Image-Based Relocalization and Alignment for Long-Term Monitoring of Dynamic Underwater Environments 〰

Beverley Gorry   Tobias Fischer   Michael Milford   Alejandro Fontan

Abstract—**Effective monitoring of underwater ecosystems is crucial for tracking environmental changes, guiding conservation efforts, and ensuring long-term ecosystem health. However, automating underwater ecosystem management with robotic platforms remains challenging due to the complexities of underwater imagery, which pose significant difficulties for traditional visual localization methods. We propose an integrated pipeline that combines Visual Place Recognition (VPR), feature matching, and image segmentation on video-derived images. This method enables robust identification of revisited areas, estimation of rigid transformations, and downstream analysis of ecosystem changes. Furthermore, we introduce the SQUIDLE+ VPR Benchmark—the first large-scale underwater VPR benchmark designed to leverage an extensive collection of unstructured data from multiple robotic platforms, spanning time intervals from days to years. The dataset encompasses diverse trajectories, arbitrary overlap and diverse seafloor types captured under varying environmental conditions, including differences in depth, lighting, and turbidity. Our code is available at: https://github.com/bev-gorry/underloc.**

I. INTRODUCTION

Underwater ecosystems like coral reefs, seagrass beds, and kelp forests support marine biodiversity, fisheries, and coastal protection. These underwater habitats are vital but face growing threats from climate change, pollution, and human activities such as overfishing and coastal expansion [1], [2]. Long-term monitoring of these habitats is essential to detect and quantify changes in the distribution and abundance of different species, helping researchers better understand the impacts of human activities and improving the long-term health and resilience of these critical ecosystems [3].

Due to the cost and limitations of diver-based monitoring, autonomous underwater vehicles offer a scalable alternative [4]. Equipped with AI and computer vision, they enable efficient, non-intrusive monitoring with advances in navigation [5], 3D mapping [6], [7], and image segmentation [8] for high-resolution data collection.

Reliable multi-year change detection enables long term monitoring of these dynamic and fragile ecosystems, requiring highly accurate registration to capture centimetre-level variations in habitats such as coral reefs [9]. In terrestrial and aerial monitoring scenarios, change detection benefits from GPS location priors, well-defined landmarks, and many
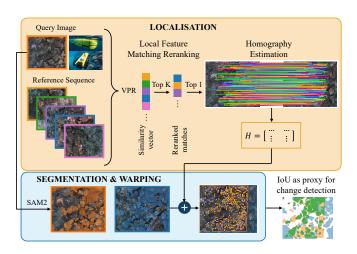
Fig. 1. **Overview of our image-based relocalization and alignment method for underwater ecosystem monitoring.** We use hierarchical Visual Place Recognition (VPR) techniques to robustly identify common locations from images captured across multi-year timescales using video from freely navigating robots. These images exhibit significant variations in environmental conditions, such as lighting, turbidity, and depth. We establish correspondences between image keypoints to estimate a rigid transformation between the images, which we then use to register segmentation masks in a common pixel space. Finally, we apply an intersection over union (IoU) metric to detect ecosystem changes over time.

decades of robotic vision techniques mostly tailored for these scenarios. However, these methods do not directly translate to underwater surveys, which must contend with unique challenges such as limited visibility, optical distortions, turbidity, caustics, and non-linear light attenuation across depths.

Early work by Eustice et al. on large-scale hull mapping laid the groundwork for underwater Simultaneous Localization and Mapping (SLAM) techniques, using both vision-based and sonar-based methods for infrastructure inspection [10], [11]. These efforts targeted applications like ship hull inspection [12], oil rig monitoring, and pipeline surveying [13], which, while distinct from ecosystem monitoring, share challenges related to long-term feature tracking [14], sensor degradation, and environmental variability [9].

In this paper, we focus on long-term environmental monitoring enabled by recent advances in visual place recognition (VPR) and semantic segmentation. VPR enables autonomous systems to recognize previously visited locations based on a set of reference images. While VPR has the potential for accurate, repeatable localization, we show that current state-of-the-art techniques are challenged by new, difficult datasets in the underwater domain, which has received very limited attention thus far. Our approach requires only a set of pre-

collected images as a reference – there is no need for SLAM, Structure-from-Motion (SfM), or geometric mapping. This makes it simpler and more scalable for long-term monitoring.

As depicted in Figure 1, our system enables cost-effective and scalable underwater ecosystem monitoring by registering locations across multi-year timescales using monocular video from autonomous marine robots. It identifies revisited areas, estimates rigid transformations, and analyzes visual changes, even under varying trajectory overlap. By integrating VPR with feature matching and image warping, our approach enhances spatial consistency and registration accuracy over time. While our method builds upon existing techniques, we are the first to combine them in this manner, enabling critical monitoring capabilities that are valuable to ecologists.

Finally, we introduce the **SQUIDLE+ VPR Benchmark**, the first large-scale benchmark for evaluating underwater VPR, leveraging publicly available data from SQUIDLE+. Through extensive experiments, we benchmark state-of-the-art VPR techniques for underwater environments, advancing precise registration for long-term ecological monitoring and improving the reliability, scalability, and effectiveness of autonomous marine surveys for scientific research and marine conservation.

