors into a fit set $\mathcal{F}$ and a check set $\mathcal{C}$ with disjoint membership. Scale estimation and optimization use only the readings and pixels of fit anchors. Images that show check anchors contribute ordinary scene features. For each check interval k we compute the relative discrepancy

$$\varepsilon_k = 100 \frac{\hat{\ell}_k - d_k}{d_k}, \quad (4)$$

where $\hat{\ell}_k$ denotes the reconstructed length of interval k , and d_k denotes its material length. We report the Mean Absolute Error (MAE) and the Root-Mean-Square Error (RMSE) of ε_k .

B. Preliminaries

We use the notation listed in Table I. The scene is static during the pass. The tape is inextensible and stays in place. Each accepted anchor shows the same material location of the tape in all of its views. A reading identifies a position along the tape but carries no georeference. The global rotation and translation of the map therefore remain a free coordinate choice.

IV. RIFT-MAP METHOD

An endpoint constraint biases the scale on bent intervals. A tape curve without image support can also take a wrong shape that biases the scale. A join through few shared features can further change the scale between two parts of the reconstruction. Therefore, we propose the Robust Integrated Flexible-Tape Mapping (RIFT-Map) method to constrain the integrated tape length only where the images support the tape curve. Fig. 3 shows its stages on the coral transect. A scene-only reconstruction first supplies cameras and points at an

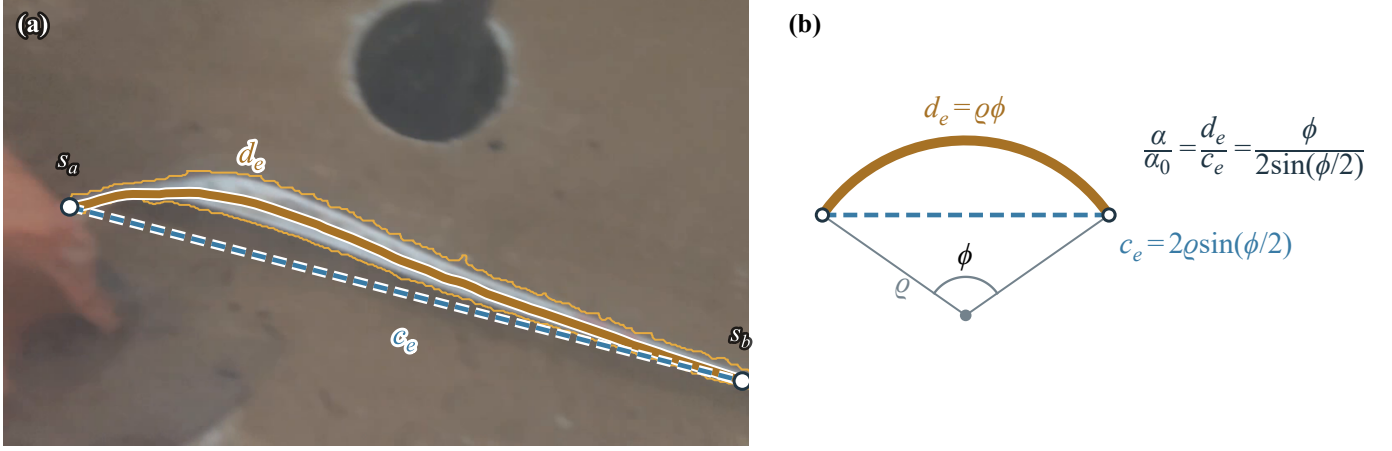


Fig. 2. Endpoint bias of a bent tape interval. (a) Pool frame in which the tape bends over a brick. The material length d_e follows the tape between two printed readings. The chord c_e joins the two readings. (b) Circular arc with the bend angle ϕ and the radius ρ that underlies (3).

arbitrary scale. The chart graph of Section IV-A links adjacent fit anchors by tape intervals and recovers their interior nodes. The metric backend of Section IV-B rescales the segments of the scene-only reconstruction, estimates the global scale from curve lengths and refines the reconstruction with arclength factors. The trajectory prior of Section IV-C limits local strain before the map is exported. Algorithm 1 summarizes the procedure.

The scene-only stage removes the tape from feature matching. The repeated tick pattern of the tape would otherwise create ambiguous matches. Semantic tape masks come from Grounding DINO detections [35] refined by SAM 2 [36]. We extract scale-invariant feature transform features outside the masks and match them sequentially with COLMAP [32]. A long sequence can break into several models. We join two models when they share at least two cameras. A join through a single shared camera also requires shared triangulated points. Their rays may disagree by at most 2° . Their relative scales may spread by at most 10%. After all joins we triangulate the union again from the full feature database. Each join fixes the relative scale of two segments from a few shared cameras or points.

A. Tape Chart Graph with Multiview Support

The graph builder triangulates each fit anchor from its observations. Anchors with a nonpositive depth or a large reprojection error are discarded. The remaining anchors are sorted by their readings. Only adjacent anchors propose an interval. The material length d_e of a proposed interval must lie between $g_{\min} = 0.05$ m and $g_{\max} = 0.65$ m.

Each interval carries a chart of five nodes $G_e = \{G_{e0}, \dots, G_{e4}\}$. The two end nodes are the triangulated anchors. The three interior nodes start at the fractions 0.25, 0.5 and 0.75 of the endpoint chord. These fractions initialize a spatial search. For every interior node we project the initial position into the frames that see the chart. A search in the semantic tape mask moves each projection by up to 70 pixels across the tape and by at most 2 pixels along it. The search keeps the candidate with the highest score of mask support

minus a penalty on its displacement. We triangulate the node from the snapped pixels. A node is accepted when it keeps at least three views, a reprojection error of at most 8 pixels and a triangulation angle of at least 0.25° . An interval enters the graph only when all three interior nodes are accepted. The graph can therefore contain gaps and several components.

B. Metric Backend

We interpolate the five nodes with a uniform Catmull-Rom spline [37] with duplicated end control points. Within segment k the four nearest nodes $P_0, \dots, P_3$ define

$$\gamma_k(t) = \frac{1}{2} [2P_1 + (P_2 - P_0)t + (2P_0 - 5P_1 + 4P_2 - P_3)t^2 + (3P_1 - P_0 - 3P_2 + P_3)t^3], \quad (5)$$

where $t \in [0, 1]$ denotes the segment parameter. Five-point Gauss-Legendre quadrature gives the chart length

$$L_e = \sum_{k=0}^3 \sum_{q=1}^5 \omega_q \|\gamma'_k(t_q)\|, \quad (6)$$

where t_q and ω_q denote the quadrature nodes and weights, respectively. Because γ'_k is linear in the nodes, the gradient of the chart length is

$$\frac{\partial L_e}{\partial G_{e\ell}} = \sum_{k=0}^3 \sum_{q=1}^5 \omega_q b'_{k\ell}(t_q) \frac{\gamma'_k(t_q)}{\|\gamma'_k(t_q)\|}, \quad (7)$$

where $b'_{k\ell}(t)$ denotes the coefficient of node ℓ in $\gamma'_k(t)$. Charts of zero length are rejected before optimization.

Each admitted chart votes for the scale with $v_e = d_e / \hat{L}_e$, where $\hat{L}_e$ denotes its length in the scene-only reconstruction. A join through one camera can leave a segment with a scale that differs by several percent from that of its neighbor. Join rescaling therefore estimates a vote m_g for every segment S_g from its own charts. It takes the largest set of votes in the segment whose largest vote exceeds its smallest by at most 10%. The median of this set gives m_g when the set holds at least two votes. Otherwise the segment takes the median of all votes. Every segment after a join is scaled by m_g / m_{g-1}

