From Beach to Cliff: Adapting CoastSnap Citizen Science for Coastal Cliff Change Detection Using ML-Assisted Image Registration and Prompted Segmentation

Alfredo Jaramillo-Velez
Joint Centre for Disasters Research
Massey University
Wellington, New Zealand
0000-0002-4901-0416

Supun Gamlath
Department of Computer Science & Engineering
University Of Moratuwa
Moratuwa, Sri Lanka
0000-0001-7565-9545

Meenambika Chandirakumar
Department of Computer Science & Engineering
University Of Moratuwa
Moratuwa, Sri Lanka
0009-0003-5627-000X

Raj Prasanna
Joint Centre for Disasters Research
Massey University
Wellington, New Zealand
0000-0002-5608-6558

Saskia de Vilder
Engineering Geology Team
Earth Sciences New Zealand
Wellington, New Zealand
0000-0002-4531-2070

Sam McColl
Engineering Geology Team
Earth Sciences New Zealand
Wellington, New Zealand
0000-0002-2805-1761

Marion L. Tan
Joint Centre for Disasters Research
Massey University
Wellington, New Zealand
0000-0003-0309-4121

Carol Stewart
College of Health
Massey University
Wellington, New Zealand
0000-0002-3781-6308

Thanuja D. Ambegoda
Department of Computer Science & Engineering
University Of Moratuwa
Moratuwa, Sri Lanka
0000-0002-5059-3479

***Abstract*—** CoastSnap is a citizen science tool that uses repeat smartphone photographs from fixed stations to monitor coastal change, yet it has rarely been applied to coastal cliffs. We test a workflow for cliff change screening using a controlled pilot CoastSnap station at Ōnaero, New Zealand (iPhone 12), integrating (i) Machine Learning-assisted ground control point (GCP) transfer for image registration, (ii) pinhole camera calibration and reprojection-based geo-rectification onto a curved cliff-surface model, and (iii) prompted segmentation for isolating cliff-related features. Auto-GCP benchmarking shows a trade-off between precision and robustness: the Scale-Invariant Feature Transform (SIFT) model achieves low localisation error when successful, but is sensitive to changes in illumination. In contrast, the Local Feature Transformer (LoFTR) model provides a higher detection yield and more stable performance, but is sensitive to the threshold used. Nevertheless, it is well suited to operational use with human-in-the-loop verification. Calibration produced consistent parameters across repeat images, with focal length estimates matching device specifications. Visual segmentation based only in Regions of Interest (RoI) often merged adjacent objects, while text-guided Segment Anything Model (LangSAM) delineated cliff and debris more reliably than fracture-like features. Preliminary detection of the supply and removal of debris highlights the potential for quantifying the frequency and magnitude of mass movements at each cliff.

***Index Terms*—** Landslides, Citizen Science, Geo-rectification, Semantic Segmentation

I. Introduction

Coastal cliffs are highly dynamic landforms shaped by marine forcing (waves and tides) and subaerial processes, and they can generate sudden, high-consequence hazards through mass movement. These hazards threaten not only beach users at the cliff toe but also infrastructure and properties located along cliff tops. Effective risk management can be supported by monitoring, especially approaches that capture changes at a sufficient cadence to detect site-specific failure patterns and evolving instability.

A widely adopted method for coastal erosion monitoring is the citizen science initiative CoastSnap [1], which uses community-contributed photographs from fixed stations to document shoreline change, predominantly on sandy beaches. CoastSnap has rarely been applied to coastal cliffs, despite the potential value of frequent observations in these environments. Extending CoastSnap to cliffs requires overcoming challenges associated with heterogeneous smartphone imagery (device optics, viewpoint offsets) and variable environmental conditions (illumination, shadows, tides), which can confound image comparison and change detection.

This study evaluates the feasibility of adapting CoastSnap for coastal cliff change detection using a workflow that integrates machine learning (ML) assisted image registration and prompted segmentation. As an initial test case, we used a CoastSnap station installed at Ōnaero, Taranaki, New Zealand, on private property, where photo frequency and device characteristics were controlled (iPhone 12). Specifically, we (i) benchmark models for automatic ground control point (GCP) identification to support image alignment, (ii) estimate camera intrinsics and extrinsics and apply reprojection-based geo-rectification, and (iii) test visual and semantic prompting (SAM-based

https://doi.org/10.54389/XOGF1409

segmentation) to isolate key cliff elements (particularly debris/talus) and support time-series interpretation.

