HOSS-BENCH: OPEN-SET CROSS-MODAL VESSEL RE-IDENTIFICATION BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Vessel re-identification aims to match a query image of a vessel against a gallery of candidate images to determine if the same vessel appears. Cross-modal RGB–SAR vessel re-identification is particularly challenging due to large appearance differences across sensing modalities. The Hybrid Optical and SAR Ship Re-Identification Dataset (HOSS) was recently released alongside a baseline method. In this work, we extend the evaluation protocol of HOSS into a realistic benchmark by (a) proposing an improved training/validation/test split; (b) introducing *open-set* evaluation, where the query image may depict an unknown vessel; (c) measuring re-identification performance in the most challenging cross-modal setting where the query vessel appears in the gallery only in a different modality, while the gallery contains multiple similar vessels in the same modality as the query. We formulate vessel re-identification as an open-set recognition problem and use open-set accuracy as the primary metric, alongside standard closed-set ReID metrics. We re-train the TransOSS model according to the new protocol and compare it with a DINOv3-based baseline, highlighting that current performance remains insufficient for realistic online open-set vessel clustering applications.

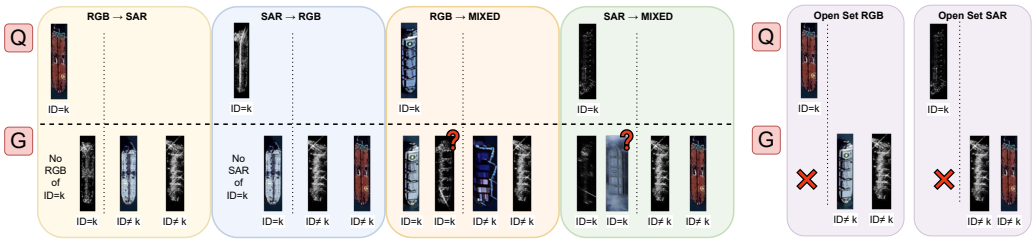

Figure 1: Illustration of HOSS-Bench evaluation modes. Queries (Q, top) are matched against a shared multi-modal gallery (G, bottom). Images with ID=k denote true positives for the query identity, while ID$\neq$ k indicate non-matching vessels. Red crosses indicate open-set queries with no corresponding gallery match. Question marks (?) denote *optional* positives, which may be present for some queries and absent for others. Mixed-modality settings allow positives from multiple modalities, whereas cross-modal settings restrict positives to the opposite modality.

1 INTRODUCTION

Maritime monitoring is essential for detecting illegal activities such as unreported fishing and smuggling. While the Automatic Identification System (AIS) is widely used for vessel tracking, a large fraction of vessels operate without reliable AIS signals due to regulatory exemptions or deliberate deactivation. As a result, satellite imagery has become a critical complementary observation source (Paolo et al., 2024). While vessel detection is well studied, RGB–SAR vessel re-identification remains challenging due to drastic cross-modal appearance differences. In this work, we introduce HOSS-Bench, a benchmark for open-set cross-modal vessel re-identification based on the HOSS dataset (Wang et al., 2025). We propose a balanced dataset split with a shared multi-modal gallery and formulate vessel re-identification as an open-set recognition problem. Using this benchmark, we

analyze baseline performance across multiple query modes and highlight the difficulty of realistic open-set RGB–SAR vessel re-identification.

2 Related Work and Limitations

The HOSS dataset (Wang et al., 2025) introduced the first dataset for RGB–SAR vessel re-identification; however, its evaluation protocol exhibits several limitations. It relies on modality-separated galleries, whereas real-world deployments typically involve mixed-modality galleries. Gallery sizes also differ substantially across evaluation modes, making RGB-to-SAR and SAR-to-RGB results not directly comparable. Moreover, the protocol assumes a closed-set formulation in which every query has a corresponding gallery match, and does not model open-set scenarios. Finally, no explicit validation split is provided. These limitations motivate HOSS-Bench, a benchmark with a balanced dataset split and a unified open-set evaluation protocol for RGB–SAR vessel re-identification.

3 HOSS-Bench: Dataset and Evaluation Protocol

3.1 Balanced Dataset Split

We propose a new balanced split of the HOSS dataset designed to support realistic and comparable RGB–SAR vessel re-identification. Unlike the original evaluation protocol, which relies on modality-separated and imbalanced query–gallery configurations, our split uses a *shared multi-modal gallery* across all evaluation modes and explicitly includes open-set queries.

Table 1: Dataset statistics for HOSS-Bench. Query counts are reported as *test / validation*. All evaluation modes share the same multi-modal gallery.