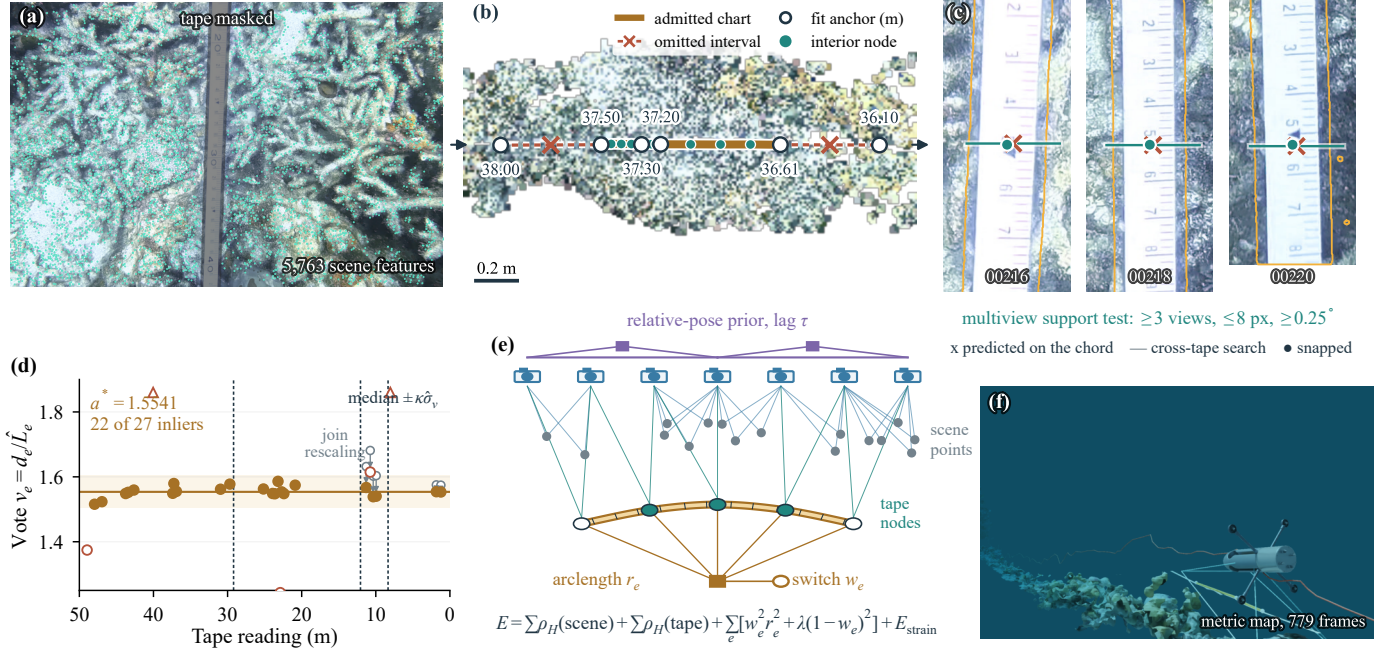


Fig. 3. Stages of RIFT-Map shown on the coral transect. (a) Scene-only structure from motion extracts features outside the semantic tape mask. (b) Adjacent fit anchors propose intervals. Each interval carries a chart of five nodes. (c) Each interior node is predicted on the chord, projected into the frames that see it and moved across the tape to the mask center. It is triangulated from the snapped pixels and passes the multiview support test with at least three views. (d) Each admitted chart votes for the scale with its material length over its curve length. Hollow grey markers show the votes before join rescaling. Dashed lines mark the joins of the scene-only stage. The median and the robust spread of the rescaled votes select the consensus. (e) Bundle adjustment couples reprojection terms of scene points and tape nodes with one switchable arclength factor per chart and relative-pose priors between frames that lie τ frames apart. (f) The output is a metric map of the transect.

about the center of the join camera. All later segments follow the same similarity transform. Each scene point follows the segment of its median observing frame. We then compute the votes again in the rescaled reconstruction.

We compute the median $\tilde{v}$ of the votes and their robust spread $\hat{\sigma}_v = 1.4826 \text{MAD}(v)$, where MAD denotes the Median Absolute Deviation [38]. Votes with $|v_e - \tilde{v}| > \kappa \hat{\sigma}_v$ are rejected with $\kappa = 2.5$. At least $N_{\min}$ inliers must remain, where $N_{\min}$ is two thirds of the admitted charts rounded up. The inlier set $\mathcal{I}$ gives the weighted least-squares scale

$$a^* = \frac{\sum_{e \in \mathcal{I}} \hat{L}_e d_e / \sigma_e^2}{\sum_{e \in \mathcal{I}} \hat{L}_e^2 / \sigma_e^2}, \quad (8)$$

where $\sigma_e = \max(0.001 \text{ m}, 0.015 d_e)$ denotes the length uncertainty of interval e . We multiply all camera centers and points by a^* before the joint optimization.

The anchors and interior nodes of the inlier charts enter bundle adjustment as tape points with their own reprojection terms. The joint objective is

$$E = \sum_{(i,j)} \rho_H \left(\|\pi_\theta(T_i X_j) - u_{ij}\|^2 \right) + \sum_{(i,\ell)} \rho_H \left(\|\pi_\theta(T_i G_\ell) - u_{i\ell}^{\text{tp}}\|^2 \right) + \sum_{e \in \mathcal{I}} [w_e^2 r_e^2 + \lambda(1 - w_e)^2], \quad (9)$$

where ρ_H denotes the Huber loss with a 1-pixel transition, $u_{i\ell}^{\text{tp}}$ denotes the snapped observation of tape node ℓ , $r_e =$

$(L_e - d_e) / \sigma_e$ denotes the normalized arclength residual, and $w_e \in [0, 1]$ denotes the reliability switch of interval e . For fixed geometry the optimal switch is

$$w_e = \frac{\lambda}{\lambda + r_e^2}. \quad (10)$$

An interval with a large normalized residual therefore loses influence. We set $\lambda = 16$ so that an interval at four standard deviations keeps half of its weight. The optimization refines all poses, scene points, tape nodes, the focal length and the radial distortion. The principal point stays fixed. The first pose fixes the rigid gauge. The arclength factors fix the scale.

C. Trajectory Prior

Imperfect tape evidence can bend the camera path locally toward a wrong interval. We therefore add relative-pose priors between frames that lie τ frames apart. The priors keep the relative motion of the scaled initial reconstruction. Their translation uncertainty is

$$\sigma_{t,ij} = \max(\sigma_{\min}, \eta \|\bar{t}_{ji}\|), \quad (11)$$

where $\bar{t}_{ji}$ denotes the initial relative translation, η denotes the relative strain scale, and $\sigma_{\min} = 0.001 \text{ m}$. Both τ and η come from the fit evidence. The lag τ is the median number of frames that see a consensus chart. The scale η is the robust spread of the scale votes divided by their median. The prior adds

$$E_{\text{strain}} = \sum_{(i,j)} \|r_{ij}^{\text{rel}}\|_{\Sigma_{ij}^{-1}}^2 \quad (12)$$

Algorithm 1 RIFT-Map

Input: Images $\{I_i\}$, tape masks, fit anchors $\mathcal{F}$ with readings s_a and pixels

Output: Metric poses $\{T_i\}$, intrinsics θ , scene points $\{X_j\}$

- 1: Reconstruct the scene from features outside the tape masks and join the segments
 - 2: Triangulate the anchors of $\mathcal{F}$
 - 3: **for** each pair of adjacent fit anchors with $g_{\min} \leq d_e \leq g_{\max}$ **do**
 - 4: Initialize three interior nodes on the chord and snap their projections to the tape mask
 - 5: Triangulate the nodes and keep the interval only if every node passes the support test
 - 6: **end for**
 - 7: Compute the chart lengths (6) and the votes $v_e = d_e / \hat{L}_e$
 - 8: Scale every segment after a join by m_g / m_{g-1} and compute the votes again
 - 9: Reject outlying votes and scale the reconstruction by a^* from (8)
 - 10: Minimize (9) plus (12) over poses, points, tape nodes, intrinsics and switches
 - 11: Freeze the cameras and triangulate the scene matches again
-

to (9), where r_{ij}^{rel} denotes the relative-pose residual with the covariance Σ_{ij} . The rotational part of the prior has a standard deviation of 180° , which leaves the relative rotation free.

For tractable optimization the joint stage uses a working subset of scene points that covers every camera. After convergence we freeze the poses and intrinsics. We discard the tape points and the working subset. We then triangulate all scene matches again at the frozen cameras.

V. EXPERIMENTS

A. Test Scenarios and Setup

We evaluate RIFT-Map in two test scenarios that Table II lists. Validation test scenario A is a pool in which a tape lies on the floor and is recorded at close range. It validates the processing chain from the tape masks to the admitted charts. Field test scenario B is a coral transect near Koh Tao in Thailand. It evaluates the metric accuracy and the registration of the full method.

Two smartphones in transparent waterproof pouches recorded a measuring tape on the floor of a pool. Recording A1 covers 0.5 to 47.5 s of the first video. Recording A2 covers 44 to 101 s of the second video. We sample both at 6 frames per second. One annotator marked the tick of every visible 10-cm reading in several frames as an anchor.

