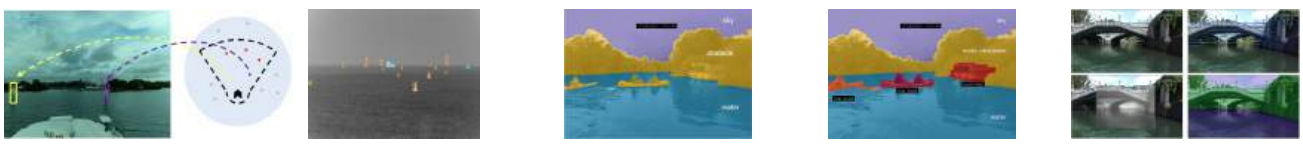
4th Workshop on Maritime Computer Vision (MaCVi): Challenge Overview

Benjamin Kiefer[1], Jan Lukas Augustin[2,3], Jon Muhovič[4], Mingi Jeong[5], Arnold Wiliem[6,7], Janez Pers[4], Matej Kristan[4], Alberto Quattrini Li[8], Matija Teršek[9], Josip Šarić[10], Arpita Vats[11], Dominik Hildebrand[12], Rafia Rahim[12], Mahmut Karaaslan[13], Ersin Kaya[13], Akib Mashrur[6], Tze-Hsiang Tang[14], Chun-Ming Tsai[15], Jun-Wei Hsieh[16], Ming-Ching Chang[17,18], Wonwoo Jo[19], Doyeon Lee[20], Yusi Cao[21], Lingling Li[21], Vinayak Nageli[22], Arshad Jamal[23], Gorthi Rama Krishna Sai Subrahmanyam[22], Jemo Maeng[24], Seongju Lee[24], Kyoobin Lee[24], Xu Liu[21], LiCheng Jiao[21], Jannik Sheikh[25], Martin Weinmann[26], Ivan Martinović[10], Jose Mateus Raitz Persch[27], Rahul Harsha Cheppally[27], Mehmet E. Belviranli[28], Dimitris Gahtidis[6], Hyewon Chun[19], Sangmun Lee[19], Philipp Gorczak[3], Hansol Kim[19], Jeeyeon Jeon[19], Borja Carrillo Perez[29], Jiahui Wang[21], Sangmin Park[24], Andreas Michel[25], Jannick Kuester[25], Bettina Felten[25], Wolfgang Gross[25], Yuan Feng[30], Justin Davis[28]
[1]LOOKOUT, [2]Helmut Schmidt University, [3]catskill GmbH, [4]University of Ljubljana, [5]Virginia Tech, [6]Shield AI, [7]Queensland University of Technology, [8]Dartmouth College, [9]Luxonis, [10]Faculty of Electrical Engineering and Computing, University of Zagreb, [11]LinkedIn, [12]University of Tuebingen, [13]Konya Technical University, [14]Schneider Electric Taiwan Co., Ltd., [15]University of Taipei, [16]National Yang Ming Chiao Tung University, [17]University at Albany, SUNY, [18]Inventec Corporation, [19]HD Korea Shipbuilding & Offshore Engineering Co., Ltd., [20]Seoul National University, [21]Xidian University, [22]Indian Institute of Technology, Tirupati, [23]Centre for Artificial Intelligence and Robotics (CAIR), Bangalore, India, [24]Gwangju Institute of Science and Technology, [25]Fraunhofer IOSB, [26]Karlsruhe Institute of Technology, [27]Kansas State University, [28]Colorado School of Mines, [29]Arquimea Research Center, [30]Independent Researcher

Abstract

The 4th Workshop on Maritime Computer Vision (MaCVi) is organized as part of CVPR 2026. This edition features five benchmark challenges with emphasis on both predictive accuracy and embedded real-time feasibility. This report summarizes the MaCVi 2026 challenge setup, evaluation protocols, datasets, and benchmark tracks, and presents quantitative results, qualitative comparisons, and cross-challenge analyses of emerging method trends. We also include technical reports from top-performing teams to highlight practical design choices and lessons learned across the benchmark suite. Datasets, leaderboards, and challenge resources are available at

https://macvi.org/workshop/cvpr26.



(a) Vision-to-Chart (b) Thermal Detection (c) Embedded Seg. (d) LaRS Panoptic (e) Multi-modal

Figure 1. Overview of MaCVi @ CVPR 2026 challenges, including Vision-to-Chart Data Association, Thermal Object Detection Maritime Collision Avoidance Dataset, Embedded Semantic Segmentation, LaRS Panoptic Segmentation, and the Multi-modal Challenge Multiaqua.

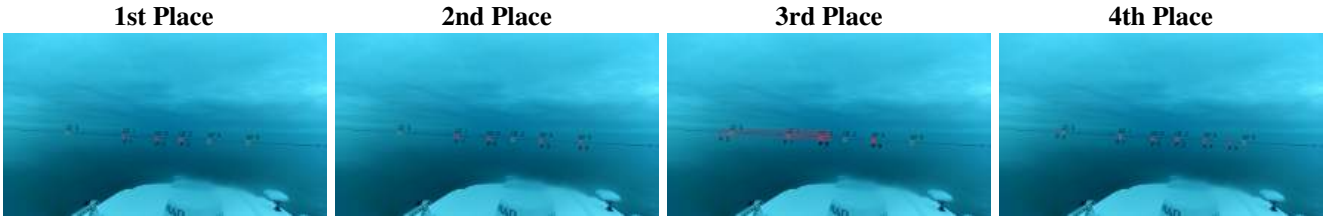


Figure 2. **Qualitative detection examples** from the Vision-to-Chart Data Association challenge. Each panel shows the overlay of predicted and ground-truth bounding boxes and predicted index of associated chart marker.

Table 1. Results for the Vision-to-Chart Data Association challenge on the held out test split. Ranking medals indicate final placement. The baseline (gray) is the organizer-provided DETR-style fusion transformer. Method names are derived from the submitted technical reports.