II. Methodology

A. Overall Workflow

The proposed methodology standardises heterogeneous CoastSnap-style citizen images to enable repeatable coastal-cliff change detection. The core design principle is to decouple geometric standardisation from semantic interpretation (i.e., we first align images to the same viewpoint, then identify and track cliff features), reducing sensitivity to smartphone variability and environmental conditions (e.g., illumination, viewpoint offsets and tides).

As summarised in Fig. 1, the workflow includes: (i) repeat image acquisition from fixed stations; (ii) quality checks; (iii) ML-assisted geometric processing (Auto-GCP identification, registration, and camera calibration for reprojection-based geo-rectification); (iv) prompted segmentation (visual prompts and SAM-based semantic prompts, including LangSAM) of cliff-related features; and (v) temporal interpretation of segmented outputs to screen for change cues.

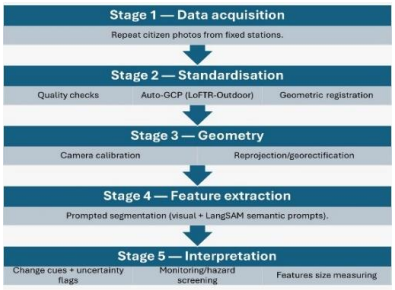

Figure 1. Overall workflow showing the five-stage pipeline from image acquisition to temporal change interpretation.

B. Ground Control Point Identification

To establish a geometric baseline, a set of reference ground control points (GCPs) was first defined on a high-quality reference photograph using clearly visible, (likely) static landmarks on the cliff and its surroundings (e.g., the top of a wooden post or a distinctive boulder corner; green marks in Fig. 2). For each reference GCP, 3D coordinates were obtained from a terrestrial laser scanning (TLS) survey to support subsequent camera calibration and geo-rectification. In total, eight GCPs were used at the Ōnaero site.

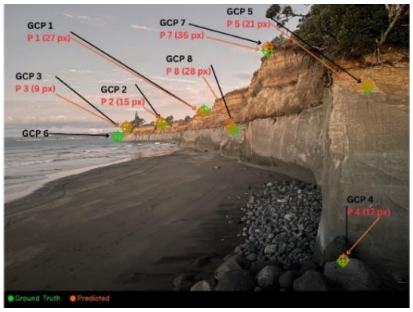

Figure 2. Example GCP localisation using the LoFTR-outdoor feature matching in a target image.

While manual annotation ensures accurate GCP locations on the reference image, the main aim of this stage is to automatically localise those same GCPs on incoming citizen photographs (orange marks in Fig. 2). This is framed as a feature detection and matching problem in which reference-image GCP positions must be transferred to each uncalibrated target image. We evaluated four feature-matching approaches spanning classical and learning-based methods: SIFT and ORB are traditional keypoint detector–descriptor pipelines (SIFT is scale- and rotation-invariant but can be sensitive to radiometric changes; Oriented FAST and Rotated BRIEF (ORB) is a fast binary alternative designed for efficiency). DIstillation with NO labels (DINOv2-large) is a self-supervised vision transformer whose learned representations can be used to derive correspondences under appearance variability. LoFTR (outdoor) is a transformer-based matcher that performs dense, detector-free local feature matching, typically providing more robust correspondences in challenging outdoor scenes

Model performance was assessed by comparing predicted GCP locations against manually annotated "ground truth" marks in each target image. A prediction was counted as a true positive only when (i) the correct GCP identifier was matched and (ii) the Euclidean pixel error was strictly below a distance threshold, $\boldsymbol{d}$ (50 px by default). With a native resolution of 4032 × 3024 px, d=50 px represents ~1.2% of the image width, providing a practical tolerance for localisation while penalising gross mismatches. Finally, to reduce the risk of residual outliers propagating into camera calibration, we applied a human-in-the-loop (HITL) quality assurance step: an operator reviews predicted GCPs, corrects them when needed, and approves the final set before calibration.