The coral transect was recorded near Koh Tao in Thailand on 5 September 2024 with the three-camera system of Wang et al. [19]. The survey names this transect Line2. Three GoPro HERO12 cameras in housings with red filters faced the reef. We use only the middle camera, which keeps the tape near the image center. The uncorrected sequence of this camera contains 779 frames of 2688 by 1512 pixels at about 0.9 frames per second. Seventy-nine numeric anchors with 324 observations cover readings from 0.91 to 49.00 m. The pass runs toward decreasing readings. Nine check intervals with material lengths of 0.1 to 0.2 m are withheld from every fit. Their midpoints lie 0.7 to 22.9 m from the start of the pass. After removal of the check anchors, 59 fit anchors pass the triangulation test. The median range from a camera to an anchor it observes is 0.43 m.

We evaluate the coral transect by a leave-one-interval-out validation along the whole pass. Every admitted chart with a material length of at least 0.1 m is withheld once. Twenty-two charts meet this length. Each method is fitted on the remaining charts. We then triangulate the nodes of the withheld chart at the frozen cameras of the fitted map from the same snapped pixels. Its arclength gives the reconstructed length $\hat{\ell}_k$ in (4). The nine check intervals are scored in the fit on all charts. For each of them the chord between the two triangulated check anchors gives $\hat{\ell}_k$. The 31 withheld intervals have midpoints at tape readings from 1.2 to 48.3 m. We obtain 95% intervals of the MAE and the RMSE from 20,000 bootstrap resamples of the intervals. A one-sided Wilcoxon signed-rank test compares the absolute discrepancies of RIFT-Map and another method on the same intervals.

We compare RIFT-Map with three alternatives that use the same admitted charts and the same consensus settings. Endpoint consensus rescales the scene-only reconstruction by a robust consensus of chord votes. Curve consensus uses the curve votes instead. Endpoint bundle adjustment adds factors that equate each chord with its material length. We also evaluate the established workflow of Wang et al. [19]. It aligned all three cameras in Metashape with 40 manually registered tape markers. Its metric scale comes from one scale bar between the markers at 41 and 42 m. We evaluate this model with its own scale and with the curve-consensus scale of the same charts. Both evaluations triangulate the nodes from rays of the middle camera. Five ablations isolate the parts of RIFT-Map. They remove join rescaling or bundle adjustment, fix the switches at one or remove the trajectory prior. The fifth ablation drops the support test from the variant without prior. It adds chord factors for the omitted intervals and takes the consensus from the chord votes of all 36 intervals. We also ran Metashape 2.3.2 twice on the 779 middle-camera frames with our tape masks to test registration. One run uses default matching. The other run uses the 3,893 image pairs verified by our scene-only stage.

We implemented RIFT-Map with COLMAP 4.2.0, PyCOLMAP and the Ceres solver. For the coral transect we use a shared camera model with a focal length, a principal point and two radial coefficients. For the pool recordings we use a shared camera model with one radial coefficient. In the pool the scene-only stage excludes a band of 40 pixels around the tape mask because the floor offers little texture beside the tape. The joint stage allows 500 iterations. Apart from the seeds of the repeated simulation, all random seeds are fixed to 1337. The experiments ran on an Intel Core i7-14650HX processor with 31 GB of memory and an NVIDIA RTX 4070 Ti SUPER graphics card with 16 GB of memory.

B. Validation Test Scenario A

Fig. 4 shows the processing chain on six pool frames. Grounding DINO and SAM 2 found the tape in 278 of the 282 frames of A1 and in 326 of the 342 frames of A2. The color of the tape gave the mask of the remaining frames. The scene-only stage registered 208 frames of A1 in three models and 334 frames of A2 in three models. Their mean

TABLE II
TEST SCENARIOS

Scenario	Recording	Camera	Frames	Tape anchors	Evaluation
A, pool	A1	Smartphone, 540 by 960	282	42 with 281 observations	Processing chain
	A2	Smartphone, 544 by 960	342	38 with 533 observations	Processing chain
B, coral reef	Line2	GoPro HERO12, 2688 by 1512	779	79 with 324 observations	31 withheld intervals

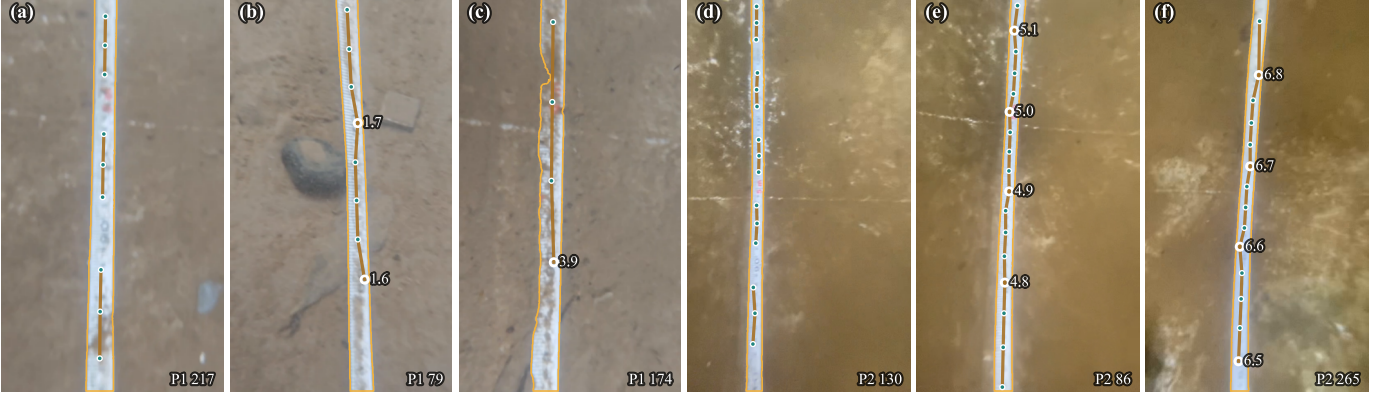


Fig. 4. Processing chain of RIFT-Map in validation test scenario A. (a) to (c) Frames of recording A1. (d) to (f) Frames of recording A2. Yellow outlines mark the semantic tape mask. White circles mark the marked readings in meters. Orange lines join the nodes of each admitted chart and teal dots mark its interior nodes snapped to the tape.

reprojection errors were 0.74 to 0.95 pixels. Within these models 17 anchors of A1 and 27 anchors of A2 passed the triangulation test. The chart graph proposed 12 intervals in A1 and 21 intervals in A2 between adjacent anchors. The multiview support test admitted 11 and 20 of them. The two omitted intervals each had an interior node with fewer than three consistent views. The interior nodes of the admitted charts kept 5 to 102 consistent views. The snapped nodes followed the center of the tape in every frame of Fig. 4. The tape lay straight on the floor in the labeled sections. Its arclength between two readings therefore equals its chord. Thirty of the 31 admitted charts had an arclength within 0.53% of their chord. Their median excess was 0.07%. One admitted chart at the start of A2 had an arclength 23.6% above its chord because its interior nodes folded along the viewing direction. The scene-only reconstructions of the pool drifted in scale along the pass. In the largest model of A2 the reconstructed length of a 10-cm interval near its end was 1.91 times that of an interval near its start. The pool therefore validates the processing chain from the masks to the admitted charts. The metric evaluation uses the coral transect.

C. Field Test Scenario B

Fig. 5 shows six frames along the transect together with the chart graph. The tape crosses branching coral, rubble, sand and a crevice in these frames. The chart graph used 59 fit anchors. It proposed 36 intervals with material lengths of 0.06 to 0.62 m (see Fig. 5(g)). Twenty-seven intervals passed the multiview support test. The other nine were omitted because an interior node kept fewer than three consistent views. Fig. 3(c) shows one interior node that passed the support test in five views. Most admitted charts were nearly straight. Twenty-five of them

had an arclength within 1.9% of their chord with a median excess of 0.18%. The remaining two charts had arclength-to-chord ratios of 1.96 and 2.05. No interval was admitted between the readings of 20.7 and 11.6 m. This stretch includes the steep part of the reef.

The scene-only stage joined four segments at frames 346, 614 and 666. The join at frame 346 rested on two shared cameras. The joins at frames 614 and 666 rested on one shared camera with one and 16 shared points, respectively. The four charts of the segment between frames 615 and 666 voted from 1.602 to 1.681. Their median of 1.6177 exceeded the medians of 1.5537 and 1.5535 of the first two segments by 4.1%. Join rescaling therefore enlarged this segment by 4.1% (see Fig. 3(d)). It enlarged the last segment by 1.4%. The consensus then kept 22 of the 27 scale votes. The median vote was 1.5537. The consensus scale was $a^* = 1.5541$.