Place	Method	Institution	P ↑	R ↑	F1 ↑	mIoU ↑	Overall ↑
①	Skyline-aware ROI-calibrated association [§A.1]	HD Korea Shipbuilding	**0.9248**	0.6934	0.7926	**0.7365**	**0.7646**
②	QueryMLP projection + DETR [§A.2]	Arquimea Research Center	0.8563	**0.7604**	**0.8055**	0.6718	0.7386
③	Dynamic chart-derived DEIMv2 [§A.3]	Xidian University team	0.4107	0.3827	0.3962	0.3705	0.3834
-	Fusion transformer (baseline)	MaCVi Organizers	0.3142	0.2925	0.3029	0.3636	0.3333
4th	IMU-conditioned query DETR [§A.4]	IIT Tirupati / CAIR-DRDO	0.2170	0.2107	0.2138	0.2499	0.2318

1. Introduction

Maritime environments, with their unique challenges such as dynamic lighting, reflections, and cluttered scenes, demand specialized computer vision techniques [9, 17, 19, 21, 42, 49]. Autonomous systems like Unmanned Surface Vehicles (USVs) rely heavily on robust vision algorithms to navigate, detect, and interpret complex surroundings [2, 22, 58]. Addressing these challenges requires not only cutting-edge algorithms but also standardized benchmarks and a collaborative research ecosystem [1, 3, 8, 27].

The 4th Workshop on Maritime Computer Vision (MaCVi 2026) builds on the momentum of previous iterations [20, 23, 24] and marks this year's workshop at CVPR 2026. The 2026 challenge suite includes Vision-to-Chart Data Association, Thermal Object Detection, LaRS Panoptic Segmentation, Embedded Semantic Segmentation, and Multimodal Semantic Segmentation.

This draft report summarizes the current challenge configuration for MaCVi 2026 and serves as the starting point for iterative updates. The rest of the paper is structured as follows: Section 2 details the general challenge protocols, while Section 3 outlines the challenge tracks. All datasets, evaluation tools, and leaderboards are available at https://macvi.org/workshop/cvpr26.

2. Challenge Participation Protocol

As in previous challenge iterations, all tracks followed a shared participation and evaluation protocol, with task-specific details provided on the individual challenge pages. Rules, datasets, evaluation code, and starter kits were distributed through the workshop website, and participants submitted predictions through the official evaluation server in the required task-specific formats, including ONNX export where applicable. Submission frequency was limited to one to three entries per day depending on the track. The evaluation server handled automated scoring and public leaderboards for most tracks, with the challenge timeline spanning December 29, 2025 to March 15, 2026 (AoE). After final verification, teams were ranked by the official track metrics and compliance requirements, and top teams were invited to contribute short technical reports for this overview paper.

3. MaCVi 2026 Challenge Tracks

MaCVi @ CVPR 2026 features five benchmark challenges. The 2026 edition puts explicit emphasis on both model quality and embedded real-time constraints.

3.1. Vision-to-Chart Data Association Challenge

This challenge studies association between visible maritime navigational aids and chart markers from a monocular RGB image. Given an image and a set of chart queries, methods must detect visible buoys and assign them to the correct markers under clutter, reflections, and long-range viewpoints. The dataset contains 4,285 training, 904 validation, and 924 test samples; the training and validation splits are public, while the test split is private.

Ranking follows [26] and uses buoy-detection Precision/Recall/F1, matched-box mIoU, and an overall score defined as the arithmetic mean of F1 and mIoU. We use the real-time organizer baseline from [26]. Submissions were limited to 250M parameters and were evaluated by the organizers for reproducibility and private-test performance.

Table 1 and Figure 2 summarize the final results. All top-

Table 2. Results for the Thermal Object Detection challenge. Ranking medals indicate final placement. The baseline (gray) is the organizer-provided Faster R-CNN with ResNet-50. Self-reported inference speed and hardware are included. Backbone and detection-level pretraining datasets are reported separately (IN = ImageNet-1K, IN-22K = ImageNet-22K, LVD = Large Vision Dataset).

Place	Method	Institution	AP ↑	AP_{50}	AP_{75}	AR_1	AR_{10}	FPS	Hardware	Backbone pretr.	Det. pretr.
1	Multi-arch ensemble + SSL [§B.1]	Schneider Electric Taiwan	**0.4868**	**0.8268**	**0.4709**	**0.3043**	**0.5937**	∼0.01	A100 40GB	IN-22K, LVD-142M	COCO, O365
2	DEIMv2 ensemble [§B.2]	U. Taipei / NYCU / UAlbany	0.4709	0.8218	0.4409	0.2987	0.5763	∼0.1	RTX 3090	LVD-1689M	COCO
3	AGAF [§B.3]	HD Korea Shipbuilding	0.4685	0.8067	0.4628	0.3033	0.5850	∼1.5	RTX 5070	LVD-1689M	COCO
-	Faster R-CNN (baseline)	MaCVi Organizers	0.3137	0.6313	0.2769	0.2417	0.4167	∼32	A30	IN	COCO

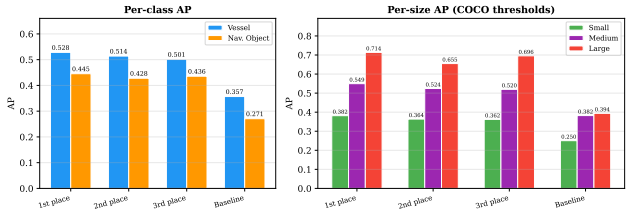


Figure 3. **Per-class and per-size AP breakdown** for the Thermal Object Detection challenge. Left: AP for *vessel* vs. *navigational object* classes. Right: AP at COCO size thresholds (small / medium / large). Vessel detection is consistently easier than navigational object detection across all methods. Small object AP remains the main bottleneck.

ranked methods outperform the organizer baseline, with the first two methods showing a substantial margin. Technical descriptions of all submissions are provided in Section A.

Across submissions, the strongest trend is the use of explicit geometric priors. Top methods project chart information into image space, define query-specific search regions, or inject chart- and IMU-derived cues into the detector. The winning method combines skyline estimation, ROI projection, assignment, and calibration in a staged pipeline; the second-place method uses a learned world-to-image projection as a compact spatial prior; and the third-place method shows that a stronger visual backbone and decoder can further improve robustness. Overall, the results suggest that physically grounded geometric conditioning is especially effective for vision-to-chart association, and that stronger detection backbones remain complementary to these priors.

3.2. Thermal Object Detection Challenge

Night-time and low-visibility conditions are critical operational scenarios for unmanned surface vehicles, yet conventional RGB-based perception systems perform poorly in such conditions. Thermal infrared imaging offers a modality that is largely invariant to ambient illumination, making it a natural complement for maritime operations. This challenge targets obstacle and vessel detection in thermal imagery.

The challenge uses the Maritime Collision Avoidance Dataset [12], which contains annotated electro-optical imagery captured from German, British, and Dutch waters between 2023 and 2025. For the MaCVi 2026 edition, the dataset is organized into two object classes (*vessel* and *navigational object*) with COCO-format bounding box annotations. The dataset comprises 704 train, 173 val, and 381 test images with hidden labels. An NVIDIA RTX 5080 GPU, sponsored by catskill GmbH, was offered as a prize for the top-performing team.