C. Camera Calibration

To map image pixels (u, v) to world coordinates (x, y, z), a camera calibration was performed using a pinhole camera model in homogeneous coordinates. The projection is expressed as:

$$s\begin{bmatrix}u\\v\\1\end{bmatrix} - K\,[R \mid t]\begin{bmatrix}x\\y\\z\\1\end{bmatrix} \quad (1)$$

where $\boldsymbol{K}$ is the intrinsic camera matrix, $\boldsymbol{R}$ is the rotation matrix parameterised by azimuth ($\boldsymbol{\alpha}$), tilt ($\boldsymbol{\tau}$), and roll ($\boldsymbol{\rho}$), $\boldsymbol{t}$ is the translation vector derived from the known camera location, and $\boldsymbol{s}$ is a scale factor. To constrain the model, we adopted standard assumptions: the principal point is fixed at the image centre, the pixel aspect ratio is set to 1:1, and skew is zero. Following [6], lens distortion was neglected because images were captured using the standard (non–wide angle) smartphone lens and modern devices typically apply in-camera distortion correction.

Under these assumptions, calibration reduces to four unknowns: focal length ($\boldsymbol{f}$), and the three orientation angles ($\boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\rho}$).

Parameters were estimated via non-linear least squares using the Trust-Region Reflective algorithm [7]. Initial values were obtained with a LookAt initialisation that

orients the camera towards the centroid of the eight GCPs, with the initial focal length set to the image width in pixels (i.e. f_0 = 4032 pixels). The optimisation minimised the reprojection error between observed 2D GCP positions and model projections, and calibration quality was summarised using the RMSE reprojection error across all eight GCPs.

To express focal length in physical units, we converted pixels to millimetres using the sensor width and image width:

$$f_{mm} = f_{px}\left(\frac{w_{sensor}}{w_{image}}\right) \quad (2)$$

where, w_{sensor} = 5.6 mm and w_{image} = 4032 pixels for the iPhone 12 standard lens.

D. Image Geo-rectification

Following camera calibration, images are geo-rectified by projecting pixels onto a three-dimensional representation of the coastal cliff surface (Fig. 3). Unlike standard CoastSnap beach rectification, which commonly assumes a planar ground surface, the cliff face is modelled as a curved, near-vertical surface to better represent its geometry at Ōnaero.

Figure 3. Schematic of the Ōnaero cliff surface model.

The across-cliff profile is represented by an n-th order polynomial, with n = 3 adopted for the Ōnaero site based on empirical fitting trials. The polynomial is fitted using the subset of GCPs located directly on the cliff face. A continuous surface mesh is then generated by extruding the fitted curve vertically between the surveyed cliff-toe (base) and cliff-top elevations, derived from the same topographic dataset used to georeference the GCPs.

Geo-rectification is performed by mapping image pixels to this 3D surface using the calibrated camera model (Eq. 1). The resulting output is a spatially standardised cliff-view image referenced to the NZTM coordinate system, enabling consistent measurement of distance along the cliff face and elevation. This provides a common geometric basis for the next stages; prompted segmentation and time-series change screening.

E. Prompted Segmentation

To delineate the cliff and isolate key cliff-related elements (cliff face, debris/talus, vegetation and fracture-like features), we applied prompted image segmentation. Image segmentation partitions an image into regions (pixel groups) corresponding to coherent objects or classes [8]. We tested two alternative methods to delineate cliff elements under heterogeneous citizen-image conditions: visual-prompt segmentation to constrain analysis to a RoI, and semantic prompt–based segmentation to extract specific feature classes using text prompts.

- *Visual Prompt Segmentation (RoI-based):* Visual prompting was first used to restrict segmentation to a cliff-focused RoI and reduce interference from sky/sea/background. For each image, a rectangular RoI was interactively defined as a bounding box around the cliff sector of interest. The approach was then extended to batches of images acquired from the same station and similar viewing configuration by reusing a single bounding box.

  Segmentation within the RoI was performed using the SAM [8]. Given an image and a bounding box prompt, SAM produces candidate masks (binary pixel maps that label which pixels belong to the object of interest) consistent with the prompt. In this study, the candidate mask with the highest score was selected as the default output for subsequent analysis.