The working subset held 80,000 scene points with 179,284 observations. The joint stage also used 22 arclength factors and 102 tape points with 493 observations. The trajectory prior had a lag of 10 frames, a strain scale of 1.29% and 78 relative-pose factors. The optimization converged after 49 successful steps without an unsuccessful step. The mean reprojection error of the working subset rose from 0.201 to 0.209 pixels. All 22 switches stayed above 0.87. After the second triangulation the map held 621,724 points with 2,626,353 observations. Its mean reprojection error was 0.679 pixels. Fig. 6 unrolls this map along the camera path. The plan view in Fig. 6(a) shows a continuous reef strip along the whole pass. The side profile in Fig. 6(b) shows the relief of the reef under the camera path at true scale. Its vertical axis points against the mean viewing direction of the camera.

All 779 frames of the coral transect lie in one map

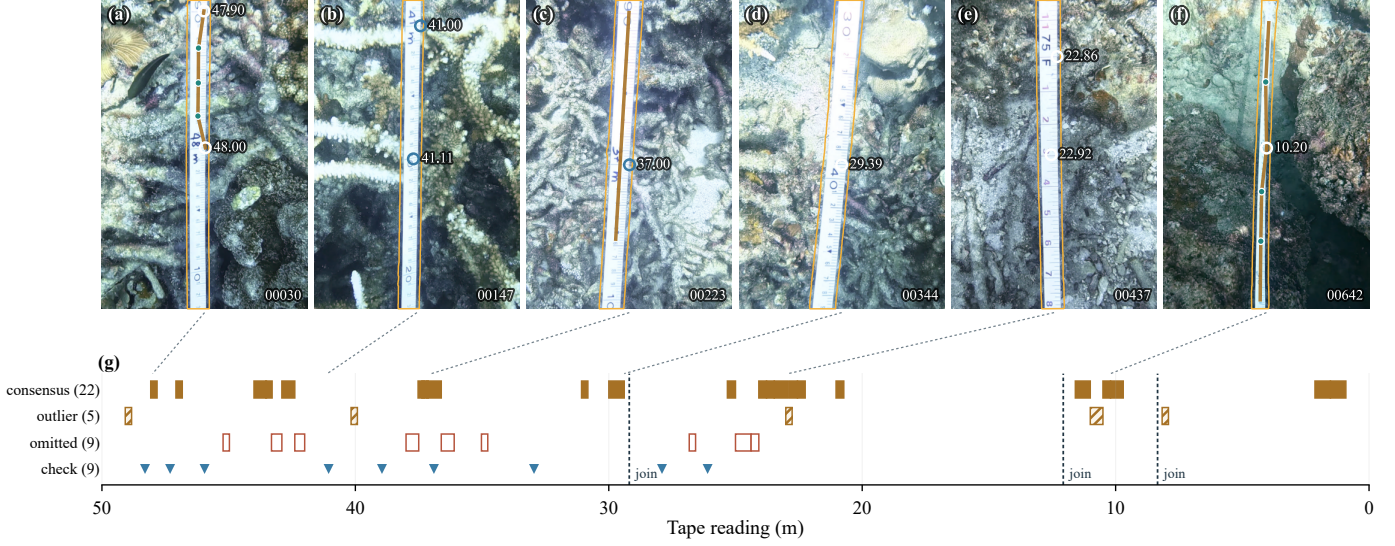


Fig. 5. Frames and chart graph of field test scenario B. (a) to (f) Frames of the middle camera at tape readings of about 48, 41, 37, 29, 23 and 10 m. Yellow outlines mark the semantic tape mask. White circles mark fit anchors and blue circles mark check anchors with their readings, in meters. Orange lines join the nodes of consensus charts and teal dots mark their interior nodes. The frames are shown after color correction. (g) Proposed intervals along the tape reading. Filled bars entered the consensus. Hatched bars passed the support test with outlying votes. Open bars were omitted by the support test. Triangles mark the midpoints of the check intervals. Dashed lines mark the joins of the scene-only stage.

TABLE III
REGISTRATION OF THE MIDDLE-CAMERA FRAMES IN FIELD TEST SCENARIO B

Method	Cameras	Registered	Components	Largest
Metashape workflow [19]	Three	779	1	779
Metashape, default matching	Middle	757	4	346
Metashape, verified pairs	Middle	755	3	589
RIFT-Map	Middle	779	1	779

(Table III). This connectivity comes from the joins of the scene-only stage. The three-camera Metashape model of Wang et al. [19] also held all 779 middle-camera frames in one component. Metashape on the middle camera alone with default matching registered 757 frames in four components (see Fig. 6(c)). Its largest component held 346 frames. With the verified pair list Metashape registered 755 frames in three components. Its largest component held 589 frames. Both middle-camera solutions left frames 592 to 613 unregistered. These frames lie where the camera climbed a steep part of the reef about 35 m along the path (see Fig. 6(b) and (c)).

We imported the metric map into the Isaac Sim robot simulator (see Fig. 7). Underwater simulators built on Isaac Sim render camera and sonar data for perception research [39]. We meshed the scene points into a colored surface by screened Poisson reconstruction [40]. The surface covers 61.7 m² of reef. The median distance of the 621,724 map points from this surface is 2.8 mm. Of these points 88.5% lie within 10 mm of the surface. Within 20 mm of it lie 96.1% of the points. The twin also holds the tape through the triangulated anchors and chart nodes and the camera path at metric scale. A model of a hybrid aerial-underwater vehicle hovers 0.36 m above the tape at the pose of survey frame 214. Fig. 7(d) shows the steep part of the reef where the middle-camera Metashape runs left frames 592 to 613 unregistered. The surface breaks into strips on this steep face.

Table IV lists the discrepancies of the 31 withheld intervals. RIFT-Map had an MAE of 1.37% and an RMSE of 2.04%. These were the lowest values of the six compared methods. Endpoint bundle adjustment came closest with an MAE of 1.47% and an RMSE of 2.13%. The Metashape workflow had an MAE of 3.51% and an RMSE of 4.78% with its scale bar. With the curve-consensus scale of the same charts it had 3.23% and 3.86%. RIFT-Map had the smaller absolute discrepancy than the Metashape workflow on 25 of the 31 intervals. Its advantage was significant at the 5% level over both Metashape evaluations and both consensus methods. It was not significant over endpoint bundle adjustment. Fig. 8(a) shows the discrepancy of every method on every interval. The scale bar of the Metashape workflow underestimated all eight intervals above the reading of 41 m by 7.3% to 10.5%. Between the readings of 39 and 21 m its discrepancies stayed within 3.1% in magnitude. After the steep part it underestimated the six intervals by 2.8% to 7.5%. All four middle-camera methods had their largest discrepancy on the 0.11-m check interval near the reading of 41.1 m with values from -7.5% to -8.0%. Without this interval RIFT-Map had an RMSE of 1.47% over the other 30 intervals, against 1.61% for endpoint bundle adjustment.

Join rescaling changed the discrepancies after the join at frame 614. Without it RIFT-Map underestimated the four charts of this segment by 1.9% to 6.6%. Join rescaling lowered their RMSE from 4.13% to 2.66%. Over all 31 intervals it lowered the MAE of RIFT-Map from 1.59% to 1.37% and the RMSE from 2.31% to 2.04%. Bundle adjustment after join rescaling lowered the MAE from 1.43% to 1.37%. The trajectory prior lowered the RMSE from 2.12% to 2.04%. The MAE was 1.37% with and without it. Fixing the switches at one changed no discrepancy by more than 0.1 percentage points. Endpoint fallback without the prior had the lowest