3.2.1. Evaluation Protocol

Detection performance is evaluated using the standard COCO object detection metrics. The primary ranking metric is Average Precision (AP), computed as AP averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05. AP_{50} serves as a tiebreaker. Additional reported metrics include AP_{75}, AR_1, and AR_{10}.

Participants submit COCO-style JSON files containing bounding boxes, class labels, and confidence scores. A maximum of three submissions per day is allowed during the challenge period. All participants are required to report inference speed (in FPS) and the hardware used for benchmarking. As baseline, we provide a Faster R-CNN with a ResNet-50 backbone pretrained on COCO.

3.2.2. Submissions, Analysis and Trends

Table 2 summarizes the overall ranking, Figure 3 reports the class and size breakdown, Figure 4 shows representative detections, and Figure 5 plots the precision–recall curves at IoU = 0.50.

Ensemble dominance. All three podium methods rely on multi-model ensembles fused via Weighted Boxes Fusion (WBF) [48]. The winning method combines 6 models spanning 5 distinct architectures (Co-DINO [56], DDQDETR [55], DINO [54], RTMDet [37], and RF-DETR [46]), the 2nd place method ensembles 11 DEIMv2 detectors trained at 4 different resolutions with different random seeds, and the 3rd place method fuses 9 RF-DETR 2XLarge sources via class-aware WBF with per-class weights for vessels and navigational objects. The complementary strategies architecture diversity, optimization diversity, and class-aware fusion all prove effective for improving recall and localization stability.

Semi-supervised learning. The 1st place method employs a three-phase training pipeline that includes pseudo-labeling on the 381 unlabeled test images followed by MixPL [4] teacher-student training. This semi-supervised component

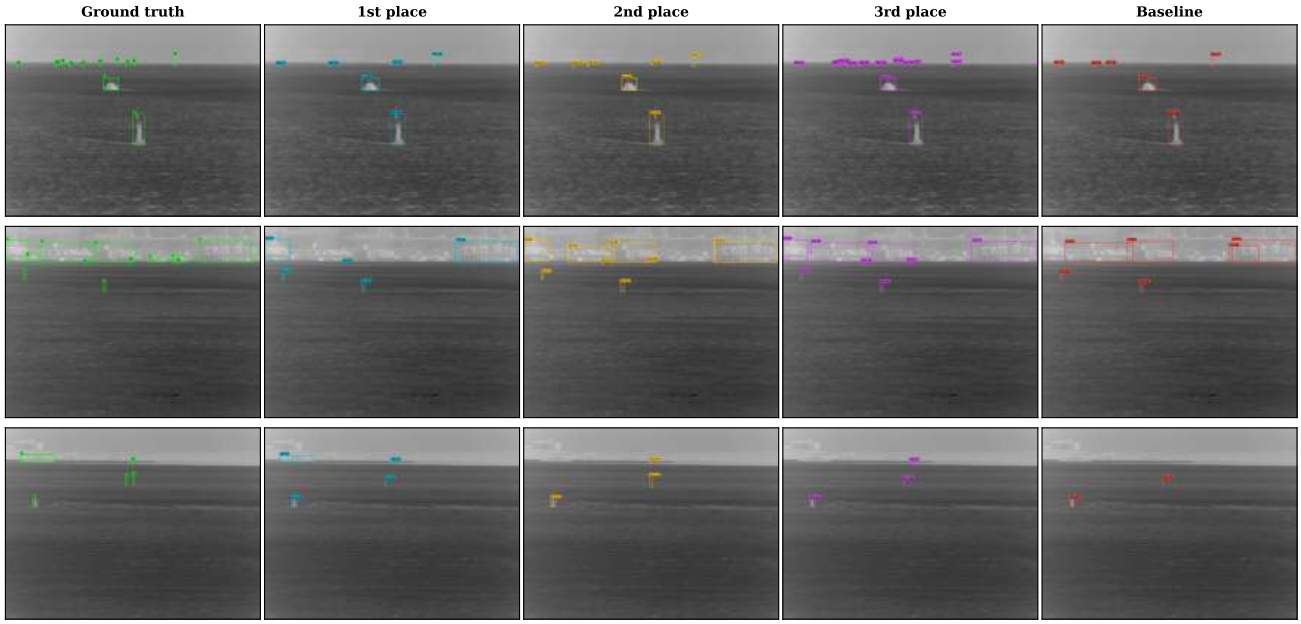


Figure 4. **Qualitative detection examples** from the Thermal Object Detection challenge. Each row shows one test image with ground truth (green) and predictions from the top three methods and baseline. Columns left to right: ground truth, 1st place, 2nd place, 3rd place, baseline. Predictions are filtered at confidence ≥ 0.3. V = vessel, N = navigational object.

demonstrates that even modest amounts of unlabeled data can meaningfully improve performance when labeled training sets are small (704 images).

Resolution and small objects. A domain-specific challenge is that 63.4% of annotated objects are smaller than 32×32 pixels (COCO "small" category), with navigational objects having a median size of only 8.3×18.6 pixels. The technical reports indicate that higher inference resolution is the single largest lever for improving detection of these small objects. The 2nd place method explores resolutions from 1024×1024 up to 1600×1600, assigning higher WBF weights to models operating at larger scales.

Contrast enhancement. Both top methods apply CLAHE (Contrast Limited Adaptive Histogram Equalization) to improve feature contrast in thermal imagery. Notably, the 1st place report observes that CLAHE benefits strong models but slightly hurts weaker ones (-0.002 to -0.008 AP), suggesting that it amplifies existing model capacity rather than providing a universal boost.

Annotation quality and domain-specific filtering. The 3rd place method demonstrates that carefully curating the training labels can rival more complex learning strategies. Fixing inconsistent wind-turbine annotations in the provided dataset yielded the single largest gain of any individual technique across all submissions (+0.027 AP). Additionally, a YOLO-based [16] horizon detector that suppresses navigational-object false positives above the estimated sky-

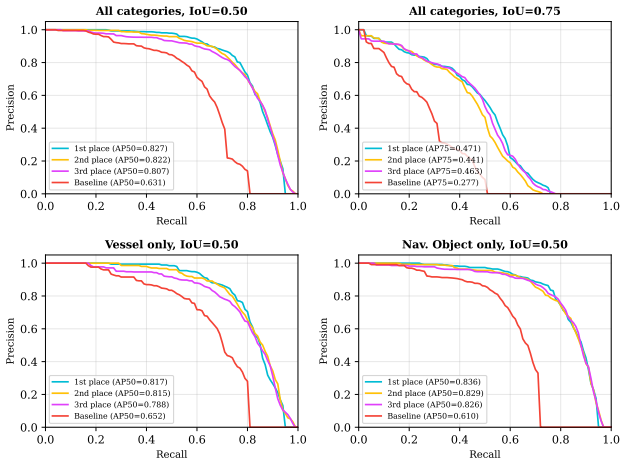


Figure 5. **Precision–Recall curves** for the Thermal Object Detection challenge. Curves are shown for each method at IoU=0.50.

line contributed a further +0.012 AP. These domain-specific adaptations—annotation refinement and geometric filtering—are complementary to the ensemble and semi-supervised strategies employed by the other top teams.