	Query (test / val)				**Gallery**	**Train**
Split	RGB→SAR	SAR→RGB	RGB→MIXED	SAR→MIXED		
IDs	27 / 20	26 / 21	36 / 21	36 / 21	64 / 51	385
Images	48 / 28	47 / 29	47 / 29	47 / 29	306 / 59	1236

We define four query modes. In **RGB→SAR**, queries are RGB images; some queries have only SAR positives in the gallery, while others have no corresponding gallery match. The **SAR→RGB** mode is defined analogously with SAR queries and RGB positives. In the mixed-modality settings, **RGB→MIXED** queries are RGB images that may have at least one RGB positive in the gallery and may additionally have SAR positives, while some queries have no corresponding gallery match; **SAR→MIXED** queries are defined analogously for SAR queries. We additionally report results on **ALL**, the union of all query modes. In all settings, the gallery additionally contains distractor images that do not correspond to any query identity.

To ensure fair comparison across modes, we balance the number of vessel identities and images in the query sets while keeping the gallery fixed. Dataset statistics for all evaluation modes are summarized in Table 1. The gallery is shared across modes and contains both RGB and SAR images.

3.2 Open-Set Evaluation Protocol

Our primary focus is *open-set* RGB–SAR vessel re-identification, where a query vessel may not appear in the gallery. For each query image, the system determines whether a matching vessel exists in the gallery. As illustrated in Figure 2, this decision is based on the minimum distance between the query embedding and all gallery embeddings in the latent space. For a given query q, we compute distances to all gallery images and identify the minimum distance d. If d exceeds a threshold θ, the query is rejected; otherwise, it is accepted and assigned the identity of the nearest gallery image. The threshold θ is selected on the validation set to maximize open-set accuracy and is fixed at test time.

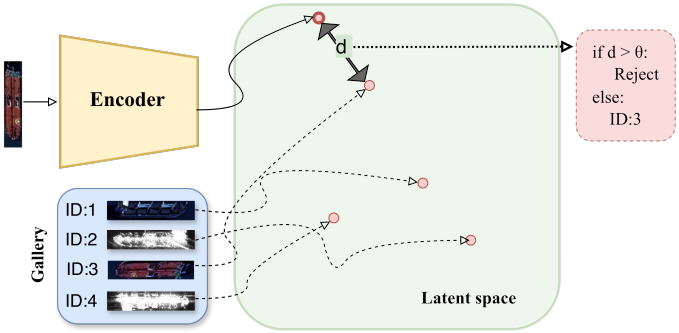

Figure 2: Overview of the operation of the baseline methods on HOSS-Bench. A query image is encoded into a latent embedding and compared against a shared multi-modal gallery. If the minimum distance to all gallery embeddings exceeds a threshold θ, the query is rejected as having no match; otherwise, the identity of the nearest gallery image is assigned.

We evaluate open-set performance using **open-set accuracy**, defined as

$$\text{acc} = \frac{\text{correct matches} + \text{correct rejections}}{\text{total number of queries}}.$$

In addition to open-set evaluation, we report standard closed-set re-identification metrics on the subset of queries that have at least one positive match in the gallery. Specifically, we compute mean Average Precision (Zheng et al., 2015), mean Inverse Negative Penalty (mINP) (Ye et al., 2022), and Rank-k accuracy (CMC) (Wang et al., 2007), which measure retrieval quality and ranking performance when a correct match is guaranteed to exist.

4 Experiments

4.1 Experimental Setup

Models and Training Details. We evaluate TransOSS (Wang et al., 2025) and DINOv3 (Siméoni et al., 2025), a strong vision foundation model. TransOSS is initialized from the official pretrained weights. To ensure a fair comparison, DINOv3 is continually pretrained using the same CLIP-style cross-modal strategy and OptiSar data as TransOSS, and both models are then fine-tuned on the training split of HOSS-Bench. All models are trained using the AdamW optimizer with a learning rate of 5×10^{-4} and batch size 32. We train DINOv3 for 200 epochs and TransOSS for 300 epochs.

Model Selection and Thresholding. During training, we jointly select the best model checkpoint and its corresponding rejection threshold θ using the validation set. At each epoch, for every validation query, we compute the minimum distance to the gallery. These distances define a set of candidate thresholds. For each candidate threshold, we evaluate validation open-set accuracy as defined above. We retain the checkpoint that achieves the highest validation accuracy together with its optimal threshold. The resulting (checkpoint, θ) pair is fixed and used for test-time inference.

4.2 Results and Analysis

Table 2 reports open-set and closed-set re-identification performance. DINOv3 outperforms TransOSS across all metrics, particularly in mixed-modality settings. The large gap between cross-modal and mixed-modality modes highlights the difficulty of realistic scenarios where the gallery is multi-modal but the true match may exist only in a different modality. Figure 3 shows that both models struggle to distinguish known from unknown vessels, and that TransOSS additionally suffers from frequent identity misclassification among accepted queries.

Online Open-Set Clustering. To approximate real-world deployment where galleries are not fixed and grow over time, we simulate an online open-set clustering scenario. Vessel observations are

Table 2: Evaluation results on HOSS-Bench. Accuracy reports open-set performance. mAP, mINP, and Rank-k are computed on the closed-set subset.