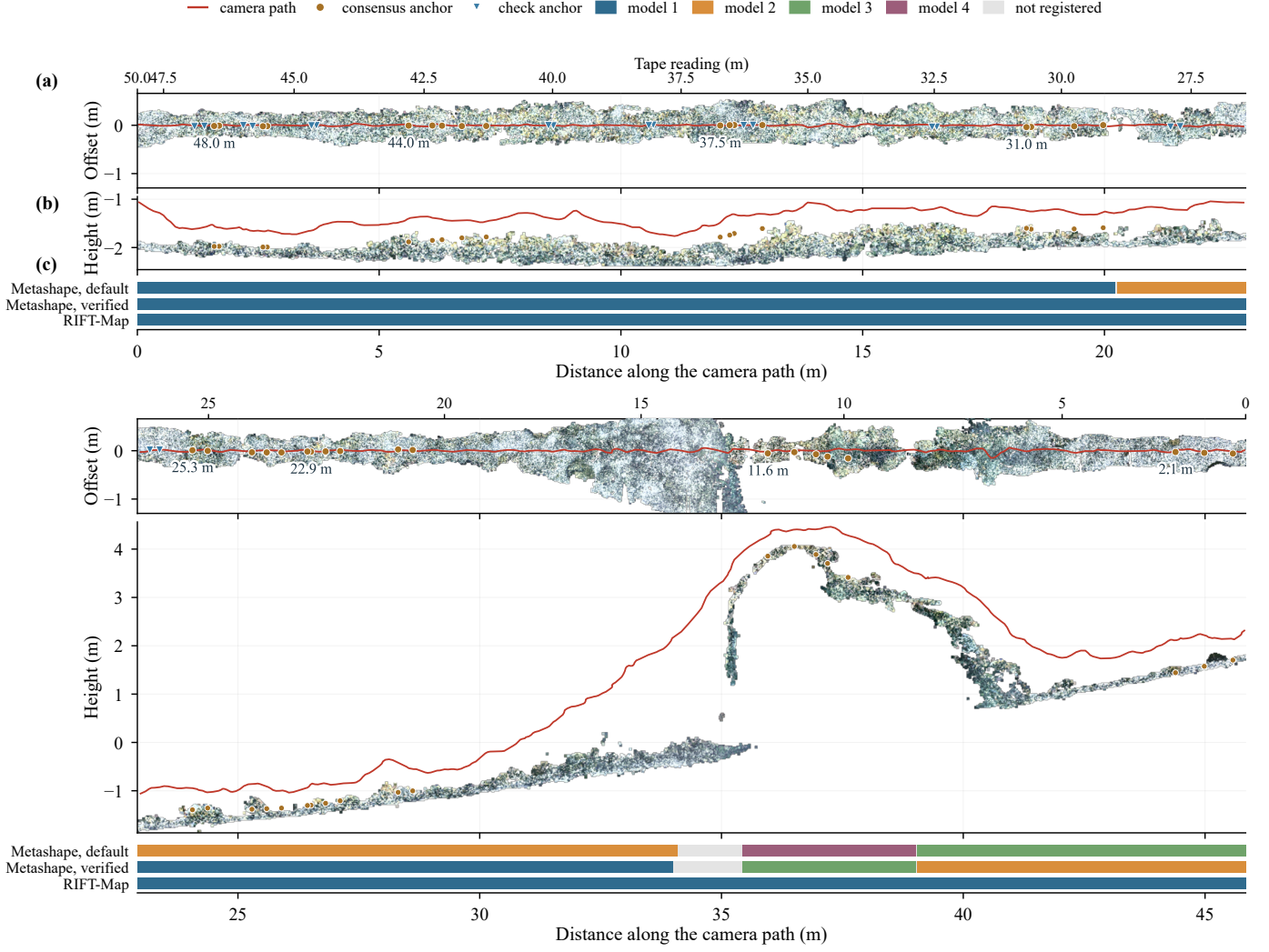


Fig. 6. Metric map of field test scenario B produced by RIFT-Map, unrolled along the camera path. The upper and lower blocks show the first and the second half of the pass. (a) Plan view with the horizontal offset from the camera path. Points take the colors of the color-corrected frames. The red line is the camera path. Orange circles mark the 36 anchors of the 22 consensus charts with selected readings. Blue triangles mark the 18 anchors of the check intervals. The upper axis gives the tape reading. (b) Side profile at true scale. (c) Registration of the 779 middle-camera frames by Metashape with default matching, by Metashape with the verified pairs and by RIFT-Map. Each color marks one connected model.

TABLE IV
DISCREPANCY OF THE 31 WITHHELD INTERVALS IN FIELD TEST SCENARIO B

Method	Metric reference	MAE (%)	RMSE (%)	RMSE interval (%)	Max $ \varepsilon_k $ (%)	p	Smaller
Three-camera Metashape [19]	Scale bar	3.51	4.78	3.53 to 5.86	10.55	0.0002	25 of 31
Three-camera Metashape [19]	Curve consensus	3.23	3.86	3.13 to 4.53	7.98	0.0002	25 of 31
Endpoint consensus	Chord consensus	1.76	2.46	1.50 to 3.33	7.45	0.012	20 of 31
Curve consensus	Arclength consensus	1.68	2.46	1.43 to 3.37	7.66	0.036	21 of 31
Endpoint bundle adjustment	Chord factors	1.47	2.13	1.22 to 3.06	7.95	0.35	16 of 31
RIFT-Map	Arclength factors	1.37	2.04	1.12 to 2.96	8.03		
Ablations of RIFT-Map							
Without join rescaling	Arclength factors	1.59	2.31	1.30 to 3.24	7.64	0.10	20 of 31
Without bundle adjustment	Arclength consensus	1.43	2.08	1.18 to 3.04	8.09	0.19	18 of 31
Switches fixed at one	Arclength factors	1.37	2.04	1.13 to 2.99	8.06	0.31	19 of 31
Without trajectory prior	Arclength factors	1.37	2.12	1.12 to 3.14	8.45	0.53	15 of 31
Endpoint fallback without prior	Mixed factors	1.23	1.95	0.99 to 2.93	7.97	0.91	15 of 31

values of all variants with an MAE of 1.23% and an RMSE of 1.95%.

The processing of the coral transect ran on the computer described above. The semantic masks of the 779 frames took 8.6 min. Feature extraction, sequential matching and guided

matching took 13.2, 13.8 and 15.7 min with eight processor threads. The incremental mapping of the largest segment took 5.9 min. The chart graph took 12 s. Join rescaling, consensus and bundle adjustment took 23 s together. The final triangulation took 46 s. One annotator marked 324 pixels of

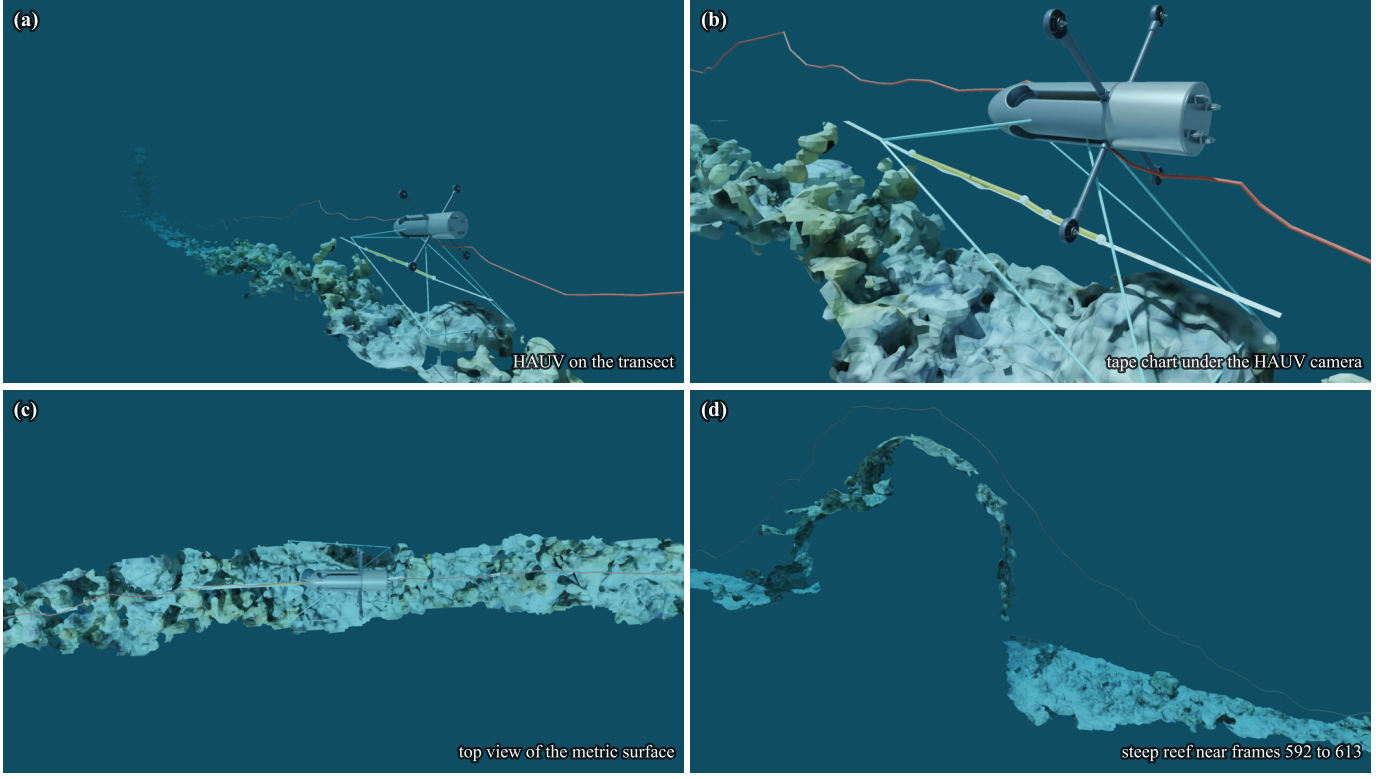


Fig. 7. Metric digital twin of field test scenario B in Isaac Sim. (a) Overview of the transect with the vehicle model above the tape. (b) Close view of a consensus chart under the camera of the vehicle. The thin lines outline the camera footprint. (c) Top view of the reef surface with the camera path. (d) The steep part of the reef near frames 592 to 613. The renderings add a depth-dependent underwater attenuation.