II. RELATED WORKS

We review image registration, 3D mapping, and VPR as key components of long-term underwater ecosystem monitoring. Traditional image registration (Section II-A) relies on GPS or multi-sensor SLAM for coarse localization before alignment refinement, while our approach uses VPR to efficiently retrieve images from monocular video for precise registration. In 3D mapping, existing methods propagate semantic information within a single trajectory, whereas we extend segmentation across multi-year datasets under varying conditions (Section II-B). Underwater VPR remains underexplored, with most techniques designed for structured terrestrial environments (Section II-C). Finally, we review underwater localization benchmarking in Section II-D.

A. Image registration for change detection

Delaunoy et al. [9] used SURF features to register and detect changes in images taken 10 months apart in a small coral reef scene. Similarly, Williams et al. [15] applied a stereo-based SLAM technique with SIFT features to co-register multiple 3D image maps collected by an autonomous underwater vehicle (AUV) during a grid survey. Their approach compared the same area over 12 hours, estimating the AUV's trajectory across overlapping grid-based surveys.

Bryson et al. [16] advanced multi-year repeat survey imagery processing and precision registration for monitoring long-term changes in benthic marine habitats. Post-processing of stereo imagery and navigation data through SLAM and 3D reconstruction generated a photo-textured model of the seafloor, orthographically projected into a geo-referenced mosaic. Their approach included mutual information optimization to improve robustness against variations in color and brightness across years.

These studies relied on GPS or multi-sensor SLAM for coarse localization before applying image registration methods. In contrast, our method efficiently retrieves image cues from the same location using videos from an uncalibrated monocular camera. This enables accurate image registration as required for monitoring applications.

B. 3D Mapping of Coral Reefs

Recent advances in underwater 3D mapping have improved coral reef reconstruction and semantic interpretation. Sauder et al. [17] proposed a pipeline to propagate semantic segmentation from images to 3D point clouds estimated via SfM from single-trajectory video streams. Sethuraman et al. [18] leveraged neural radiance fields (NeRFs) for physics-informed novel view synthesis and image restoration, addressing water column effects like attenuation and backscattering. Wang et al. [7] combined visual-inertial odometry with real-time 3D reconstruction to generate dense maps on resource-constrained AUVs, achieving results comparable to offline methods. Song et al. [19] introduced TURTLMap, designed for real-time localization and mapping in low-texture underwater regions.

Unlike these approaches that typically focus on single-time mapping, require controlled camera paths, or address specific visibility challenges, our method enables propagation of semantic segmentation across images captured in different years under varying conditions with arbitrary trajectory overlaps. Our pipeline prioritizes image-to-image registration for direct ecological comparison using only monocular video input, making it more accessible for practical long-term monitoring applications.
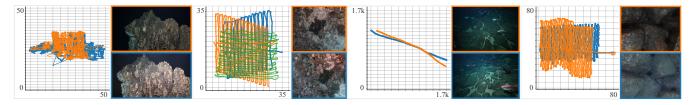
C. Visual Place Recognition

Despite its potential, underwater VPR remains largely underexplored and is rarely applied in practical settings. Most VPR pipelines [20]–[25] are developed and tested in structured environments, with limited application to unstructured settings—and almost none for underwater imagery. A thorough review of terrestrial place recognition techniques is beyond the scope of this paper, but excellent surveys exist for the interested reader [26]–[29].

Of interest for this paper are AnyLoc [30], which is one of the few VPR techniques evaluated on the Eiffel Tower underwater dataset [31], and MegaLoc [32], a versatile image retrieval model that achieved state-of-the-art performance in VPR, visual localization, and landmark retrieval by leveraging a diverse set of data and training techniques.

Place recognition is often used in a hierarchical pipeline, where the top K retrieved database images are reranked using more computationally intensive local feature matching. A widely used example is hloc [33], which uses NetVLAD [20] for image retrieval and SuperPoint [34] with SuperGlue [35] pre-trained on outdoor scenes for local matching.

Several works specifically address place recognition in underwater environments. Maldonado-Ramirez et al. [36] propose an unsupervised VPR method using a convolutional autoencoder to learn compact representations from salient landmarks detected by a visual attention algorithm. Burguera

(a) Eiffel Tower 2018-2020   (b) Okinawa 2016-2017-2018   (c) Tasman Fracture 2018/12/04-06   (d) St Helens 2011-2013

Fig. 2. **GPS trajectories (in meters) and environmental differences across underwater datasets.** The visualized trajectories depict query sequences (orange) and database sequences (blue) for four underwater datasets: (a) the *Eiffel Tower* (2018–2020) from the Mid-Atlantic Ridge hydrothermal vent, (b) *Okinawa* (2016–2017–2018), capturing mesophotic coral reef environments before and after Typhoon Trami, (c) the *Tasman Fracture* (2018), showcasing deep-sea benthic habitats, and (d) *St Helens* (2011–2013), featuring images recorded during a transition movement in shallow barren zones. To analyze the unique overlap between sequences covering the same area but following different trajectories, we include a third trajectory in the Okinawa dataset. Accompanying RGB image pairs for each trajectory illustrate corresponding locations, highlighting the significant appearance variations that challenge automated VPR systems. Notably, even correctly matched locations exhibit substantial visual differences due to variations in viewpoint, lighting conditions, and actual ecosystem changes over time, further emphasizing the complexity of long-term visual place recognition in underwater environments.

et al. [37] introduce a deep network for fast, robust underwater loop detection with clustered SIFT features and unsupervised training. While these methods focus on underwater VPR and evaluate self-recorded robotic sequences, our approach is tested on large, publicly available benchmarks covering diverse sequences, robotic platforms, and camera setups, aligning more closely with general VPR evaluation.

