A Progressive Three-Stage Learning Framework for Underwater Visual Pose Estimation

Tianxiang Cao, Zheyu Liu, Ying Tan, and Ye Pu
Faculty of Engineering and Information Technology (FEIT)
University of Melbourne, Melbourne, Australia
{caot2, zheyul}@student.unimelb.edu.au, {yingt, ye.pu}@unimelb.edu.au

***Abstract*—Underwater visual pose estimation remains challenging due to severe image degradation and the scarcity of labeled underwater data. Building on our recent two-stage framework, we propose a three-stage learning framework that exploits increasingly realistic supervision to bridge the domain gap between in-air and underwater images. The framework learns from synthetic underwater images generated from labeled in-air datasets, adapts to real underwater images through self-supervised learning, and is refined using pseudo-labels from pretrained 3D foundation models. Experiments on Aqualoc-based pairwise pose estimation and SeaThru-NeRF dense mapping demonstrate improved pose, trajectory, and reconstruction accuracy over 3D foundation models while maintaining a lightweight, real-time model.**

***Index Terms*—Underwater pose estimation, transfer learning**

I. Introduction

Underwater visual pose estimation is important for robotic navigation, inspection, and mapping [1]. Although learning-based correspondence methods such as SuperPoint [2] and MicKey [3] achieve strong performance in terrestrial environments, their performance degrades underwater due to image degradation caused by absorption, scattering, turbidity, and illumination changes [4], together with the scarcity of labeled underwater pose datasets [1]. Existing underwater methods mainly improve feature robustness through image enhancement or appearance-based learning [5], while adapting geometry-aware pose estimation to underwater environments remains unexplored.

Building on our recent two-stage learning framework MK-UW [19], which learns from synthetic labeled data and adapts to real underwater images through self-supervised learning, this paper introduces a third learning stage that exploits pseudo-labels generated by pretrained 3D foundation models. The resulting progressive three-stage learning framework combines supervised, self-supervised, and pseudo-supervised learning to exploit increasingly realistic supervision without requiring manually labeled underwater poses. Our contributions are: (i) a pseudo-label refinement stage that provides multi-view geometric supervision using pretrained 3D foundation models; (ii) an evaluation of representative foundation models as pseudo-label generators; and (iii) experiments on Aqualoc-based pairwise pose estimation and SeaThru-NeRF dense mapping, showing improved pose, trajectory, and reconstruction accuracy with a real-time, lightweight model compared with 3D foundation models.

II. Method

Our framework extends the two-stage framework [19] with a third learning stage using pseudo-labels from pretrained 3D foundation models (Fig. 1), providing multi-view geometric supervision without manually labelled underwater poses.

A. Base Two-Stage Framework (Stages I–II)

Stages I and II follow our previous framework [19]. Stage I performs supervised transfer learning on synthetic underwater images generated from labelled in-air RGB-D data, while Stage II adapts the model to real underwater images through self-supervised learning. Although these stages improve robustness to underwater appearance, they lack real 3D geometric supervision, motivating Stage III.

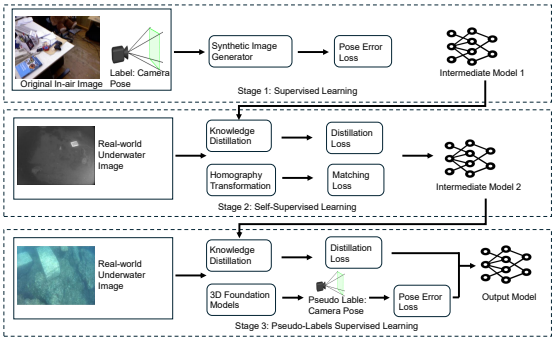

Fig. 1. Three-stage learning framework: (I) supervised transfer, (II) self-supervised adaptation, and (III) pseudo-supervised refinement.

B. Stage III: Pseudo-label Refinement

Stage II adapts the model to real underwater images through homography consistency, but it cannot capture the true three-dimensional viewpoint changes encountered in underwater environments. To address this limitation, Stage III introduces pseudo-supervision using pretrained 3D foundation models. Specifically, a pretrained 3D foundation model generates pseudo-labels from real underwater image pairs, providing multi-view geometric supervision without requiring manually labelled underwater poses.

Generating pseudo-labels. Given a real underwater image pair (I, I'), a pretrained 3D foundation model predicts a pseudo ground-truth relative pose $\tilde{T} = (\tilde{\mathbf{R}}, \tilde{\mathbf{t}})$ from the estimated scene geometry. Pseudo-labels are generated offline and incur no additional training or inference cost.