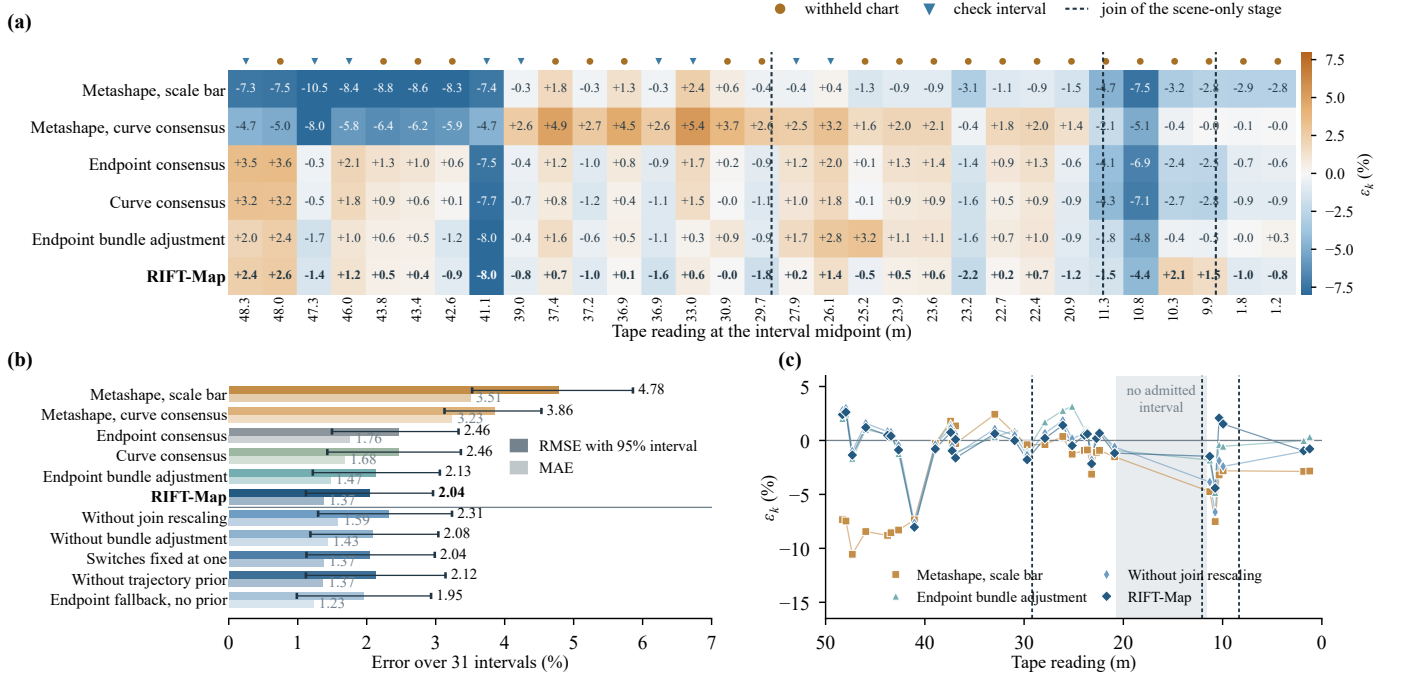


Fig. 8. Transect-wide validation in field test scenario B. (a) Signed discrepancy of each method on the 31 withheld intervals, ordered along the tape. Circles mark withheld charts and triangles mark check intervals. Dashed lines mark the joins of the scene-only stage. (b) RMSE with its 95% bootstrap interval and MAE of the compared methods and of the ablations of RIFT-Map. (c) Discrepancy along the tape. The shaded stretch holds no admitted interval.

79 anchors for RIFT-Map. The Metashape project of Wang et al. held 40 registered tape markers.

D. Controlled Curvature Analysis

The two real scenarios cannot vary the bend of the tape at a fixed scene. We therefore simulate 50 cameras and 250 scene points above a 3.9-m tape. Forty anchors lie at a spacing of 0.1 m along the tape. We generate a straight tape, a curved tape and a mixed tape. The curved tape follows a sinusoid with an amplitude of 0.08 m and a period of 0.4 m. The mixed tape stays straight for 1.6 m and then follows a raised cosine with the same amplitude and period. Fig. 9(a) shows the three profiles. Anchor pixels carry Gaussian noise with a standard deviation of 0.12 pixels. Curve samples carry noise with a standard deviation of 0.06 pixels. At the simulated range of 3 m and the focal length of 1,100 pixels the anchor noise corresponds to 0.33 mm. On the coral transect a tick marked 2 pixels off at the median range of 0.43 m moves by 0.35 mm. We split twenty intervals into ten fit intervals and ten check intervals that alternate along the tape. The five measured curve samples of each interval form its chart. Three arms share one scene reconstruction, one set of tape observations and one curve-consensus scale. The consensus uses $\kappa = 3$ and $N_{\min} = 5$. The consensus arm stops after scaling. The endpoint arm adds factors that equate each endpoint chord with its material length. The arclength arm adds the continuous factors of (9) without switches. Every arm then evaluates the same ten check curves at its frozen cameras. We repeat each tape with 100 random seeds. Each seed draws a new scene, new camera poses and new pixel noise. We also sweep the amplitude of the curved tape from 0.5 to 4 cm with 50 seeds per amplitude. The camera-scale error is the absolute deviation of the least-squares scale ratio between reconstructed and true camera centers from one. The camera error is the root-mean-square distance between reconstructed and true camera centers without alignment.

Table V lists the means and standard deviations over the 100 seeds. On the curved tape the endpoint arm enlarged the reconstruction by 3.16% on average. The consensus and arclength arms kept the camera-scale error at 0.22% and 0.23%, respectively. The arclength arm had a smaller camera-scale error than the endpoint arm in all 100 seeds. On the mixed tape the arclength arm had the lowest camera-scale error of 0.10%. The consensus arm reached 0.15% and the endpoint arm 0.26%. The arclength arm was more accurate than the consensus arm in 65 of the 100 seeds. On the straight tape the endpoint arm had the lowest camera-scale error of 0.07% because chord and arclength coincide there. Fig. 9(d) shows that the endpoint bias grew with the bend amplitude. It followed the mean ratio of material length to chord of the fit intervals. At an amplitude of 2 cm this ratio predicts a bias of 0.43%. The endpoint arm reached 0.44% there. The two curve arms stayed below 0.2% up to an amplitude of 4 cm. All arms recovered all ten check curves in every repetition.