3.2.3. Discussion and Challenge Winners

The winners of the Thermal Object Detection challenge are listed in Table 2. All three winning methods leverage large-scale ensembles (6–11 models) as their core strategy,

Table 3. Results for the USV-based Panoptic Segmentation Challenge measured in overall panoptic quallity (PQ) and separate for thing and stuff classes. Competing methods are compared to a baseline Mask2Former model [5] and the top-performing PanSR method [59].

Place	Team	Method	Section	PQ	SQ	RQ	PQ_{st}	PQ_{th}
★	MaCVi Team	PanSR (Swin-L)	-	57.3	75.9	66.9	95.4	43.0
1	FER Zagreb	M2F-DINOv3	§C.1	**53.5**	**74.3**	**63.3**	**95.8**	**37.7**
2	Fraunhofer IOSB - BIG	MaskDINOv3	§C.2	48.3	72.4	57.7	94.9	30.9
3	KTSFOE	ThingSeg-DGCR	§C.3	42.6	71.0	51.1	91.8	24.2
4th	PanopticMaritimeScenery	M2F-Mot3-SwinL-FPN	-	42.2	73.5	48.7	95.3	22.3
5th	NTNU	RF-DETR	-	40.3	69.9	48.0	93.3	20.4
6th	Dt deep-square	NewSR	-	36.0	73.1	40.8	94.3	14.2
-	MaCVi Team	M2F (Swin-B)	-	41.4	75.2	47.1	92.5	22.3

confirming that ensemble-based fusion remains a dominant approach for detection challenges with limited training data. The 1st place method additionally demonstrates that semi-supervised learning provides meaningful gains when labeled data is scarce, while the 3rd place method shows that annotation quality and domain-specific filtering (§B.3) can be competitive alternatives to semi-supervised learning. Across all submissions, small object detection remains the central difficulty: the extreme size distribution of thermal maritime objects (median navigational object area $\approx$154 px^2) pushes current detector architectures to their limits. Detailed method descriptions are provided in §B.1, §B.2, and §B.3.

3.3. LaRS Panoptic Segmentation Challenge

This challenge continues the LaRS panoptic benchmark and requires parsing USV-view scenes into *stuff* classes and *thing* instances. The *stuff* classes include water, sky, and static obstacles, while the *thing* classes include eight types of dynamic obstacles: boat/ship, buoy, row boat, swimmer, animal, paddle board, float and other. As in previous editions, strong performance requires balancing semantic consistency, instance-level detection, and robustness to small obstacles.

3.3.1. Evaluation Protocol

Submitted methods are ranked based on panoptic quality (PQ) averaged over 11 classes and evaluated on 1,203 test images from the LaRS dataset. Different from standard evaluation, dynamic obstacle detections inside static obstacle regions are not additionally penalized as false positives [60]. For training, participants are allowed to use the LaRS training and validation sets, which consist of 2,605 and 198 labeled images respectively, as well as other datasets if disclosed. To provide additional insights, we factorize the panoptic quality into recognition quality (RQ) and segmentation quality (SQ) [25], as well as separately evaluate stuff (PQ_{st}) and thing (PQ_{th}) classes.

3.3.2. Submissions, Analysis and Trends

The challenge received 26 submissions from 6 different teams. Table 3 shows the results for the top submission from each team, along with the performance of the state-of-the-art method PanSR [59] (out of competition) and our Mask2Former baseline with Swin-B backbone. Results from the remaining submissions are available on the public leaderboard on the MaCVi website. Further, we analyze the results with a focus on the top three performing teams, for which detailed technical reports are provided in the appendix C.

We observe that four of the six submitted methods outperform our baseline, while none surpass PanSR, the state-of-the-art method for maritime panoptic segmentation, which achieves 57.3 PQ on the LaRS test set. The top-performing submission, M2F-DINOv3, comes close with 53.5 PQ and even exceeds PanSR on stuff classes (PQ_{st}). However, the lower performance on thing classes underscores the importance of dynamic obstacle detection in maritime scenes. The second-ranked submission, MaskDINOv3, achieves 48.3 PQ, trailing the winner by 5.2 PQ points, while the third-ranked ThingSeg-DGCR reaches 42.6 PQ, trailing by 10.9 PQ points. Performance on thing classes appears to be the main factor differentiating the top three methods as well, with MaskDINOv3 and ThingSeg-DGCR falling behind M2F-DINOv3 by roughly 7 and 13 PQ_{th} points, respectively.

Performance by scene attributes. To further understand these differences, we now examine stratified performance according to different scene attributes including environment, illumination, reflections, and conditions. Figure 6 shows that the global ranking is preserved across most of the scenarios. Interestingly, the top two submissions outperform the SOTA method PanSR in certain challenging scenarios, such as night illumination and heavy reflections. The winning M2F-DINOv3 further surpasses PanSR in overexposed scenes, as well as under rain, fog, and in the presence of plants or debris. These results indicate that large-scale backbone pretraining can enhance model robustness in difficult scenarios.

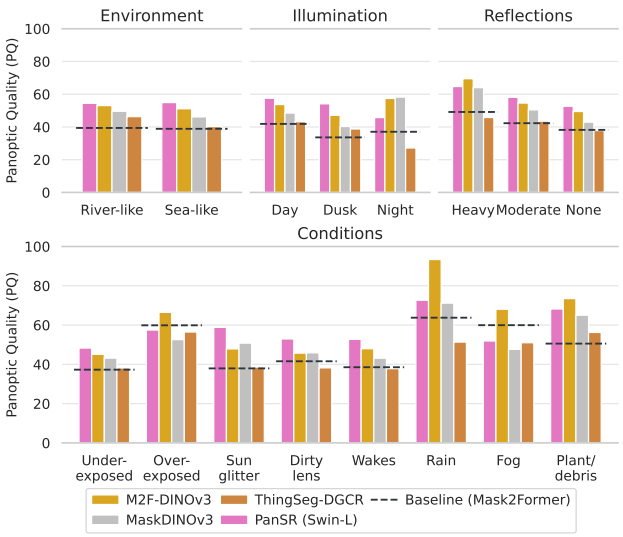


Figure 6. Panoptic quality stratified according to scene attributes.