Selecting reliable pseudo-labels. Foundation models may produce unreliable pseudo-labels for degraded underwater images. We therefore assign each pseudo-label a reliability weight $w \in [0,1]$ based on the generator confidence c and a geometric self-consistency check. Specifically, the average reprojection error is defined in [3]:

$$e_{\text{geo}} = \frac{1}{|\Omega_s|} \sum_{\mathbf{x}\in\Omega_s} \left\| \pi\Big(\tilde{T}\,\tilde{D}(\mathbf{x})\,K^{-1}\dot{\mathbf{x}}\Big) - \mathbf{m}(\mathbf{x}) \right\|_2, \tag{1}$$

where K is the camera intrinsic matrix, which is non-singular. The notion of $\pi(\cdot)$ denotes perspective projection, $\dot{\mathbf{x}}$ is the homogeneous pixel coordinate, and $\mathbf{m}(\mathbf{x})$ is the corresponding point in I'. The reliability weight is

$$w = c\,\mathbb{1}[e_{\text{geo}} < \tau]\,\exp(-e_{\text{geo}}/\sigma), \tag{2}$$

which rejects unreliable pseudo-labels and smoothly down-weights the remaining samples. Here$\mathbb{1}[\cdot]$ is the indicator function, which equals 1 when $e_{\text{geo}} < \tau$ and 0 otherwise, τ is the reprojection-error threshold, and σ is a scale parameter controlling the exponential decay of the reliability weight.

Learning from pseudo-labels. The model is refined by minimizing the following loss

$$\mathcal{L}_{\text{stage3}} = \mathbb{E}_{J\sim P(J)}\Big[w\,\ell_{\text{pose}}\big(h(J),\tilde{T}\big)\Big] + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{3}$$

where ℓ_{pose} is the pose loss introduced in Stage I and $\mathcal{L}_{\text{reg}}$ is the regularization term used to preserve the Stage II representation [19]. Stage III is performed after Stage II so that refinement operates on features already adapted to real underwater images. To preserve the descriptor quality learned in Stage II, the final objective becomes $\mathcal{L} = \lambda_{\text{ps}}\mathcal{L}_{\text{stage3}} + \mathcal{L}_{\text{stage2}}$, where $\mathcal{L}_{\text{stage2}}$ is defined in [19] and λ_{ps} is a positive constant.

III. Experimental Results

We evaluate MK-UW on an underwater visual pose estimation benchmark from the Aqualoc dataset [7]. Performance is measured using TranE, RotE, RepE, PrecP, PrecV, AucP, and AucV (defined in [19]).

A. Effect of the Pseudo-label Generator

We compare three 3D foundation models—VGGT [13], MASt3R [20], and DUSt3R [21]—as pseudo-label generators while keeping Stages I–II fixed. Table I compares their performance with the two-stage baseline (S1+S2).

TABLE I
EFFECT OF THE STAGE III PSEUDO-LABEL GENERATOR ON THE UNDERWATER MAPFREE BENCHMARK. S1+S2 IS THE TWO-STAGE BASE MK-UW; THE OTHER THREE ARE THE FULL THREE-STAGE MK-UW.

Generator $\mathcal{G}$	TranE↓	RotE↓	RepE↓	PrecP↑	AucP↑	PrecV↑	AucV↑
S1+S2	1.35	15.06	142.65	13.92	29.30	37.58	57.92
DUSt3R [21]	1.37	17.40	146.40	11.11	22.70	31.05	47.97
MASt3R [20]	1.37	14.28	153.36	13.19	26.53	35.72	56.57
VGGT [13]	1.35	14.79	133.90	15.76	30.07	37.14	59.16

B. Downstream Application: Dense Mapping

We further evaluate MK-UW as the tracking front-end of an online dense mapping system on the SeaThru-NeRF dataset [15]. MK-UW estimates relative camera poses, while Depth Anything V2 [14] predicts per-frame depth and TSDF fusion [18] incrementally reconstructs the scene. We report relative pose error (RPE-R, RPE-t), absolute trajectory error (ATE), absolute rotation error (ARE), and reconstruction completeness and accuracy (Comp and C@2%), using the COLMAP reconstruction [16] as reference. As offline references, we additionally compare against VGGT and MASt3R.

TABLE II
DENSE MAPPING RESULTS ON THE CURAÇAO SEATHRU-NERF SEQUENCE. ONLINE METHODS ARE COMPARED WITH OFFLINE REFERENCES. BEST ONLINE RESULTS ARE SHOWN IN BOLD.