D. Underwater Localization Benchmarking

Boittiaux et al. [31] introduced the Eiffel Tower dataset, a deep-sea dataset for long-term visual localization. Ferrera et al. [38] presented AQUALOC to support visual–inertial–pressure SLAM for underwater vehicles, providing offline trajectories from SfM for comparison with real-time localization methods.

Joshi et al. [39] proposed a Visual-Inertial SLAM pipeline, evaluated on an artificial wreck off South Carolina and in Florida's caverns and caves. Singh et al. [40] introduced a dataset for benchmarking refractive camera model estimation, featuring a time-synchronized 5-camera/IMU setup on a remotely operated vehicle (ROV) in a controlled pool environment.

Angelakis et al. [41] used animal-borne video and movement data from Australian sea lions to map benthic habitats, addressing challenges in vessel-based surveys, which are costly, time-consuming, weather-dependent, and impractical for deep-sea environments. Joshi et al. [42] presented a 3D water quality mapping system for shallow waters using a BlueROV2 with GPS and a water quality sensor, enabling location correction by resurfacing when errors occur.

In this paper, we make a step towards standardised benchmarking of VPR methods in underwater scenarios, akin to efforts in terrestrial VPR [43], [44]. Specifically, we propose to leverage sequences freely available on `https://squidle.org/`. SQUIDLE+ hosts the largest repository of openly accessible georeferenced marine images, enabling the construction of datasets tailored to specific scientific questions and sourced from various platforms, campaigns, and globally distributed deployments. By leveraging this vast data source, our work establishes the foundation for a dedicated

TABLE I

SQUIDLE+ VPR BENCHMARK ORIGINATES FROM VARIOUS ROBOTIC PLATFORMS FOLLOWING TRAJECTORIES WITH ARBITRARY OVERLAP, SPANNING TIME DIFFERENCES RANGING FROM DAYS TO YEARS. THE DATA INCLUDES DIVERSE SEAFLOOR TYPES CAPTURED UNDER VARYING SCENE CONDITIONS (E.G., DEPTH, LIGHTING, TURBIDITY).

	Dataset	Year	# Images (k)	Dur. (h)	Dist. (km)	Avg. Depth (m)	Seafloor	Platform	Loc. Radius (m)
SQUIDLE+	Okinawa	2018	2.8	01:16	0.7	37	Mesophotic coral	SOTON/UTOK OPLab AUV	5.0
		2017	3.0	02:04	0.7	36			
		2016	2.1	00:58	0.5	38			
	Tasman Fracture	2018/12/04	1.2	00:57	2.3	885	Seamount habitats	CSIRO O&A MRITC Towed Stereo Camera	10.0
		2018/12/06	1.3	00:56	2.3	885			
	St Helens	2013	6.5	01:48	2.9	12	Shallow barren zones	IMOS AUV Sirius	3.0
		2011	5.5	01:30	1.9	12			
		2009	6.4	01:47	2.7	10			
	Eiffel Tower (Mid-Atlantic Ridge)	2020	3.9	03:18	0.9	1.7k	Hydrothermal vent edifice	EMSO-Azores ROV	0.5
		2018	5.4	17:19	1.0	1.7k			
		2016	3.6	03:24	0.9	1.7k			
		2015	4.9	01:26	0.3	1.7k			

underwater VPR benchmark, facilitating the development and rigorous evaluation of VPR pipelines.

III. TECHNICAL APPROACH

This section details our approach to enable change detection in underwater scenes using video streams from uncalibrated monocular cameras.

Given two sequential image sets, Q and D, from a video stream, we first find the top K matches in D for each Q_i using global descriptors obtained via state-of-the-art VPR techniques [20], [22], [23], [25], [30], [32]. Using LightGlue [45], we rerank these matches via the inlier count, and filter out image pairs that have a large reprojection error to avoid false matches. We then extract semantic segmentation masks from the matched images, and warp the images using the homography matrix. We use the intersection-over-union (IoU) metric to approximate location similarity.

A. Visual Place Recognition – Global Retrieval Stage

We extract image descriptors using standard VPR techniques and compute the similarity matrix by evaluating the

L2 norm between all image pairs: $S(Q, D) \in \mathbb{R}^{q \times d}$.

For each query image Q_i, we select the top K database images D_k with the highest similarity scores, forming the set of initial matches:

$$\mathcal{M}_K = \{(Q_i, D_k) \mid D_k \in \text{TopK}(S(Q_i, D), K)\}, \quad (1)$$

where $\text{TopK}(\cdot, K)$ denotes the function that selects the K most similar matches.

B. Keypoint Correspondences – Local Refinement Stage

We refine matches by using LightGlue to establish keypoint correspondences $\mathcal{K}_{qd}$ in image space between SuperPoint features in $\mathcal{M}_K$:

$$\mathcal{K}_q = \{\boldsymbol{k}_q \mid \boldsymbol{k}_q \in \mathbb{R}^3\}, \quad \mathcal{K}_d = \{\boldsymbol{k}_d \mid \boldsymbol{k}_d \in \mathbb{R}^3\}, \quad (2)$$

where $\boldsymbol{k}_q$ and $\boldsymbol{k}_d$ represent keypoints expressed in homogeneous coordinates.