Detection rate by obstacle size. Figure 7 shows the detection rates of dynamic obstacles as a function of object size. It also shows the normalized object frequency for each size bin (blue). Detection rates are lowest for the smallest objects (first bin), which also contains the largest number of instances. This is where the gap between the submitted methods and the SOTA PanSR is largest, highlighting the impact of PanSR's object-centric proposal head and training scheme, that addresses the oversegmentation and mask drifting in inference typical for Mask-Dino-based models, and improves small object detection and dense scenarios.

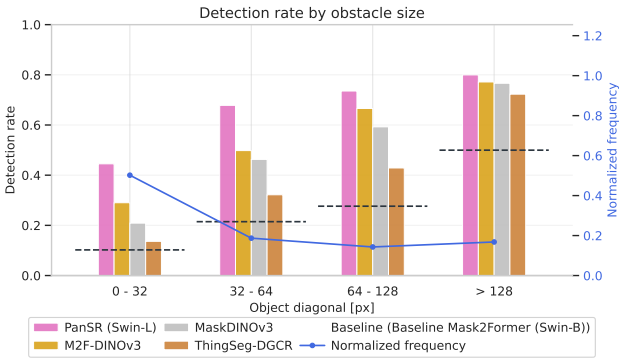


Figure 7. Dynamic obstacle detection rate stratified according to object size.

Qualitative results. Figure 8 presents qualitative results on three LaRS test scenes for the top-performing methods. The largest differences appear in the third scene, which contains many small objects. Even PanSR fails to detect all objects labeled in the ground truth, and the detection rate is lower for the other methods. Interestingly, MaskDINOv3 detects more objects than M2F-DINOv3 but struggles to segment the sea, highlighting differences in how Mask2Former and MaskDINO generate mask proposals.

3.3.3. Discussion and Challenge Winners

The winners of the LaRS Panoptic Segmentation challenge are listed in Table 3. All three methods outperform our baseline and show notable improvements in both stuff and thing segmentation. The performance of the winning method is comparable to the winner of the previous challenge, yet it still falls short of our baseline specialized for maritime panoptic segmentation. The top two methods highlight the strong trend of using mask-transformer meta-architectures with large-scale, self-supervised backbones. While self-supervised pretraining improves robustness under challenging conditions such as rain and fog, detecting small objects in crowded scenes remains challenging. Further gains may come from combining large-scale self-supervised pretraining [47] with targeted small-object detection strategies [59].

3.4. Embedded Semantic Segmentation Challenge

Recent maritime obstacle segmentation methods often rely on expensive, power-hungry hardware, making them unsuitable for small, energy-constrained USVs. Building on previous editions [23, 24], this challenge targets methods that balance segmentation quality and real-time efficiency and evaluates them on a real-world Luxonis embedded device OAK4 [36].

3.4.1. Evaluation Protocol

The Embedded Semantic Segmentation Challenge follows the LaRS evaluation protocol [60] with additional deployment constraints for embedded execution. Methods predict per-pixel labels for water, sky, and obstacles and are evaluated with navigation-oriented rather than conventional segmentation metrics.

Following LaRS, static obstacle detection is measured by water-edge accuracy (μ), while dynamic obstacle detection is summarized by F1. To balance dynamic obstacle detection and segmentation quality, the primary ranking metric is the combined quality score

$$Q = \mathrm{F1} \cdot \mathrm{mIoU}. \tag{1}$$

Because submissions must run on the target embedded platform, models must be exportable to ONNX with a static computation graph, use only supported operations, accept a single 768×384 image normalized with ImageNet [7] statistics, and achieve at least 30 FPS on the target device.

For server-side evaluation, submitted models are quantized to INT8 using the LaRS validation split, compiled for device execution, and evaluated after standard resizing and padding, with outputs resized back to the original resolution before scoring. Final rankings are computed with the LaRS

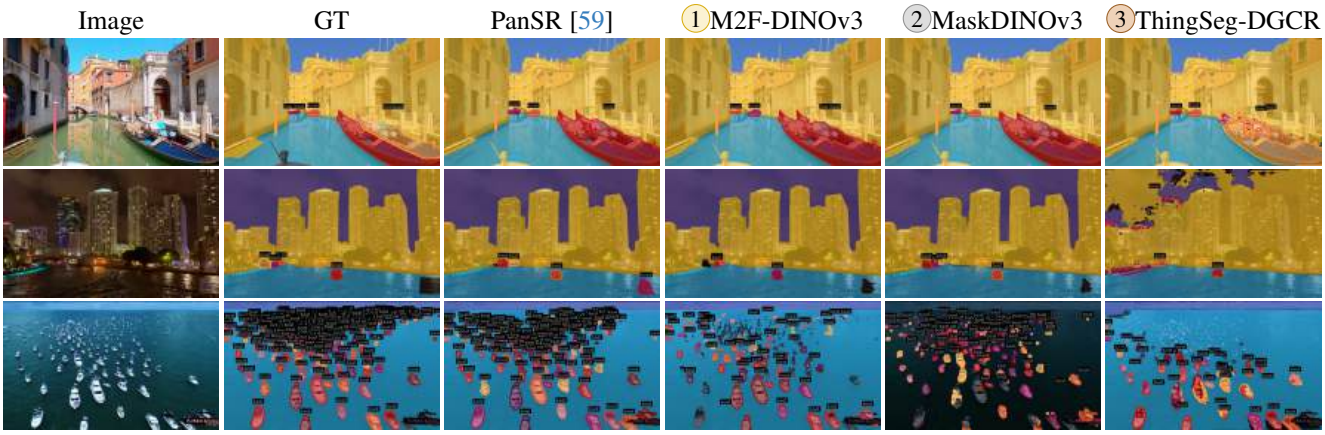


Figure 8. Qualitative results of the top-performing methods in the LaRS panoptic segmentation challenge.

Table 4. Overview of the submissions for the USV-based Embedded Obstacle Segmentation challenge. For comparison, last year's winning method is included inline in gray and is not assigned placements. The best results among current non-gray entries are denoted in bold.