	Method	RPE-R° ↓	RPE-t [m]↓	ATE [m]↓	ARE° ↓	Comp %↓	C@2%↑
Online	MicKey	5.09	0.062	0.071	37.5	4.69	4.3
	MK-UW(S1+S2)	1.29	0.046	0.059	29.2	4.08	18.0
	MK-UW$_{\text{MASt3R}}$	**1.09**	0.039	**0.052**	14.9	1.92	71.4
	MK-UW$_{\text{VGGT}}$	1.10	**0.038**	0.053	**11.8**	**1.47**	**79.7**
Offline	VGGT	0.42	0.004	0.007	3.1	0.88	97.5
	MASt3R	1.40	0.045	0.098	16.0	1.95	53.3

Quantitative Table II reports trajectory and mapping results. Among the online methods, performance improves consistently from MicKey to the two-stage framework and further with Stage III refinement. Mapping accuracy (C@2%) increases from 4.3% to 18.0% and then to 71–80%, while trajectory and completeness errors decrease accordingly. The VGGT-distilled model achieves the best performance among the online methods and outperforms the offline MASt3R baseline across all metrics. Although offline VGGT achieves the highest accuracy, it reconstructs the entire sequence jointly and thus serves as an upper bound rather than an online method. Its superior geometry also makes it the most effective pseudo-label generator for Stage III refinement.

Qualitative. Fig. 2 shows the two-stage variants of MK-UW (Stage I and Stage I+II) present clearer geometric features but exhibit a scale inconsistency that distorts the map; adding the Stage III pseudo-label refinement distilled from MASt3R or VGGT largely removes this scale drift. The VGGT-distilled model produces the best reconstruction, while the MASt3R-distilled model introduces spurious blue points in near-field regions, showing that the model conflates the strongly blue water background with foreground structure during training.

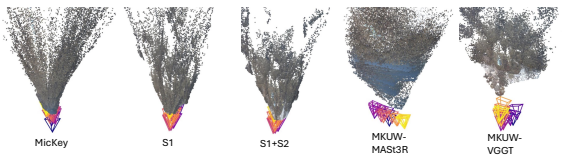

Fig. 2. Qualitative dense reconstruction on the Curaçao SeaThru-NeRF sequence. Stage III removes the scale inconsistency of the earlier models, with the VGGT-distilled model producing the best reconstruction. The MASt3R-distilled model introduces spurious near-field blue points.

REFERENCES

[1] R. M. Eustice, H. Singh, and J. J. Leonard, "Visually navigating the RMS Titanic with SLAM information filters," in *Robotics: Science and Systems*, 2008.

[2] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF CVPR Workshops*, 2018, pp. 224–236.

[3] A. Barroso-Laguna, S. Munukutla, V. Prisacariu, and E. Brachmann, "Matching 2D images in 3D: Metric relative pose from metric correspondences," in *Proc. IEEE/CVF CVPR*, 2024.

[4] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proc. IEEE/CVF CVPR*, 2019, pp. 1682–1691.

[5] J. Yang et al., "Knowledge-distillation-based underwater feature extraction network for visual pose estimation," *IEEE J. Ocean. Eng.*, 2025.

[6] E. Arnold, J. Wynn, S. Vicente, G. Garcia-Hernando, A. Monszpart, V. Prisacariu, D. Turmukhambetov, and E. Brachmann, "Map-free visual relocalization: Metric pose relative to a single image," in *Proc. ECCV*, 2022, pp. 690–708.

[7] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, "AQUALOC: An underwater dataset for visual-inertial-pressure localization," *Int. J. Robot. Res.*, vol. 38, no. 14, pp. 1549–1559, 2019.

[8] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proc. IEEE/CVF CVPR*, 2018, pp. 6723–6732.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE ICCV*, 2011, pp. 2564–2571.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, 2006, pp. 404–417.

[12] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024.

[13] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual geometry grounded transformer," in *Proc. IEEE/CVF CVPR*, 2025.

[14] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," in *Proc. NeurIPS*, 2024.

[15] D. Levy, A. Peleg, N. Pearl, D. Rosenbaum, D. Akkaynak, S. Korman, and T. Treibitz, "SeaThru-NeRF: Neural radiance fields in scattering media," in *Proc. IEEE/CVF CVPR*, 2023, pp. 56–65.

[16] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE/CVF CVPR*, 2016, pp. 4104–4113.

[17] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *Vision Algorithms: Theory and Practice*, 2000, pp. 298–372.

[18] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proc. ACM SIGGRAPH*, 1996, pp. 303–312.

[19] T. Cao, Y. Tan, and Y. Pu, "MK-UW: A two-stage transfer learning framework for underwater metric-keypoint pose estimation," unpublished manuscript, 2025. [Online]. Available: https://tianxiangcao.github.io/Papers/MK_UW.pdf.

[20] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3D with MASt3R," in *Proc. ECCV*, 2024.

[21] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUSt3R: Geometric 3D vision made easy," in *Proc. IEEE/CVF CVPR*, 2024, pp. 20697–20709.