- *Semantic Segmentation:* As the alternative approach, we applied text-prompted segmentation using class descriptors such as “cliff”, “vegetation” and “debris”. We used LangSAM, which combines GroundingDINO for text-guided object localisation with SAM for mask generation [8,9]. This enables class-wise mask extraction driven by semantic prompts, supporting consistent identification of cliff-related features across subsequent images.

F. Tracking debris change

A key objective of cliff-feature segmentation is to support the identification of geomorphic change, particularly changes in debris/talus at the cliff toe that represent supply from landslides or removal or reworking by wave action. As an initial step, we used semantic prompt–based segmentation to extract debris masks from images acquired at different time steps and then explored change by comparing masks between consecutive dates.

Debris change was assigned into three categories: (i) persistent/common areas (debris present in both epochs), (ii) newly added areas (debris detected only in the later image), and (iii) removed areas (debris detected only in the earlier image). Because apparent differences can be influenced by environmental variability (e.g., tidal stage, wetness and illumination), these outputs are treated as preliminary indicators. Ongoing work will therefore incorporate two mitigation steps: (1) filtering image pairs to comparable acquisition conditions where possible (e.g., similar tidal stage and limited shadowing), and (2) applying quality gates to exclude images affected by sea spray, blur or glare prior to change comparison.

III. Preliminary Results

A. Ground Control Point Identification

The four (GCP) feature-matching approaches were evaluated using the Ōnaero image set. Performance is summarised as detection accuracy at multiple pixel-distance thresholds ***d***, where a detection is considered correct when it matches the correct GCP identifier and falls within ***d*** pixels of the manually annotated location. Table 1 reports accuracy at selected thresholds, together with the mean distance error (MDE) computed over correct detections.

Table 1. Auto-GCP identification accuracy (%) at selected pixel-distance thresholds and mean distance error (MDE).

Model	Pixel-distance thresholds					MDE (px)
	10	20	30	50	100	
LoFTR	7	19	39	53	68	25
SIFT	44	47	49	50	53	7
ORB	27	32	38	43	49	18
DINOv2	4	11	16	30	54	32

Two distinct localisation behaviours were identified. SIFT provides the highest precision when it succeeds: 44 % of GCP instances are correctly identified within 10 px, and it yields the lowest MDE (7 px). Accuracy increases only marginally as the threshold relaxes (to 53% at 100 px), indicating that SIFT tends to either localise a GCP very accurately or fail to transfer it. Inspection of the low-accuracy reference–target pairs in Fig. 4 (dark-red cells), by cross-checking the corresponding images, indicates that these failures (low accuracy) are predominantly associated with low illumination and/or marine spray affecting image clarity; this suggests that SIFT performance is sensitive to such environmental conditions.

In contrast, LoFTR exhibits the strongest robustness to uncontrolled conditions. Although it trails SIFT at tight thresholds (e.g., 7% at 10 px), accuracy increases steeply with tolerance, reaching 53% at 50 px and 68% at 100 px (highest among all models). This indicates broader GCP coverage at the expense of greater spatial scatter (MDE 25 px), which is expected in a fully automated transfer setting. ORB showed intermediate behaviour (up to 48.61% at 100 px, MDE 18 px). DINOv2 performed poorly at operationally useful thresholds (≤50 px; 30% at 50 px) and produced the largest MDE (32 px), despite improving to 54.17% at 100 px, suggesting limited suitability for precise point localisation.

More broadly, the confusion matrices (Fig. 4) indicate that GCP identification accuracy depends not only on target-image quality but also on the choice of reference image. Images captured under clear, well-lit (but not overexposed) conditions consistently yield higher detection rates across targets, whereas overcast or low-illumination reference images degrade matching performance for most targets. Conversely, some target images remain difficult to match regardless of reference quality, typically those affected by strong shadows, marine spray or unusual lighting.