Place	Institution	Method	Section	FPS	**Q↓**	μ	Pr	Re	F1	mIoU
$*$	DLMU	RSOS-Net	[24]	85.1	64.2	72.5	66.4	67.9	67.1	95.7
1	Independent Researcher	DSOS-Net	§D.1	66.5	**61.9**	**68.3**	**62.6**	69.7	**66.0**	93.8
2	Independent / K-State	PIDNet-S	§D.2	67.6	52.1	67.8	54.1	56.4	55.2	**94.4**
3	Colorado School of Mines	RSOS_R50	§D.3	**96.3**	46.4	61.5	41.6	67.0	51.3	90.3
-	Colorado School of Mines	YXHR_medium	§D.3	64.9	45.3	64.3	42.6	64.3	51.3	88.4
5th	DT Deep-Square	newlight	-	130.1	39.4	56.3	32.1	**75.8**	45.1	87.3
6th	UCLA	EdgeModel	-	32.4	18.5	54.9	13.9	59.9	22.5	82.3

metrics above, and average embedded throughput is reported alongside accuracy.

3.4.2. Submissions, Analysis and Trends

We received 20 submissions from 5 teams. Final rankings assign one placement per team using the best-ranked submission when a team enters multiple models. Table 4 also includes last year's winning method [24] for comparison, but that entry is not part of the ranking.

The top three methods highlight complementary strategies for embedded maritime segmentation. DSOS-Net (§D.1) leads overall by combining a DINOv3-pretrained ConvNeXt backbone [33, 47] with a lightweight RSOS-Net-style decoder [50] and a two-stage training schedule. PIDNet-S (§D.2) emphasizes quantization-friendly real-time deployment and Copy-Paste obstacle augmentation for small and rare obstacles. RSOS_R50 (§D.3) revisits RSOS-Net in PyTorch with a ResNet-50 backbone and lightweight multi-scale context aggregation.

Compared to last year's reference model, the top-ranked methods do not surpass the best Q or F1 scores, but they remain above the embedded speed threshold. Among this year's ranked entries, DSOS-Net achieves the best quality score and F1, PIDNet-S the best mIoU, and RSOS_R50 the highest throughput of the top three, highlighting the trade-off between quality and throughput.

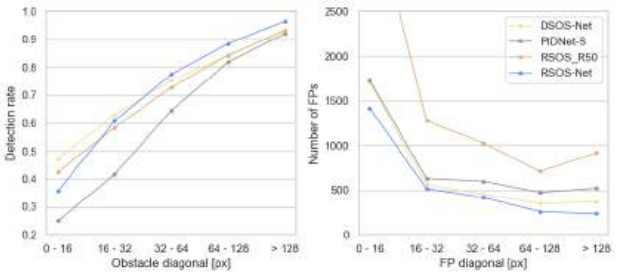


Figure 9. Detection rate and FP as a function of obstacle size for embedded segmentation methods. Only the three winning and the best-performing method from last year's challenge are shown.

Across obstacle sizes in Figure 9, DSOS-Net is strongest on the smallest obstacles, with RSOS_R50 also competitive, whereas last year's RSOS-Net remains strongest on medium and large obstacles. RSOS-Net and DSOS-Net also produce the fewest false positives, while RSOS_R50 generates more small-object false alarms. Under scene conditions and environments (Figures 10 and 11), last year's RSOS-Net remains

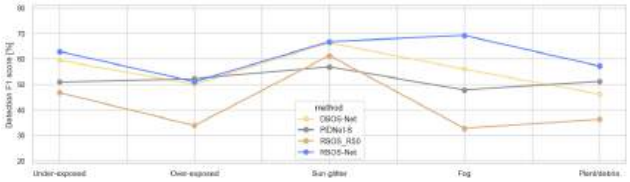


Figure 10. Detection F1 under challenging visual conditions for embedded segmentation methods. Only the three winning and the best-performing method from last year's challenge are shown.

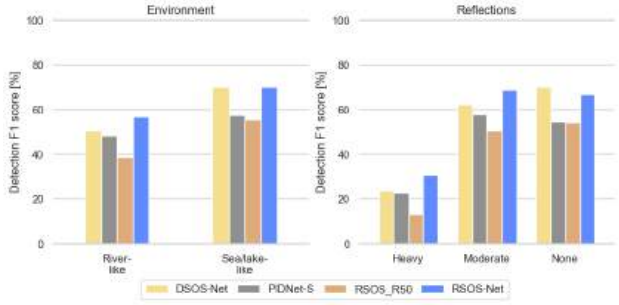


Figure 11. Detection F1 across scene environments and reflection levels for embedded segmentation methods showing the three winning and the best-performing method from last year's challenge.

strongest in many difficult settings, DSOS-Net is the best current entry overall and strongest when reflections are absent, and PIDNet-S is generally more stable than RSOS_R50 but less accurate. Overall, DSOS-Net is the strongest ranked method this year, while last year's RSOS-Net remains a strong reference.

3.4.3. Discussion and Challenge Winners

Table 4 lists the winners of the Embedded Semantic Segmentation challenge. DSOS-Net leads overall, PIDNet-S prioritizes quantization-friendly real-time deployment, and RSOS_R50 remains competitive when throughput is prioritized.

3.5. Multimodal Semantic Segmentation Challenge

This track studies segmentation with synchronized RGB, thermal, and LiDAR inputs on MULTIAQUA. Methods predict four classes—*static obstacle*, *dynamic obstacle*, *water*, and *sky*—and must remain reliable when one or more modalities are degraded or missing.

3.5.1. Evaluation Protocol

Participants submitted predictions for the labeled validation/daytime split and the hidden test/nighttime split. The main challenge is two-fold: methods must exploit auxiliary modalities such as thermal and LiDAR while still relying on RGB when it is informative. Ranking is based on M, the mean of the two mIoU scores; per-split mIoU and dynamic-obstacle IoU are also reported to capture nighttime robustness and safety-critical obstacle detection.

3.5.2. Submissions, Analysis and Trends

The challenge received 36 submissions from 3 teams, with only each team's best entry used for the final ranking. Technical reports for the top-performing methods are provided in Appendix E. Two of the submitted methods use strong pretrained backbones from the SAM family, giving them a strong start with the RGB input, while potentially also helping with interpreting other modalities. As shown by the winner, GatedMemorySAM, explicitly modeling difficult visibility conditions is a necessary part of training robust multimodal models.

3.5.3. Discussion and Challenge Winners

The winners of the Multimodal Semantic Segmentation challenge are listed in Table 5. GatedMemorySAM performs best overall and even outperforms the MaCVi reference on the validation split. The other two methods remain competitive on daytime data but degrade markedly at night, showing that robust cross-modal fusion under degraded sensing remains difficult.

4. Conclusion

This summary presents the MaCVi 2026 challenges and we see a trend towards maritime perception methods that transfer across tasks, sensing conditions, and deployment constraints. Future iterations will introduce more generalist challenges.