We then re-rank the matches based on the inlier count as a new similarity measure, selecting the best match per query:

$$\mathcal{M}_1 = \{(Q_i, D_1) \mid D_1 = \text{TopK}(\text{inliers}(Q_i, D_k), 1)\}. \quad (3)$$

C. Homography Estimation and Outlier Filtering

From the image correspondences $\mathcal{K}_{qd}$ in the best match subset $\mathcal{M}_1$, we estimate the homography matrix $H \in \mathbb{R}^{3 \times 3}$ and project the keypoints from one image onto their corresponding locations in the other. The reprojection error e_r is computed as the mean root-mean-square (RMSE) distance between keypoints projected onto the opposite image:

$$e_r = \frac{1}{2\sqrt{|\mathcal{K}_{qd}|}} \left(\sqrt{\sum_{(\boldsymbol{k}_q, \boldsymbol{k}_p) \in \mathcal{K}_{qd}} \|\boldsymbol{k}_q - H\boldsymbol{k}_p\|^2} \right.$$
$$\left. + \sqrt{\sum_{(\boldsymbol{k}_q, \boldsymbol{k}_p) \in \mathcal{K}_{qd}} \|H^{-1}\boldsymbol{k}_q - \boldsymbol{k}_p\|^2} \right), \quad (4)$$

where $|\mathcal{K}_{qd}|$ denotes the number of matched keypoints. To account for scale variations—such as when the query image is a zoomed-in version of the database image or vice versa—we *average* bidirectional errors by computing transformations in both directions.

Finally, we discard image pairs with a reprojection error greater than $\chi = 10$ pixels, ensuring only geometrically consistent matches are retained:

$$\mathcal{M}_\chi = \{(Q_i, D_1), \ e_r(Q_i, D_1) \leq \chi\}. \quad (5)$$

D. 2D Warping of Segmentation Masks

To simulate a potential change detection method for visual monitoring on the registered image pairs $\mathcal{M}_\chi$, we automatically extract segmentation masks for each image using Segment-Anything 2 (SAM2) [46]. For simplicity in our experiments, all segmented instances within an image are merged into a single unified mask. We then use the homography matrix to warp the masks into a common image space, enabling pixel-level comparison using pixel intersection over union (IoU) as a similarity proxy.

IV. EXPERIMENTAL SETUP

A. VPR Baselines

For global retrieval, we evaluate state-of-the-art visual place recognition (VPR) methods implemented in [47], including MixVPR [25], CosPlace [22], NetVLAD [20], AnyLoc [30], MegaLoc [32], and CricaVPR [23]. While our pipeline supports all models referenced in [47], we select this subset due to their strong performance and widespread adoption in terrestrial VPR. For local feature refinement, we use SuperPoint [34] keypoints in combination with LightGlue [45].

To assess whether the retrieval results, based on GPS-derived ground truth, correspond to meaningful place recognition, we compare the evaluated VPR methods against a random guesser, which serves as a lower performance bound. The random guesser is implemented as a Monte Carlo experiment: for each query image Q_i, we randomly select K database matches, repeating this selection n times per query. The Recall@K metric is then defined as the proportion of iterations where at least one correct match was retrieved among the K selected database images, normalized by the total number of trials, $|Q| \cdot n$, where $|Q|$ is the number of query images and n is chosen to be sufficiently large. To establish an upper performance bound, we employ a brute-force baseline that performs exhaustive local feature matching on every query-database image pair using LightGlue. In this case, the number of keypoint matches serves as the similarity measure between image pairs.

B. SQUIDLE+ VPR Benchmark

In this work, we propose using data from SQUIDLE+ to advance underwater VPR by leveraging a vast collection of unstructured data. This dataset originates from various robotic platforms executing diverse trajectory patterns with arbitrary overlap, spanning time differences ranging from days to years. The data includes diverse seafloor types captured under varying out-of-distribution conditions of the scene (e.g., depth, lighting, turbidity).

Table I provides details of the selected sequences, while Figure 2 illustrates the robot trajectories and presents sample RGB images depicting the scene variations.

We process the sequences by selecting only those captured with downward-looking cameras, the most common setup for exploring seafloor environments. This distinguishes our benchmark from most datasets in structured environments captured with forward-facing cameras, where the majority of VPR techniques are trained [30], making it more comparable to aerial datasets [48].

To ensure consistent processing times across datasets, we downsample the RGB images to a resolution of approximately 640×480 pixels, adjusted to maintain the original aspect ratio. We include AUV sequences with GPS signals to evaluate our retrieval pipeline using an image-independent ground truth for visual localization assessment [49].

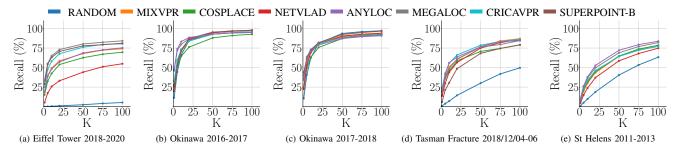Our dataset includes the following sequences:

(a) Eiffel Tower 2018-2020 (b) Okinawa 2016-2017 (c) Okinawa 2017-2018 (d) Tasman Fracture 2018/12/04-06 (e) St Helens 2011-2013

Fig. 3. **Recall@K performance for VPR methods across underwater datasets.** Experimental results illustrating the probability that a correct match appears within the top K retrieved candidates. We compare six VPR methods—MixVPR, CosPlace, NetVLAD, AnyLoc, MegaLoc, and CricaVPR—against a random guesser to assess whether the retrieval results, given our GPS-based ground truth, correspond to meaningful location identification. Additionally, we include the brute-force SuperPoint approach (which performs feature matching on all possible image pairs) to establish an upper performance bound. † In the Okinawa dataset, the random guesser's performance approaches that of VPR methods for $K \gtrsim 10$ due to the lawnmower trajectory pattern (see Fig. 2), which increases image density. This allows random selection to occasionally retrieve correct matches, even without visual correspondence. While this does not compromise the GPS-based ground truth, it underscores that VPR methods are most discriminative at lower K values (see Section V-A). ‡ Due to its significantly higher computational cost, we do not include the SuperPoint brute-force approach in the St Helens dataset.