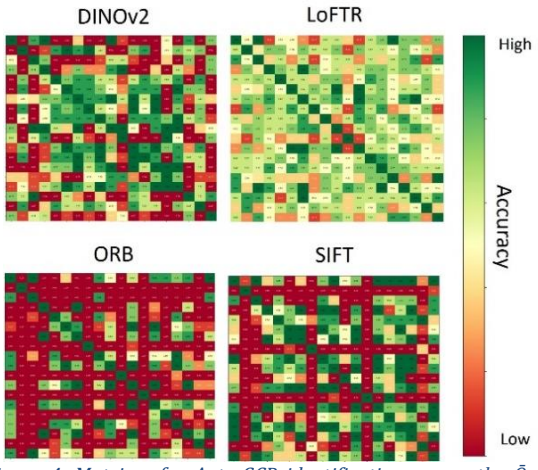

Figure 4. Matrices for Auto-GCP identification across the Ōnaero image set. Rows and columns correspond to the 19 images, used as reference (rows) and target (columns), respectively. Each cell reports the detection accuracy for that reference–target pair relative to manually annotated GCP ground-truth positions.

Overall, the confusion matrices in the Fig. 4 highlight a trade-off between precision (SIFT) and robustness/coverage (LoFTR), which is critical when processing uncontrolled citizen imagery. Based on these results, LoFTR was adopted as the primary Auto-GCP detector; residual localisation error is subsequently mitigated through the HITL review step prior to camera calibration.

B. Camera Calibration and Geo-rectification

Camera calibration was applied independently to each image from the Ōnaero site using the eight TLS-referenced GCPs. Table 2 summarises the estimated parameters and their variability across the calibrated image set. The focal length estimate is highly consistent (mean 3045.11 px, SD 23.43 px). Converting to physical units (Eq. 2) yields $f \approx$ 4.23 mm, which closely matches the nominal lens specification of 4.2mm, supporting the validity of the estimated intrinsics.

Table 2. Camera calibration summary. Reported statistics (mean and standard deviation) are computed across all calibrated images using the eight GCPs.

Parameter	Mean	Standard Deviation
Focal length (pixels)	3045	23.43
Azimuth (degrees)	186.26	0.22
Tilt (degrees)	6.28	1.06
Roll (degrees)	−93.39	0.60
Reprojection RMSE (pixels)	31	8.11

Orientation parameters show that station geometry is repeatable, with low dispersion in azimuth (SD 0.22°) and roll (SD 0.60°). Tilt varies more (SD 1.06°), consistent with small differences in how user seat the smartphone in the cradle, which primarily affects the vertical aiming angle. Across all images, the mean reprojection error is 30.95 px (SD 8.11 px), corresponding to ~0.8% of the image width; an example overlay is shown in Fig. 5.

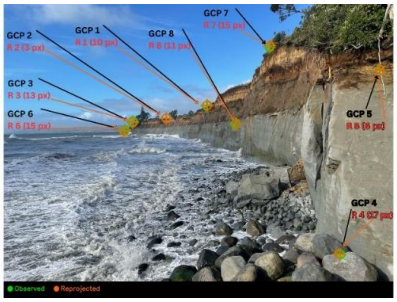

Figure 5. Example reprojection overlay for a calibrated Ōnaero image. Observed GCP locations (green markers) are compared against their model-projected positions (orange markers), with labels indicating the pixel reprojection error for each GCP.

Using the estimated camera parameters, each image was mapped onto the site-specific cliff surface model to generate an NZTM-referenced, geo-rectified cliff-view output. This spatial standardisation supports more consistent segmentation and time-separated comparison. An example result is shown in Fig. 6.

Figure 6. Example geo-rectified cliff image for the Ōnaero site, obtained by projecting the smartphone photograph onto the fitted cubic-polynomial cliff surface and registering the output to the NZTM coordinate system.

C. Segmentation

This section summarises preliminary outcomes for both RoI-constrained visual-prompt segmentation (SAM) and text-guided semantic segmentation (LangSAM).

- *Visual-prompt segmentation (SAM, RoI-based):* Bounding-box prompting was used to constrain SAM to a cliff-focused region of interest. Figure 7 illustrates a typical outcome: the RoI is defined by a rectangular prompt, and SAM returns candidate masks within that region, from which the highest-scoring mask was selected. We found that this approach reliably delineated the cliff as a whole, but it often merged adjacent objects into a single mask. A recurrent failure mode was the inability to separate the cliff face from overlying vegetation, resulting in a combined cliff–tree segment (Fig. 7), which limits interpretability for feature-specific tracking. Moreover, it was not able to segment the cliff into separate cliff-elements.