Acknowledgments. We thank all participating teams, contributors, and organizers supporting the MaCVi initiative. We also thank catskill GmbH for sponsoring the RTX 5080 GPU prize, Luxonis for sponsoring the embedded challenge prize, and LOOKOUT for data support.

Table 5. Results for the USV-based Multimodal Segmentation Challenge. We report M, validation/test mIoU, and dynamic-obstacle IoU.

Place	Team	Method	Section	M	mIoU (*val*)	IoU_{do} (*val*)	mIoU (*test*)	IoU_{do} (*test*)
★	MaCVi Team	SWIN-B + Mask2Former	-	83.4	90.8	70.1	76.0	37.0
1	GIST AI LAB	GatedMemorySAM	§E.1	**82.1**	**93.3**	**78.8**	**70.9**	**35.4**
2	Xidian University	Adapted MFNet-SAM-LoRA	§E.2	64.2	90.9	70.3	37.6	7.4
3	Fraunhofer IOSB	Modified DustNet	§E.3	62.8	87.9	60.7	37.6	13.5

References

[1] Domenico Bloisi. Background modeling and foreground detection for maritime video surveillance. *Chapter in Handbook on Background Modeling and Foreground Detection for Video Surveillance: Traditional and Recent Approaches, Implementations, Benchmarking and Evaluation, Chapman and Hall/CRC*, pages 14–1, 2014. 2

[2] Borja Bovcon and Matej Kristan. WaSR–A Water Segmentation and Refinement Maritime Obstacle Detection Network. *IEEE Transactions on Cybernetics*, pages 1–14, 2021. 2

[3] Borja Bovcon, Jon Muhovič, Janez Perš, and Matej Kristan. The mastr1325 dataset for training deep usv obstacle detection models. In *Int. Conf. Intell. Robots and Systems*, pages 3431–3438. IEEE, 2019. 2

[4] Zeming Chen, Wenwei Zhang, Xinjiang Wang, Kai Chen, and Zhi Wang. Mixed pseudo labels for semi-supervised object detection. *arXiv preprint arXiv:2312.07006*, 2023. 3, 10

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5, 11

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 14

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6, 16

[8] Chaitra Desai, Sujay Benur, Ujwala Patil, and Uma Mudenagudi. Rsuigm: Realistic synthetic underwater image generation with image formation model. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(1), Dec. 2024. 2

[9] Antonio-Javier Gallego, Antonio Pertusa, Pablo Gil, and Robert B Fisher. Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras. *Journal of Field Robotics*, 36(4):782–796, 2019. 2

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 14

[11] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2918–2928, 2021. 14

[12] Philipp Gorczak, Thomas Lübcke, Martin Portier, and Helmut Schmid. Maritime collision avoidance dataset germany, english channel, and the netherlands. In *Journal of Physics: Conference Series*, volume 3123, page 012024. IOP Publishing, 2025. 3

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 11

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 15

[15] Shihua Huang, Yongjie Hou, Longfei Liu, Xuanlong Yu, and Xi Shen. Real-time object detection meets dinov3. *arXiv preprint arXiv:2509.20787*, 2025. 12

[16] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, Jan. 2023. 4, 11

[17] Urška Kanjir, Harm Greidanus, and Krištof Oštir. Vessel detection and classification from spaceborne optical images: A literature survey. *Remote sensing of environment*, 207:1–26, 2018. 2

[18] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan De Geus. Your vit is secretly an image segmentation model. In *Proceedings of the computer vision and pattern recognition conference*, pages 25303–25313, 2025. 11

[19] Benjamin Kiefer, Dominik Hildebrand, Rafia Rahim, Mahmut Karaaslan, Michael DeFilippo, Ersin Kaya, and Andreas Zell. Real-time radar–vision association via monocular distance estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2026. 2

[20] Benjamin Kiefer, Matej Kristan, Janez Perš, Lojze Žust, Fabio Poiesi, Fabio Andrade, Alexandre Bernardino, Matthew Dawkins, Jenni Raitoharju, Yitong Quan, Adem Atmaca, Timon Höfer, Qiming Zhang, Yufei Xu, Jing Zhang, Dacheng Tao, Lars Sommer, Raphael Spraul, Hangyue Zhao, Hongpu Zhang, Yanyun Zhao, Jan Lukas Augustin, Eui-ik Jeon, Impyeong Lee, Luca Zedda, Andrea Loddo, Cecilia Di Ruberto, Sagar Verma, Siddharth Gupta, Shishir Muralidhara, Niharika Hegde, Daitao Xing, Nikolaos Evangeliou, Anthony Tzes, Vojtěch Bartl, Jakub Špaňhel, Adam Herout, Neelanjan Bhowmik, Toby P. Breckon, Shivanand Kundargi, Tejas Anvekar, Ramesh Ashok Tabib, Uma Mudenagudi, Arpita Vats, Yang Song, Delong Liu, Yonglin Li, Shuman Li, Chenhao Tan, Long Lan, Vladimir Somers, Christophe De Vleeschouwer, Alexandre Alahi, Hsiang-Wei Huang, Cheng-Yen Yang, Jenq-Neng Hwang, Pyong-Kun Kim, Kwangju Kim, Kyoungoh Lee, Shuai Jiang, Haiwen Li, Zheng Ziqiang, Tuan-Anh Vu, Hai Nguyen-Truong, Sai-Kit Yeung, Zhuang Jia, Sophia Yang, Chih-Chung Hsu, Xiu-Yu Hou, Yu-An Jhang, Simon Yang, and Mau-Tsuen Yang. 1st workshop on maritime computer vision (macvi) 2023: Challenge results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 265–302, January 2023. 2

[21] Benjamin Kiefer, David Ott, and Andreas Zell. Leveraging synthetic data in object detection on unmanned aerial vehicles. *arXiv preprint arXiv:2112.12252*, 2021. 2

[22] Benjamin Kiefer, Yitong Quan, and Andreas Zell. Approximate supervised object distance estimation on unmanned surface vehicles, 2025. 2

[23] Benjamin Kiefer, Lojze Žust, Matej Kristan, Janez Perš, Matija Teršek, Arnold Wiliem, Martin Messmer, Cheng-Yen Yang, Hsiang-Wei Huang, Zhongyu Jiang, Heng-Cheng Kuo, Jie Mei, Jenq-Neng Hwang, Daniel Stadler, Lars Sommer, Kaer Huang, Aiguo Zheng, Weitu Chong, Kanokphan Lertniphonphan, Jun Xie, Feng Chen, Jian Li, Zhepeng Wang, Luca Zedda, Andrea Loddo, Cecilia Di Ruberto, Tuan-Anh