TABLE II

RECALL@K PERFORMANCE FOR VPR METHODS ACROSS UNDERWATER DATASETS. WE USE GREEN TO HIGHLIGHT THE BEST R@10 RESULT FOR EACH DATASET AND ORANGE FOR THE SECOND-BEST R@10 RESULT EXCLUDING SUPERPOINT BRUTE-FORCE. ANYLOC, MEGALOC, AND CRICAVPR PERFORM COMPARATIVELY. WE CHOSE MEGALOC FOR SUBSEQUENT EXPERIMENTS AS IT IS SIGNIFICANTLY FASTER THAN ANYLOC AND PERFORMS SLIGHTLY BETTER THAN CRICAVPR ON AVERAGE. † WE COMPARE RESULTS USING R@10, AS THIS IS THE LARGEST VALUE THAT PROVIDES MEANINGFUL COMPARISONS GIVEN THE GPS-BASED GROUND TRUTH (SEE EXPLANATION IN FIGURE 3).

Methods	Eiffel			Okinawa 2016-2017			Okinawa 2017-2018			Tasman Fracture			St Helens		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MixVPR	19.2	39.7	49.5	21.9	52.6	70.9	25.3	55.8	69.5	15.5	38.5	51.1	7.5	19.8	29.9
CosPlace	15.2	32.5	42.5	27.5	51.1	64.3	22.8	49.5	63.4	20.5	35.9	45.6	6.9	21.2	31.4
NetVLAD	5.2	17.4	25.8	25.9	56.8	73.1	22.7	53.1	69.2	13.8	29.7	43.9	4.6	14.7	23.8
AnyLoc	18.6	38.7	48.3	46.1	73.8	83.3	34.3	60.1	70.6	19.3	41.4	56.0	10.3	26.4	37.9
MegaLoc	33.7	55.1	65.1	47.2	72.0	79.7	40.2	63.9	73.0	18.4	36.8	48.5	9.5	25.6	36.5
CricaVPR	25.1	47.7	58.1	32.6	59.5	73.1	37.2	60.2	71.0	23.8	42.7	55.6	9.3	23.4	33.5
Superpoint Brute-Force	28.9	54.3	63.1	19.6	51.5	71.3	43.8	61.3	71.8	5.9	21.3	30.5	-	-	-

TABLE III

AVERAGE COMPUTATION TIME PER QUERY (MS) FOR EACH VPR METHOD. ALL ONE-STAGE VPR TECHNIQUES ARE SIGNIFICANTLY FASTER THAN THE BRUTE-FORCE SUPERPOINT APPROACH. OUR HIERARCHICAL METHOD, COMBINING MEGALOC AND SUPERPOINT, IS ×100 FASTER THAN BRUTE-FORCE SUPERPOINT.

Methods	Avg. Time per Query (ms)
MixVPR	8
CosPlace	14
NetVLAD	21
AnyLoc	428
MegaLoc	49
CricaVPR	38
Superpoint-B	27150
AnyLoc + Superpoint-H	49 + 226

Okinawa[1]: This dataset captures the mesophotic coral seafloor off Sesoko Island, near Okinawa, over a two-year period using the AUV *Tuna-Sand* [50]. The 2018 sequence was recorded days after impact from Typhoon Trami (Paeng). Visual changes from damage to the environment hinder relocalization, emphasizing the need for a robust pipeline to document damage and recovery.

Tasman Fracture[2]: This dataset consists of benthic stereo still imagery collected to map the distribution of epibenthic fauna and habitats on seamounts off Tasmania [51]. The imagery was acquired at a depth of 885 meters over a two-day period using a deep-towed camera.

St Helens[3]: This dataset comprises benthic stereo-imaging surveys conducted with the AUV *Sirius* to monitor long-term urchin population dynamics in St Helens, northeast Tasmania [52]. As part of a bi-yearly monitoring program, the AUV surveyed shallow barren/kelp transition zones, which are included in our benchmark.

We include the **Eiffel Tower**[4] [31] dataset alongside the SQUIDLE+ sequences to provide a curated reference specifically designed for benchmarking long-term underwater visual localization. The dataset consists of images captured during four visits to the same hydrothermal vent edifice over a five-year period. Camera poses and a common scene geometry were estimated using navigation data and SfM. In contrast to [30], where only a small subset of images was used (∼1%), we incorporate entire sequences for a more comprehensive

[1] https://oceanperception.com/impact/expeditions/

[2] https://data.csiro.au/collection/60886
[3] https://www.researchgate.net/publication/294422138_Stereo-imaging_AUV_detects_trends_in_sea_urchin_abundance_on_deep_overgrazed_reefs
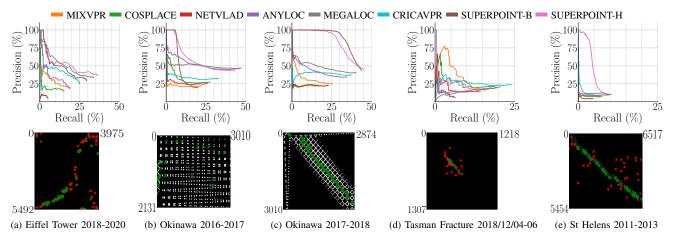[4] https://www.seanoe.org/data/00810/92226/