VI. DISCUSSION

A. Effect of Tape Curvature and Evidence

In the pool the multiview support test admitted 31 of 33 proposed intervals. On the coral transect it admitted 27 of 36 intervals. In both scenarios it omitted an interval only where

TABLE V
CONTROLLED CURVATURE RESULTS OVER 100 SEEDS

Tape	Arm	Scale (%)	Camera (mm)	MAE (%)	RMSE (%)
Straight	Endpoint	0.07 ± 0.06	2.3 ± 1.3	0.08 ± 0.03	0.10 ± 0.03
	Consensus	0.11 ± 0.07	2.9 ± 1.5	0.09 ± 0.03	0.10 ± 0.03
	Arclength	0.11 ± 0.07	2.9 ± 1.5	0.09 ± 0.03	0.10 ± 0.03
Curved	Endpoint	3.16 ± 0.28	71.6 ± 6.4	3.04 ± 0.29	3.06 ± 0.29
	Consensus	0.22 ± 0.19	5.4 ± 4.1	0.34 ± 0.13	0.41 ± 0.14
	Arclength	0.23 ± 0.19	5.5 ± 4.1	0.34 ± 0.13	0.40 ± 0.13
Mixed	Endpoint	0.26 ± 0.11	6.2 ± 2.3	0.39 ± 0.11	0.44 ± 0.12
	Consensus	0.15 ± 0.10	3.7 ± 2.2	0.26 ± 0.08	0.33 ± 0.10
	Arclength	0.10 ± 0.08	2.8 ± 1.7	0.24 ± 0.07	0.32 ± 0.09

an interior node kept fewer than three consistent views. On the straight pool tape 30 of the 31 admitted curves kept their straight shape within 0.53%. The test checks the views of each node but not the shape of the curve. The folded chart in the pool shows such a case. Its curve length gives an outlying scale vote, which the robust consensus of Section IV-B is designed to reject. The controlled analysis isolates the geometric effect of the tape model. Over 100 repetitions an endpoint constraint biased the scale by 3.16% on a tape with a bend amplitude of 8 cm. This bias followed the ratio of material length to chord that (3) describes for a circular arc. Curve lengths avoided it at every amplitude. On the coral transect most admitted intervals were almost straight with a median arclength excess of 0.18% over the chord. Chord and curve votes therefore gave similar scales on this transect. Chord factors for the nine omitted intervals even lowered the discrepancy of the variant without prior. On a bent tape the same fallback would bring back the endpoint bias of (3).

B. Effect of Join Rescaling

The largest gain on the coral transect came from join rescaling. The join at frame 614 rested on one shared point. The segment after it carried a scale about 4% smaller than that of its neighbors. A single global scale left the charts of this segment up to 6.6% too short. Join rescaling corrected this step from the tape votes of the segment. Endpoint bundle adjustment also reduced this error because its chord factors could rescale the segment locally. Over all 31 intervals its MAE and RMSE stayed above those of RIFT-Map. The trajectory prior and the switches had small effects on this transect. All switches stayed above 0.87 because the consensus had already removed the outlying votes. The connected 779-frame map comes from the joins of the scene-only stage. Join rescaling gives these joins a metric basis from the tape.

C. Limitations

The fit charts and the check intervals of the coral transect were inspected during the development of the method. No interval was admitted between the readings of 20.7 and 11.6 m, where the camera climbed the steep part of the reef. The metric shape of this stretch rests on the scene-only reconstruction and the trajectory prior. On the coral transect the reference is the material length of the tape rather than an independently measured rigid distance. A short interval also carries the error of its two marked ticks. The interval near the reading of

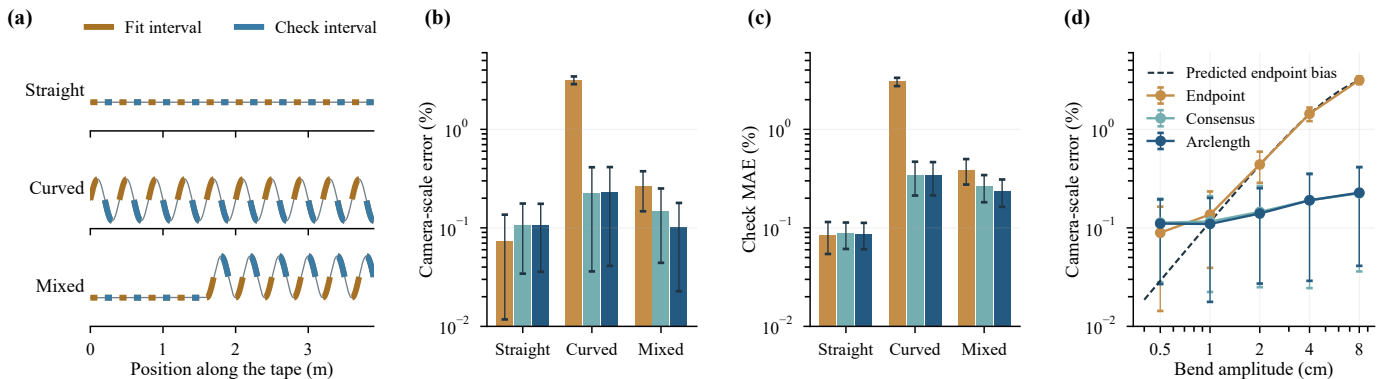


Fig. 9. Controlled curvature analysis. (a) Side profiles of the three simulated tapes with their fit and check intervals. The vertical axis is exaggerated. (b) Camera-scale error of the three arms over 100 seeds per tape. (c) MAE of the ten check curves. Bars show means and whiskers show standard deviations. (d) Camera-scale error of the curved tape against its bend amplitude. The dashed line gives the endpoint bias predicted by the ratio of material length to chord. The vertical axes are logarithmic.

41.1 m gave a discrepancy of about -8% in every middle-camera method. One transect from one survey cannot establish a general accuracy. Weakly connected sequential surveys can also bend into a dome [12]. The Metashape workflow used three cameras with its own calibration and one scale bar. This comparison therefore differs in more than one factor. The effective pinhole calibration leaves refraction at the housing port unmodeled [27], [29]. The tape constrains the geometry near its course more strongly than the geometry of the reef far from it. The readings of the tape are marked by hand.

D. Acquisition Guidance

The pool recordings show the limits of consumer video for metric monocular mapping. Their 960-pixel frames of a weakly textured floor gave the scene-only reconstruction a strong scale drift. The coral transect with 2688-pixel frames and rich reef texture kept the mean length discrepancy of RIFT-Map at 1.37%. Surveys that rely on the tape should therefore record at full resolution with a fixed focal length. Textured reef or rubble beside the tape keeps the scene-only reconstruction consistent. Steep relief can break a pass into segments that share only one camera. Readable tape intervals on both sides of such a break let join rescaling correct its scale.

In future work, rigid distances measured on the reef away from the tape would test the absolute accuracy of RIFT-Map. A refractive camera model would reduce the calibration error caused by the housing port [29]. Vehicles that follow the tape at a constant altitude would supply views of the tape with a controlled baseline. The Isaac Sim twin of the transect provides a metric scene for simulated surveys of such vehicles [39].

VII. CONCLUSION

We have proposed the Robust Integrated Flexible-Tape Mapping (RIFT-Map) method for metric monocular mapping of reef transects. RIFT-Map treats the transect tape as a curve of known material length. The method admits only intervals with multiview interior support. Join rescaling corrects the relative

scale of the scene-only segments from their tape curves. A robust consensus of curve lengths sets the scale. Switchable arclength factors then refine the map. On a straight pool tape 30 of its 31 admitted curves stayed within 0.53% of their chord. On a 50-m coral transect it built a single metric map from 779 frames of one camera. Its MAE over 31 intervals withheld along the whole pass was 1.37%, the lowest of six methods. In 100 simulated repetitions its curve-length model avoided the 3.16% scale bias of endpoint constraints on a curved tape.

In our future work, we will validate RIFT-Map against rigid distances measured on the reef. We will also add a refractive camera model for surveys by underwater vehicles.

ACKNOWLEDGMENT

The authors thank the survey team of Wang et al. for the imagery and the Metashape project of the coral transect.

REFERENCES

- [1] S. English, C. Wilkinson, and V. Baker, Eds., *Survey Manual for Tropical Marine Resources*, 2nd ed. Townsville, QLD, Australia: Australian Institute of Marine Science, 1997.
- [2] J. Hill and C. Wilkinson, *Methods for Ecological Monitoring of Coral Reefs: A Resource for Managers*. Townsville, QLD, Australia: Australian Institute of Marine Science, 2004, version 1.
- [3] J. H. R. Burns, D. Delparte, R. D. Gates, and M. Takabayashi, "Integrating structure-from-motion photogrammetry with geospatial software as a novel technique for quantifying 3D ecological characteristics of coral reefs," *PeerJ*, vol. 3, p. e1077, 2015, doi: 10.7717/peerj.1077.
- [4] J. X. Leon, C. M. Roelfsema, M. I. Saunders, and S. R. Phinn, "Measuring coral reef terrain roughness using 'Structure-from-Motion' close-range photogrammetry," *Geomorphology*, vol. 242, pp. 21–28, Aug. 2015, doi: 10.1016/j.geomorph.2015.01.030.
- [5] D. Menacho, S. Torres, L. Valle, D. Arce, and C. Carranza, "Enhanced underwater 3-D reconstruction: Case study at Los Organos reef using 3dFeatureUp and 3dVisualAppUp," *IEEE Journal of Oceanic Engineering*, vol. 51, no. 1, pp. 745–761, Jan. 2026, doi: 10.1109/JOE.2025.3617910.
- [6] C. D. Storlazzi, P. Dartnell, G. A. Hatcher, and A. E. Gibbs, "End of the chain? Rugosity and fine-scale bathymetry from existing underwater digital imagery using structure-from-motion (SfM) technology," *Coral Reefs*, vol. 35, no. 3, pp. 889–894, 2016, doi: 10.1007/s00338-016-1462-8.
- [7] W. Figueira, R. Ferrari, E. Weatherby, A. Porter, S. Hawes, and M. Byrne, "Accuracy and precision of habitat structural complexity metrics derived from underwater photogrammetry," *Remote Sensing*, vol. 7, no. 12, pp. 16883–16900, 2015, doi: 10.3390/rs71215859.