Metric	Model	RGB→SAR	SAR→RGB	RGB→MIXED	SAR→MIXED	All
Accuracy	TransOSS	45.8	36.2	61.7	48.9	48.1
	DINOv3	45.8	40.4	89.4	70.2	61.4
mAP	TransOSS	5.7	15.8	42.4	42.3	26.3
	DINOv3	25.6	12.5	64.8	61.5	40.9
mINP	TransOSS	5.5	13.9	12.3	34.9	16.6
	DINOv3	25.1	11.6	25.7	49.4	27.9
Rank-1	TransOSS	0.0	4.0	64.0	36.0	25.7
	DINOv3	19.2	4.0	96.0	68.0	46.5
Rank-10	TransOSS	19.2	40.0	92.0	76.0	56.4
	DINOv3	46.2	32.0	100.0	84.0	65.3

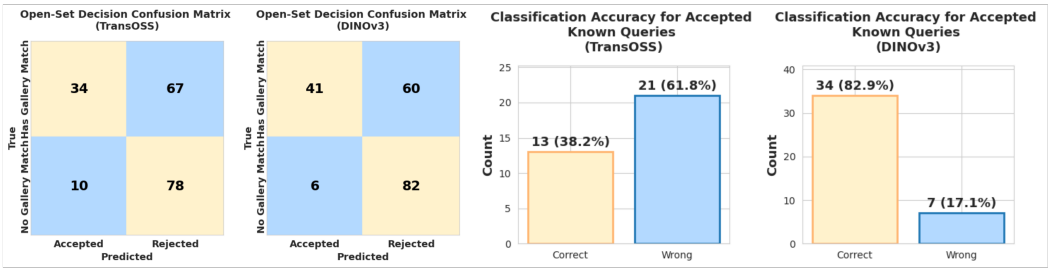

Figure 3: Open-set decision analysis. Left: confusion matrices for open-set accept/reject decisions on TransOSS and DINOv3. Right: identity classification accuracy for correctly accepted known queries. Results highlight the difficulty of separating known and unknown vessels.

processed sequentially, starting from an empty gallery, and each embedding is either assigned to the nearest existing cluster if its distance is below the open-set threshold θ, or used to initialize a new cluster otherwise. Using ground-truth vessel identities, we evaluate clustering quality with ARI (Hubert & Arabie, 1985), NMI (Strehl & Ghosh, 2002), homogeneity, and completeness (Rosenberg & Hirschberg, 2007). Table 3 shows that both models significantly over-estimate the true number of vessel identities, producing substantially more clusters than the ground truth, while DINOv3 consistently outperforms TransOSS across all clustering metrics.

Table 3: Online open-set clustering performance on HOSS-Bench. All results are reported as mean $\pm$ std over 10 random seeds. The ground-truth number of clusters is 85.

Model	# Discovered Clusters	ARI ↑	NMI ↑	Homogeneity ↑	Completeness ↑
TransOSS	214.0 $\pm$ 1.5	0.227 $\pm$ 0.006	0.861 $\pm$ 0.002	0.938 $\pm$ 0.004	0.796 $\pm$ 0.001
DINOv3	182.1 $\pm$ 2.0	0.437 $\pm$ 0.014	0.901 $\pm$ 0.002	0.960 $\pm$ 0.004	0.848 $\pm$ 0.003

5 Conclusion

We introduced HOSS-Bench, a benchmark for open-set cross-modal vessel re-identification that addresses key limitations of prior closed-set and modality-restricted evaluations. HOSS-Bench features a unified open-set evaluation protocol with a shared multi-modal gallery, and includes challenging cross-modal query modes that reflect realistic operational conditions. Experimental results with DINOv3 and TransOSS reveal substantial performance gaps between cross-modal and mixed-modality settings, as well as significant challenges in both open-set recognition and online identity discovery. We hope HOSS-Bench will motivate the community to develop more robust methods for open-set object re-identification under realistic conditions.

REFERENCES

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Fernando Paolo, David Kroodsma, Jennifer Raynor, Tim Hochberg, Pete Davis, Jesse Cleary, Luca Marsaglia, Sara Orofino, Christian Thomas, and Patrick Halpin. Satellite mapping reveals extensive industrial activity at sea. *Nature*, 625:85–91, 01 2024. doi: 10.1038/s41586-023-06825-8.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Jason Eisner (ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/D07-1043/.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

Han Wang, Shengyang Li, Jian Yang, Yuxuan Liu, Yixuan Lv, and Zhuang Zhou. Cross-modal ship re-identification via optical and sar imagery: A novel dataset and method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7873–7883, October 2025.

Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(06):2872–2893, June 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3054775. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3054775.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, Dec 2015. doi: 10.1109/ICCV.2015.133.