Fig. 4. **Top row:** Precision-Recall curves for Best-Single-Match VPR. The *SuperPoint-H* method outperforms one-stage VPR approaches, achieving performance closer to the SuperPoint brute-force approach while significantly reducing computational cost (see Table III). **Bottom row:** The background represents the binary ground truth in black and white, with green crosses indicating true positives and red crosses indicating false positives for the best match per query using the *SuperPoint-H* method. Positive matches are filtered to exclude those with a reprojection error greater than 10 pixels, corresponding to a precision of 39% for the Eiffel Tower dataset, 99% for Okinawa, 22% for Tasman Fracture, and 72% for St Helens.

evaluation.

For all sequences, we obtain the ground truth matrix where each pair of images is considered to depict the same place if their spatial distance is smaller than a predefined localization radius (see Table I).

C. Evaluation Metrics

We assess VPR methods using established metrics, as they are widely recognized in the field and provide a clear framework for evaluating performance. This approach enables us to contextualize the behavior of the methods within the standards of the research community while also highlighting their effectiveness on challenging underwater datasets.

Recall@K: The probability that at least one of the top K matches, ranked by similarity for a given query, is a correct match (i.e., a true positive).

Precision-Recall Curves: A precision-recall curve evaluates VPR performance when considering only the highest-ranked match per query. This analysis highlights the trade-off between precision and recall, providing insights into the reliability of the top-ranked match in correctly identifying revisited locations.

V. RESULTS AND DISCUSSION

A. Hierarchical Image Retrieval

Global Retrieval Stage. Figure 3 illustrates the Recall@K performance for eight baselines, including the random guesser and brute-force SuperPoint approach (see Section IV-A).

A notable observation in the Okinawa dataset is that for $K \gtrsim 10$, the random guesser's performance begins to approach that of specialized VPR methods. This phenomenon stems from the structured lawnmower trajectory pattern (see Figure 2), which creates a high spatial density of images. In such dense patterns, even random selection can occasionally retrieve images within our defined localization radius, despite lacking the visual correspondence that VPR methods identify. This does not indicate an issue with our ground truth, but

rather demonstrates that for highly structured survey patterns, the discriminative advantage of VPR methods is most evident at lower K values. The precision-recall curves in Figure 4 further illustrate this distinction, showing that VPR methods achieve substantially higher precision than random selection when evaluating the highest-ranked matches.

Table II highlights the best Recall@10 results and the second-best for each dataset, providing a comparative assessment across methods. MegaLoc is significantly faster than AnyLoc and performs slightly better than CricaVPR on average. We use R@10 for evaluation, as higher values introduce ambiguity due to the limitations of GPS-based ground truth, as discussed above.

Brute-force SuperPoint establishes an upper bound for retrieval performance across almost all datasets but incurs a significantly higher computational cost. Given this overhead, brute-force SuperPoint is excluded from experiments on the St Helens dataset.

Local Refinement Stage. Based on prior experiments, we select MegaLoc as the global image retrieval method in our hierarchical approach. Figure 4 shows precision-recall curves for the best single-match results across eight baselines. Our hierarchical method, combining MegaLoc with SuperPoint, achieves performance comparable to brute-force SuperPoint (average precision of 16% for our hierarchical method vs 18% for brute-force) while significantly reducing computational overhead (×100 faster). A key advantage is its ability to first narrow down candidate matches before feature-based refinement, yielding substantial computational savings (see Table III).

The bottom row of Figure 4 illustrates the ground truth evaluation, where correctly identified matches are marked in green and false positives appear in red. To ensure reliable evaluation, we threshold with reprojection errors smaller than 10 pixels for a match to be considered valid. This filtering results in precision levels of 39% for the Eiffel Tower dataset, 99% for Okinawa, 22% for Tasman Fracture, and 72% for
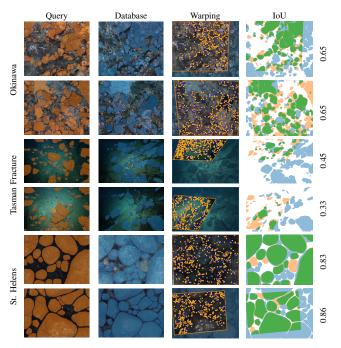
Query   Database   Warping   IoU

Okinawa — 0.65
Okinawa — 0.65
Tasman Fracture — 0.45
Tasman Fracture — 0.33
St. Helens — 0.83
St. Helens — 0.86

Fig. 5. **Qualitative results of warping segmentation masks and reprojection error, accompanied by intersection over union (IoU) scores for aligned masks. Leftmost column:** Query RGB images from each SQUIDLE+ dataset overlaid with SAM2 segmentation masks. **Second column:** Database match obtained using our MegaLoc + SuperPoint hierarchical method, overlaid with SAM2 segmentation masks. **Third column:** Query images warped onto the selected database image using the homography estimated from LightGlue keypoints. Actual keypoints from the database image are plotted as blue circles, while projected keypoints from the query image are shown as orange crosses. The reprojection distance is marked with a yellow dashed line. **Rightmost column:** Query masks warped using the estimated homography and overlaid onto the database masks. Areas of intersection are colored in green, while non-overlapping query and database masks remain orange and blue, respectively.