- [8] P. Rossi, C. Castagnetti, A. Capra, A. J. Brooks, and F. Mancini, "Detecting change in coral reef 3D structure using underwater photogrammetry: Critical issues and performance metrics," *Applied Geomatics*, vol. 12, no. Suppl. 1, pp. 3–17, 2020, doi: 10.1007/s12518-019-00263-w.
- [9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2004, doi: 10.1017/CBO9780511811685.
- [10] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular SLAM," in *Proc. Robotics: Science and Systems (RSS)*, Zaragoza, Spain, 2010, doi: 10.15607/RSS.2010.V1.010.
- [11] M. Bryson *et al.*, "Characterization of measurement errors using structure-from-motion and photogrammetry to measure marine habitat structural complexity," *Ecology and Evolution*, vol. 7, no. 15, pp. 5669–5681, 2017, doi: 10.1002/ece3.3127.
- [12] K. Istenič, N. Gracías, A. Arnaubec, J. Escartín, and R. García, "Scale accuracy evaluation of image-based 3D reconstruction strategies using laser photogrammetry," *Remote Sensing*, vol. 11, no. 18, p. 2093, 2019, doi: 10.3390/rs11182093.
- [13] E. Nocerino *et al.*, "Coral reef monitoring by scuba divers using underwater photogrammetry and geodetic surveying," *Remote Sensing*, vol. 12, no. 18, p. 3036, 2020, doi: 10.3390/rs12183036.
- [14] A. Purser *et al.*, "Ocean floor observation and bathymetry system (OFOBS): A new towed camera/sonar system for deep-sea habitat surveys," *IEEE Journal of Oceanic Engineering*, vol. 44, no. 1, pp. 87–99, Jan. 2019, doi: 10.1109/JOE.2018.2794095.
- [15] Z. Xu, M. Haroutunian, A. J. Murphy, J. Neasham, and R. Norman, "An integrated visual odometry system for underwater vehicles," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 3, pp. 848–863, Jul. 2021, doi: 10.1109/JOE.2020.3036710.
- [16] O. Pizarro, R. M. Eustice, and H. Singh, "Large area 3-D reconstructions from underwater optical surveys," *IEEE Journal of Oceanic Engineering*, vol. 34, no. 2, pp. 150–169, Apr. 2009, doi: 10.1109/JOE.2009.2016071.
- [17] M. Johnson-Roberson, O. Pizarro, S. B. Williams, and I. Mahon, "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys," *Journal of Field Robotics*, vol. 27, no. 1, pp. 21–51, 2010, doi: 10.1002/rob.20324.
- [18] K. Mizuno *et al.*, "Development of an efficient coral-coverage estimation method using a towed optical camera array system [Speedy Sea Scanner (SSS)] and deep-learning-based segmentation: A sea trial at the Kujukushima islands," *IEEE Journal of Oceanic Engineering*, vol. 45, no. 4, pp. 1386–1395, Oct. 2020, doi: 10.1109/JOE.2019.2938717.
- [19] J. Wang *et al.*, "Multi-scale approach for coral condition assessment and *Drupella* sp. identification using a versatile low-cost camera system and point cloud semantic segmentation," *Marine Pollution Bulletin*, vol. 224, p. 119072, 2026, doi: 10.1016/j.marpolbul.2025.119072.
- [20] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2012, pp. 1879–1884, doi: 10.1109/IROS.2012.6385590.
- [21] J. Sauder, G. Banc-Prandi, A. Meibom, and D. Tuia, "Scalable semantic 3D mapping of coral reefs with deep learning," *Methods in Ecology and Evolution*, vol. 15, no. 5, pp. 916–934, 2024, doi: 10.1111/2041-210x.14307.
- [22] D. T. I. Bayley and A. O. M. Mogg, "A protocol for the large-scale analysis of reefs using Structure from Motion photogrammetry," *Methods in Ecology and Evolution*, vol. 11, no. 11, pp. 1410–1420, Oct. 2020, doi: 10.1111/2041-210x.13476.
- [23] F. Neyer, E. Nocerino, and A. Gruen, "Monitoring coral growth—The dichotomy between underwater photogrammetry and geodetic control network," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2, pp. 759–766, 2018, doi: 10.5194/isprs-archives-xlii-2-759-2018.
- [24] C. Campos, R. Elvira, J. J. Gómez Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021, doi: 10.1109/TRO.2021.3075644.
- [25] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, "AQUALOC: An underwater dataset for visual-inertial-pressure localization," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1549–1559, 2019, doi: 10.1177/0278364919883346.
- [26] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: An underwater SLAM system using sonar, visual, inertial, and depth sensor," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Macau, China, 2019, pp. 1861–1868, doi: 10.1109/IROS40897.2019.8967703.
- [27] T. Treibitz, Y. Schechner, C. Kunz, and H. Singh, "Flat refractive geometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 51–65, Jan. 2012, doi: 10.1109/TPAMI.2011.105.
- [28] O. Kahmen, R. Rofallski, and T. Luhmann, "Impact of stereo camera calibration to object accuracy in multimedia photogrammetry," *Remote Sensing*, vol. 12, no. 12, p. 2057, 2020, doi: 10.3390/rs12122057.
- [29] M. She, F. Seegräber, D. Nakath, and K. Köser, "Refractive COLMAP: Refractive structure-from-motion revisited," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2024, pp. 12 816–12 823, doi: 10.1109/IROS58592.2024.10802043.
- [30] A. Bartoli, Y. Gérard, F. Chadebecq, T. Collins, and D. Pizarro, "Shape-from-template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2099–2118, 2015, doi: 10.1109/TPAMI.2015.2392759.
- [31] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Vision Algorithms: Theory and Practice*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2000, vol. 1883, pp. 298–372, doi: 10.1007/3-540-44480-7_21.
- [32] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 4104–4113, doi: 10.1109/CVPR.2016.445.
- [33] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, Mar. 1964, doi: 10.1214/aoms/1177703732.
- [34] P. Agarwal, G. D. Tipaldi, L. Spinello, C. Stachniss, and W. Burgard, "Robust map optimization using dynamic covariance scaling," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Karlsruhe, Germany, 2013, pp. 62–69, doi: 10.1109/ICRA.2013.6630557.
- [35] S. Liu *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, ser. Lecture Notes in Computer Science, vol. 15105, 2024, pp. 38–55, doi: 10.1007/978-3-031-72970-6_3.
- [36] N. Ravi *et al.*, "SAM 2: Segment anything in images and videos," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025, arXiv: 2408.00714.
- [37] E. Catmull and R. Rom, "A class of local interpolating splines," in *Computer Aided Geometric Design*, R. E. Barnhill and R. F. Riesenfeld, Eds. New York, NY, USA: Academic Press, 1974, pp. 317–326, doi: 10.1016/B978-0-12-079050-0.50020-5.
- [38] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, Dec. 1993, doi: 10.1080/01621459.1993.10476408.
- [39] J. Song, H. Ma, O. Bagoren, A. V. Sethuraman, Y. Zhang, and K. A. Skinner, "OceanSim: A GPU-accelerated underwater robot perception simulation framework," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2025, pp. 1526–1533, doi: 10.1109/IROS60139.2025.11246878.
- [40] M. Kazhdan and H. Hoppe, "Screened Poisson surface reconstruction," *ACM Transactions on Graphics*, vol. 32, no. 3, pp. 1–13, 2013, doi: 10.1145/2487228.2487237.