Vu, Hai Nguyen-Truong, Tan-Sang Ha, Quan-Dung Pham, Sai-Kit Yeung, Yuan Feng, Nguyen Thanh Thien, Lixin Tian, Andreas Michel, Wolfgang Gross, Martin Weinmann, Borja Carrillo-Perez, Alexander Klein, Antje Alex, Edgardo Solano-Carrillo, Yannik Steiniger, Angel Bueno Rodriguez, Sheng-Yao Kuan, Yuan-Hao Ho, Felix Sattler, Matej Fabijanić, Magdalena Šimunec, and Nadir Kapetanović. 2nd workshop on maritime computer vision (macvi) 2024: Challenge results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 869–891, January 2024. 2, 6

[24] Benjamin Kiefer, Lojze Zust, Matej Kristan, Janez Pers, Matija Tersek, Uma Mudenagudi, Chaitra Desai, Arnold Wiliem, Marten Kreis, Nikhil Akalwadi, Zhiqiang Zhong, Zhe Zhang, Sujie Liu, Xuran Chen, Yang Yang, Matej Fabijanic, Fausto Ferreira, Seongju Lee, Shanliang Yao, Himanshu Kumar, Aurelius Marcus, Gregor Novak, Yuan Feng, Annie Cheng, Thien Nguyen, Jannik Sheikh, Josip Saric, Zhuoxiao Li, Yutang Lu, Yipeng Lin, Xiang Yang, Ching-Heng Cheng, Ali Awad, Jon Muhovič, Yitong Quan, Junseok Lee, Kyoobin Lee, Runwei Guan, Xiaoyu Huang, Yi Ni, Tzu-Yu Lin, Chia-Ming Lee, Chih-Chung Hsu, Andreas Michel, Wolfgang Gross, Nan Jiang, Fei Feng, Evan Lucas, Ashraf Saleem, Yu-Fan Lin, and Martin Weinmann. 3rd workshop on maritime computer vision (macvi) 2025: Challenge results. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 1542–1569, February 2025. 2, 6, 7, 14

[25] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 9396–9405. IEEE Computer Society, June 2019. 5

[26] Marten Kreis and Benjamin Kiefer. Real-time fusion of visual and chart data for enhanced maritime vision, 2025. 2

[27] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojíř, Roman Pflugfelder, Gustavo Fernández, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, 2016. 2

[28] Reeve Lambert, Jalil Chavez-Galaviz, Jianwen Li, and Nina Mahmoudian. ROSEBUD: A deep fluvial segmentation dataset for monocular vision-based river navigation and obstacle avoidance. *Sensors*, 22(13):4681, 2022. 14

[29] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation, Dec. 2022. 12

[30] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 11

[31] C. Liao, X. Zheng, Y. Lyu, H. Xue, Y. Cao, J. Wang, K. Yang, and X. Hu. Memorysam: Memorize modalities and semantics with segment anything model 2 for multi-modal semantic segmentation. *arXiv preprint arXiv:2503.06700*, 2025. 15

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 12, 16

[33] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 7, 13

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 16

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 12

[36] Luxonis. OAK4-D. https://docs.luxonis.com/hardware/products/OAK%204%20D. Luxonis documentation. Accessed: 2026-04-08. 6

[37] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 3, 10

[38] Andreas Michel, Martin Weinmann, Jannick Kuester, Faisal Alnasser, Tomas Gomez, Mark Falvey, Rainer Schmitz, Wolfgang Middelmann, and Stefan Hinz. Dustnet++: Deep learning-based visual regression for dust density estimation. *International Journal of Computer Vision*, 133(7):4220–4244, 2025. 16

[39] Andreas Michel, Martin Weinmann, Fabian Schenkel, Tomas Gomez, Mark Falvey, Rainer Schmitz, Wolfgang Middelmann, and Stefan Hinz. Dustnet: Attention to dust. In *DAGM German Conference on Pattern Recognition*, pages 211–226. Springer, 2023. 16

[40] Jon Muhovič and Janez Perš. MULTIAQUA: A multimodal maritime dataset and robust training strategies for multimodal semantic segmentation, 2025. 14

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 11

[42] Dilip K Prasad, Huixu Dong, Deepu Rajan, and Chai Quek. Are object detection assessment criteria ready for maritime computer vision? *IEEE Transactions on Intelligent Transportation Systems*, 21(12):5295–5304, 2019. 2

[43] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations (ICLR)*, 2024. 15

[44] N. Ravi, V. Gabber, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolber, 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. *arXiv preprint arXiv:2408.00714*, 2024. 15

[45] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 12

[46] Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: Neural architecture search for real-time detection transformers. In *ICLR*, 2026. 3, 10, 11, 12

[47] 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. 6, 7, 11, 12, 13

[48] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021. 3, 10, 11

[49] Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Seadronessee: A maritime benchmark for detecting humans in open water. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2260–2270, 2022. 2

[50] Ning Wang, Yuan Feng, Lixin Tian, and Yi Wei. Rsosnet: Real-time surface obstacle segmentation network for uncrewed waterborne vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 27(1):1052–1065, 2026. 7, 13, 14

[51] Jiacong Xu, Zixiang Xiong, and Shankar P. Bhattacharyya. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19529–19539, 2023. 13

[52] X. Xu, J. Yang, W. Shi, S. Ding, L. Luo, and J. Liu. Physaug: A physical-guided and frequency-based data augmentation for single-domain generalized object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21815–21823, 2025. 15

[53] Shanliang Yao, Runwei Guan, Zhaodong Wu, Yi Ni, Zile Huang, Ryan Wen Liu, Yong Yue, Weiping Ding, Eng Gee Lim, Hyungjoon Seo, et al. Waterscenes: A multi-task 4d radar-camera fusion dataset and benchmarks for autonomous driving on water surfaces. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 14

[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 3, 10

[55] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *CVPR*, 2023. 3, 10

[56] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *ICCV*, 2023. 3, 10

[57] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. 1994. 16

[58] Lojze Žust and Matej Kristan. Learning maritime obstacle detection from weak annotations by scaffolding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 955–964, 2022. 2

[59] Lojze Žust and Matej Kristan. PanSR: An object-centric mask transformer for panoptic segmentation. *arXiv preprint arXiv:2412.10589*, 2024. 5, 6, 7

[60] Lojze Žust, Janez Perš, and Matej Kristan. LaRS: A diverse panoptic maritime obstacle detection dataset and benchmark. In *International Conference on Computer Vision (ICCV)*, 2023. 5, 6, 12, 13, 14