St Helens.

To quantify computational efficiency, Table III presents the average computation time per query for each VPR method. By combining MegaLoc with SuperPoint in a hierarchical fashion, our method achieves ×100 speedup over the brute-force SuperPoint approach, demonstrating a favorable trade-off between accuracy and efficiency. All experiments were conducted on a Dell Precision 3680 Tower equipped with an NVIDIA RTX 4090 GPU.

B. Segmentation Warping & Qualitative Results

Figure 5 presents the results of warping segmentation masks generated with SAM2 using the homography transformation estimated from LightGlue keypoint correspondences. We evaluate the alignment between the two warped masks using an intersection over union (IoU) metric, where the intersection is defined as the number of shared pixels between the query and database masks, and the union represents the total number of pixels covered by both masks in the warped image.

This simple evaluation provides a qualitative comparison, demonstrating how a change detection method for underwater ecosystem monitoring can be seamlessly integrated with our image-based registration approach. However, certain limitations affect the accuracy of our change detection: (1)

inconsistencies in the segmentation results, (2) the inability to differentiate between warping inaccuracies, appearance changes, and actual structural modifications in the scene.

Qualitative observations reveal that scenes with minimal changes, such as the rocky environment in St Helens, achieve higher IoU values. In contrast, the significant appearance changes caused by Typhoon Trami in the Okinawa sequence result in lower IoU values, capturing the impact of environmental disturbances. Finally, the Tasman Fracture sequence demonstrates the capability of our approach to register images despite strong viewpoint variations. However, these variations also introduce inconsistencies in image segmentation, leading to further reductions in IoU scores. This highlights the potential of combining our approach with more streamlined comparison methods to enable change detection under extreme viewpoint changes.

VI. CONCLUSIONS AND FUTURE WORK

We present the first comprehensive benchmark for VPR in underwater environments, addressing a key gap in long-term marine monitoring. Our hierarchical approach, combining global descriptors with local feature refinement, achieves robust performance across diverse underwater scenes over time spans from days to years. Furthermore, precise image registration enables pixel-level comparison between temporally separated observations—critical for marine conservation, where the vast scale of underwater habitats makes manual assessment impractical.

Future work will focus on fully automating the benchmarking process by leveraging all available resources within SQUIDLE+, enabling the most extensive evaluation of underwater localization pipelines to date. Futhermore, we motivate the development of reliable temporal maps integrated with automatic change detection tools, ensuring scalable and efficient monitoring of underwater ecosystems.

REFERENCES

[1] O. Hoegh-Guldberg and J. F. Bruno, "The impact of climate change on the world's marine ecosystems," Science, vol. 328, no. 5985, pp. 1523–1528, 2010.

[2] T. P. Hughes et al., "Coral reefs in the anthropocene," Nature, vol. 546, no. 7656, pp. 82–90, 2017.

[3] E. M. Madin, E. S. Darling, and M. J. Hardt, "Emerging technologies and coral reef conservation: Opportunities, challenges, and moving forward," Frontiers in Marine Science, vol. 6, p. 727, 2019.

[4] M. Dunbabin and L. Marques, "Robots for environmental monitoring: Significant advancements and applications," IEEE Robotics & Automation Magazine, vol. 19, no. 1, pp. 24–39, 2012.

[5] B. Zhang, D. Ji, S. Liu, X. Zhu, and W. Xu, "Autonomous underwater vehicle navigation: A review," Ocean Engineering, vol. 273, 2023.

[6] E. Iscar, K. A. Skinner, and M. Johnson-Roberson, "Multi-view 3D reconstruction in underwater environments: Evaluation and benchmark," in OCEANS, 2017.

[7] W. Wang, B. Joshi, N. Burgdorfer, K. Batsosc, A. Q. Lid, P. Mordohaia, and I. Rekleitisb, "Real-time dense 3D mapping of underwater environments," in IEEE International Conference on Robotics and Automation, 2023, pp. 5184–5191.

[8] S. Raine, R. Marchant, B. Kusy, F. Maire, and T. Fischer, "Image labels are all you need for coarse seagrass segmentation," in IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5943–5952.

[9] O. Delaunoy, N. Gracias, and R. Garcia, "Towards detecting changes in underwater image sequences," in OCEANS, 2008, pp. 1–8.

[10] R. Eustice, H. Singh, J. J. Leonard, M. R. Walter, and R. Ballard, "Visually navigating the RMS Titanic with SLAM information filters," in Robotics: Science and Systems, vol. 2005, 2005, pp. 57–64.

[11] R. M. Eustice, H. Singh, and J. J. Leonard, "Exactly sparse delayed-state filters," in IEEE International Conference on Robotics and Automation, 2005, pp. 2417–2424.

[12] A. Kim and R. Eustice, "Pose-graph visual SLAM with geometric model selection for autonomous underwater ship hull inspection," in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, pp. 1559–1565.

[13] N. Fairfield, G. Kantor, and D. Wettergreen, "Real-time SLAM with octree evidence grids for exploration in underwater tunnels," Journal of Field Robotics, vol. 24, no. 1-2, pp. 03–21, 2007.

[14] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: A multi-sensor fusion-based underwater SLAM system," The International Journal of Robotics Research, vol. 41, no. 11-12, pp. 1022–1042, 2022.