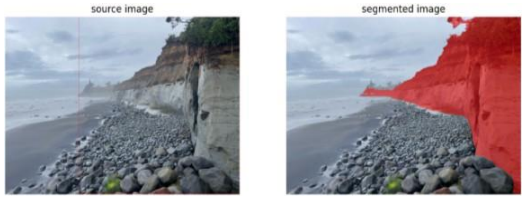

Figure 7. Example of RoI-constrained SAM segmentation.

- *Semantic prompt–based segmentation (LangSAM).* The text-guided segmentation using LangSAM (GroundingDINO + SAM) performed better. Across the Ōnaero test images, semantic prompting produced more interpretable class-wise masks than visual-prompt segmentation. Performance depended on both scene conditions and configuration. We assessed sensitivity to (i) text prompts (e.g., cliff, debris, vegetation), (ii) varying environmental conditions, (iii) cliff viewing geometry (direction/illumination), and (iv) the detection/box threshold.

Table 3 compares representative outputs for the prompts: cliff, debris and vegetation under thresholds of 0.25 and 0.40. The results indicate that the optimal threshold is image- and prompt class-dependent: increasing the threshold can reduce false detections in cluttered scenes but may also suppress weaker targets, particularly for smaller or low-contrast features like “fractures” or “cracks”.

Table 3. Representative LangSAM outputs for the text prompts cliff, debris and vegetation under two detection thresholds (0.25 and 0.40). The final row illustrates typical failure cases for small-scale, low-contrast features (e.g., joints/cracks), which are more sensitive to thresholding and image conditions.

Text Prompts	Threshold	
	0.25	0.40
Cliff		
Debris		
Vegetation		
Fractures, Joints, Cracks		

Finally, we evaluated semantic segmentation on geo-rectified images. Figure 8a shows that, in favourable cases, geo-rectification supports clean separation among major classes (cliff, sky/sea and vegetation). However, Figure 8b highlights a consistent limitation: segmentation may fail to separate the cliff toe from wave foam when colours and textures blend, generating ambiguity along the waterline.

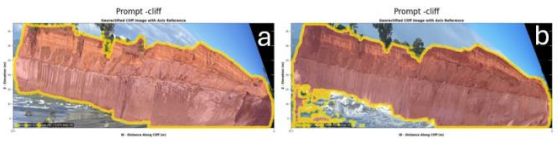

Figure 8. Semantic segmentation on geo-rectified images. (a) Successful separation of major classes under favourable conditions; (b) failure case where marine spray and/or sea–cliff texture blending degrades segmentation at the cliff toe.

D. Tracking debris change

To explore the feasibility of time-separated change screening using text-prompted segmentation, debris masks were extracted for two survey dates using the "debris" prompt and then compared to highlight spatial differences in debris extent at the cliff toe (Fig. 9). In this proceeding we focus on debris/talus as a first target feature because it provides a high-contrast, spatially coherent signal that is comparatively robust to citizen-image variability and directly interpretable in terms of cliff-toe deposition and reworking/removal. Extending the same framework to quantify change on the cliff face (e.g., fresh scars or surface retreat) is feasible but requires additional controls for illumination/wetness and more stringent masking, and is therefore left for ongoing work.

Figure 9 presents a representative example (Aug 2024 vs June 2025). While the results are preliminary and may be influenced by environmental variability (e.g., tidal stage, wetness and illumination), the approach demonstrates the potential to generate rapid, visually interpretable indicators of debris change from repeat citizen photographs. Some misclassification remains (for example, areas flagged as "removed" debris may partly reflect acquisition artefacts rather than true loss) however the output is promising and provides a basis for refinement. Ongoing work will quantify sensitivity to acquisition conditions and validate image-derived change estimates against independent measurements (e.g., TLS-derived surface change) where available.

Figure 9. Tracking debris changes from text-prompted segmentation. (a) Aug 2024; (b) June 2025; (c) debris change map: red = newly detected debris, green = persistent/common debris, blue = removed debris.

IV. Discussion

This study supports the feasibility of extending CoastSnap-style citizen imagery from beaches to coastal cliff change screening, while highlighting constraints for operational use. Auto-GCP benchmarking shows a clear trade-off between precision and robustness: SIFT localises points accurately when successful but degrades under diffuse illumination, whereas LoFTR achieves higher overall detection and greater stability across lighting conditions, making it better suited to uncontrolled citizen imagery - particularly when paired with a lightweight HITL verification step to correct residual outliers.

Camera calibration results indicate that intrinsic and extrinsic parameters estimated from repeat smartphone images are broadly consistent, with focal length estimates agreeing with device specifications and low variability in azimuth/roll consistent with a fixed cradle. Although reprojection RMSE remains significant, surface-based geo-rectification provides practical spatial standardisation that supports downstream segmentation and time-separated comparison.

Segmentation results show that RoI-only visual prompting can isolate the cliff sector but often merges adjacent objects (e.g., vegetation and cliff face). Text-guided segmentation (LangSAM) produces more interpretable masks for cliff and debris/talus, while fine, low-contrast features (joints/cracks) remain challenging. Geo-rectification generally improves consistency, though sea–cliff blending at the toe remains a failure mode. Preliminary debris change detection shows that citizen photos can provide time-separated, interpretable change cues, but results are sensitive to environmental variability (e.g., marine spray) and require independent validation.

V. Conclusion

We present a preliminary workflow to adapt CoastSnap citizen imagery for coastal cliff monitoring, integrating ML-assisted GCP matching, camera calibration with reprojection-based geo-rectification, and prompted segmentation for feature extraction and time-separated debris-change screening. LoFTR provided the most robust Auto-GCP transfer under variable conditions (with HITL QA to manage localisation scatter), calibration produced stable camera parameters consistent with device specifications, and semantic prompt–based segmentation outperformed RoI-only approaches for cliff and debris delineation.

Key limitations include sensitivity to illumination and environmental variability, persistent sea–cliff ambiguity at the toe, and weak performance for thin, low-contrast features (joints/cracks). Apparent debris change can also be confounded by tides and wetness.

VI. Next steps

Next steps will prioritise independent validation and scalability testing. First, we will validate geo-rectification performance and image-derived debris-change proxies against terrestrial laser scanning (TLS) measurements where available. Second, we will strengthen robustness to acquisition variability by implementing stricter quality gates (e.g., blur/glare screening and sea-spray filtering) and improved masking (e.g., sky/sea exclusion and cliff-toe ambiguity handling), alongside adaptive prompting and threshold selection.

To assess transferability beyond the controlled pilot setting, we will test whether comparable results can be achieved under multi-user, multi-device conditions (i.e., different smartphones and placement variability) and across additional stations and acquisition modes (fixed stations and mobile/free-roaming images). Finally, we will extend change screening beyond debris to additional cliff-relevant indicators, including fresh erosion sources/scars, fracture-like features, and cliffline recession. With these developments, the approach could complement episodic surveys by providing time-separated evidence to support cliff monitoring programmes and, ultimately, contribute to early-warning decision frameworks.

REFERENCES

[1] M. D Harley and Kinsela, M. A. “CoastSnap: A global citizen science program to monitor changing coastlines”. Continental Shelf Research, vol. 245. https://doi.org/10.1016/j.csr.2022.104796, 2022

[2] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, Nov 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: an efficient alternative to sift or surf,” 11 2011, pp. 2564–2571.

[4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2024. [Online]. Available: https://arxiv.org/abs/2304.07193

[5] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, “Loftr: Detector- free local feature matching with transformers,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8918–8927.

[6] M. D. Harley, M. A. Kinsela, E. Sanchez-Garcia, and K. Vos, “Shoreline change mapping using crowd-sourced smartphone images,” Coastal Engineering, vol. 150, pp. 175–189, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378383918304551

[7] T. F. Coleman and Y. Li, “An interior trust region approach for nonlinear minimization subject to bounds,” SIAM Journal on Optimization, vol. 6, no. 2, pp. 418–445, 1996. [Online]. Available: https://doi.org/10.1137/0806023

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, “Segment anything,” arXiv:2304.02643, 2023.

[9] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu et al., “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” arXiv preprint arXiv:2303.05499, 2023.