[15] S. B. Williams, O. Pizarro, M. V. Jakuba, I. Mahon, S. D. Ling, and C. R. Johnson, "Repeated AUV surveying of urchin barrens in north eastern tasmania," in IEEE International Conference on Robotics and Automation, 2010, pp. 293–299.

[16] M. Bryson, M. Johnson-Roberson, O. Pizarro, and S. Williams, "Automated registration for multi-year robotic surveys of marine benthic habitats," in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 3344–3349.

[17] 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.

[18] A. V. Sethuraman, M. S. Ramanagopal, and K. A. Skinner, "WaterNeRF: neural radiance fields for underwater scenes," in OCEANS, 2023.

[19] J. Song, O. Bagoren, R. Andigani, A. Sethuraman, and K. A. Skinner, "Turtlmap: Real-time localization and dense mapping of low-texture underwater environments with a low-cost unmanned underwater vehicle," in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2024, pp. 1191–1198.

[20] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1437–1451, 2018.

[21] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14 136–14 147.

[22] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2022, pp. 4878–4888.

[23] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16 772–16 782.

[24] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17 658–17 668.

[25] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "MixVPR: Feature mixing for visual place recognition," in IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2998–3007.

[26] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," Pattern Recognition, vol. 113, p. 107760, 2021.

[27] C. Masone and B. Caputo, "A survey on deep visual place recognition," IEEE Access, vol. 9, pp. 19 516–19 547, 2021.

[28] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" in International Joint Conference on Artificial Intelligence, 2021, p. 4416–4425.

[29] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual place recognition: A tutorial," IEEE Robotics & Automation Magazine, vol. 31, no. 3, pp. 139–153, 2024.

[30] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "AnyLoc: Towards universal visual place recognition," IEEE Robotics and Automation Letters, vol. 9, no. 2, pp. 1286–1293, 2024.

[31] C. Boittiaux, C. Dune, M. Ferrera, A. Arnaubec, R. Marxer, M. Matabos, L. V. Audenhaege, and V. Hugel, "Eiffel tower: A deep-sea underwater dataset for long-term visual localization," The International Journal of Robotics Research, vol. 42, no. 9, pp. 689–699, 2023.

[32] G. Berton and C. Masone, "Megaloc: One retrieval to place them all," 2025. [Online]. Available: https://arxiv.org/abs/2502.17237

[33] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12 716–12 725.

[34] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.

[35] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," 2020.

[36] A. Maldonado-Ramírez and L. A. Torres-Mendez, "Learning ad-hoc compact representations from salient landmarks for visual place recognition in underwater environments," in IEEE International Conference on Robotics and Automation, 2019, pp. 5739–5745.

[37] A. Burguera and F. Bonin-Font, "Visual loop detection in underwater robotics: an unsupervised deep learning approach," in IFAC World Congress, 2020, pp. 14 656–14 661.

[38] 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.

[39] B. Joshi, M. Xanthidis, S. Rahman, and I. Rekleitis, "High definition, inexpensive, underwater mapping," in IEEE International Conference on Robotics and Automation, 2022, pp. 1113–1121.

[40] M. Singh and K. Alexis, "Online refractive camera model calibration in visual inertial odometry," in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2024, pp. 12 609–12 616.

[41] N. Angelakis, G. L. Grammer, S. D. Connell, F. Bailleul, L. M. Durante, R. Kirkwood, D. Holman, and S. D. Goldsworthy, "Using sea lion-borne video to map diverse benthic habitats in southern australia," Frontiers in Marine Science, vol. 11, p. 1425554, 2024.

[42] K. Joshi, T. Liu, A. Williams, M. Gray, X. Lin, and N. Chopra, "3D water quality mapping using invariant extended kalman filtering for underwater robot localization," arXiv preprint arXiv:2409.11578, 2024.

[43] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, "Deep visual geo-localization benchmark," in IEEE Conference on Computer Vision and Pattern Recognition, 2022.

[44] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," International Journal of Computer Vision, pp. 1–39, 2021.

[45] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," IEEE/CVF International Conference on Computer Vision, pp. 17 581–17 592, 2023.

[46] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024.

[47] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in IEEE/CVF International Conference on Computer Vision, 2023, pp. 11 080–11 090.

[48] M. Schleiss, F. Rouatbi, and D. Cremers, "Vpair–aerial visual place recognition and localization in large-scale outdoor environments," arXiv preprint arXiv:2205.11567, 2022.

[49] E. Brachmann, M. Humenberger, C. Rother, and T. Sattler, "On the limits of pseudo ground truth in visual camera re-localisation," in IEEE/CVF International Conference on Computer Vision, 2021, pp. 6218–6228.

[50] T. Nakatani, T. Ura, Y. Ito, J. Kojima, K. Tamura, T. Sakamaki, and Y. Nose, "AUV TUNA-SAND and its exploration of hydrothermal vents at Kagoshima Bay," in OCEANS, 2008.

[51] C. Untiedt, F. Althaus, A. Williams, and K. Maguire, "Tasmanian seamount hill u: benthic imagery and annotations," Data Collection, 2023.

[52] S. Ling, I. Mahon, M. Marzloff, O. Pizarro, C. Johnson, and S. Williams, "Stereo-imaging AUV detects trends in sea urchin abundance on deep overgrazed reefs," Limnology and Oceanography: Methods, vol. 14, no. 5, pp. 293–304, 2016.