

Overview of FathomNetCLEF 2026: Positive-Unlabeled Object Detection in Marine Images

Laura V. Chrobak^{1,*}, Kevin Barnard¹ and Kakani Katija¹

¹*Research and Development, Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA*

Abstract

The first edition of FathomNetCLEF 2026, organized within LifeCLEF, focused on positive-unlabeled (PU) object detection in marine images. Participants were challenged to train on a sparsely labeled dataset while performance was evaluated on a fully annotated test set spanning 32 marine classes. The competition aimed to advance methods for marine object detection in settings where training data frequently contain many unlabeled positive instances, a common real-world condition that typically leads to an excess of false positives. The challenge saw a total of 452 entrants, 145 participants, and 102 teams. This overview paper describes the challenge design, dataset, and task structure. It then examines the methods employed by select teams that submitted working papers and summarizes key competition outcomes. Together, the submitted approaches demonstrate meaningful progress in tackling PU object detection, with top approaches offering novel insights over prior work in this applied setting of underwater imagery.

Keywords

LifeCLEF, FathomNetCLEF, FathomNet, fine-grained visual categorization, positive unlabeled, object detection, machine learning, computer vision, marine, ocean

1. Introduction

Deep-sea exploration and monitoring increasingly depend on large-scale image analysis to understand species diversity, distribution, and ecological interactions. Yet, marine imagery presents unique challenges for computer vision, including high taxonomic diversity, visually similar morphologies, and environments that range from benthic habitats to the open midwaters. A particularly acute issue is the partial and inconsistent annotation of animals in ocean imagery. Unlike many terrestrial datasets, marine image annotation is typically performed by experts with highly specialized knowledge, who may only label the taxa they study. This leads to a mix of fully annotated regions and numerous positive but unlabeled instances within the same image. In other words, “absence of a label” does not mean “absence of an object of interest.” To illustrate this, Figure 1 details an image with initial sparse annotations versus full-coverage annotation.

The FathomNetCLEF 2026 challenge addresses this real-world challenge by focusing on positive-unlabeled learning for object detection. Developing robust methods in this setting will reduce reliance on costly expert annotations, accelerate dataset curation, and broaden the usability of large-scale archives like the FathomNet Database [1]. Ultimately, improving detection under positive-unlabeled regimes will support scalable biodiversity monitoring, enhance conservation strategies, and enable marine researchers to extract ecological insights from decades of already-collected, legacy imagery.

Participants were tasked with developing object detection systems that can operate effectively in a positive-unlabeled learning regime, identifying both the presence and location of marine organisms across diverse habitats. To succeed, participants were expected to explore approaches such as positive-unlabeled, semi-supervised, and self-supervised learning. The challenge not only benchmarks the state of the art in PU detection but also aims to produce practical methods that can generalize to large-scale, heterogeneous marine datasets, directly supporting ongoing conservation and ecological research.

CLEF 2026 Working Notes, 9 – 12 September 2026, Jena, Germany

*Corresponding author.

✉ lchrobak@mbari.org (L. V. Chrobak); kbarnard@mbari.org (K. Barnard); kakani@mbari.org (K. Katija)

🆔 0009-0005-6419-9965 (L. V. Chrobak); 0000-0002-1327-6749 (K. Barnard); 0000-0002-7249-0147 (K. Katija)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

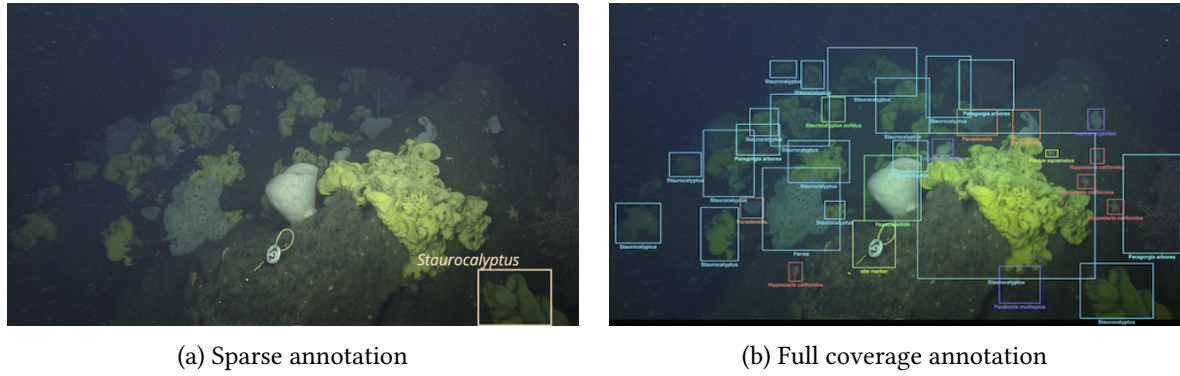


Figure 1: FathomNet sparse annotations. An example image with the (a) original sparse annotation and the (b) full-coverage annotation found in the test set.

2. Challenge Description

2.1. Dataset

The dataset is derived from the FathomNet Database [1], which contains as of February 2026, more than 400k images and 1M expert-curated bounding boxes (bboxes) across more than 2.5k marine taxonomic, geologic, and equipment classes. The dataset is formatted to adhere to the COCO Object Detection standard [2]. For the challenge, participants were provided with a training set of 6,463 images with both labeled and unlabeled positive instances. By contrast, the test set comprises 1,425 fully annotated images (Table 1). To ensure broad accessibility, the entire dataset is less than 50 GB. The FathomNet Database contains a wide spectrum of class labels. For this challenge, our expert team consolidated these labels upstream into 32 grouped categories inspired by the FathomVerse app, which are based on morphological and taxonomic features [3] and are illustrated with visual examples in Figure 3. The entire dataset includes only annotations using these 32 consolidated categories; the original FathomNet labels and the label mapping were not shared.

Figure 2 shows the annotation count for each class in both the training and test sets on a log scale. As is common in ecological datasets, the class distribution is highly long-tailed, with a small number of abundant classes and many rare ones. Notably, for a large proportion of classes, the test set contains more annotated instances than the training set. This imbalance is intentional and reflects the positive-unlabeled (PU) design of the dataset, where the test set provides more comprehensive annotations than the sparsely labeled training data.

Table 1
FathomNetCLEF 2026 dataset statistics.

Subset	Bounding Boxes	Images	Mean bboxes/image
Training	22,225	6,463	3.44
Test	16,603	1,425	11.65

2.2. Evaluation

The evaluation metric for this challenge was mean Average Precision (mAP), applied to a full coverage test set of 1,425 images. Following the COCO definition [2], AP is averaged over 10 Intersection over Union (IoU) thresholds ranging from 0.50 to 0.95 at 0.05 intervals.

This test set was generated by expert taxonomists over a four-month period. The images span both benthic and mid-water scenes and were annotated with the goal of achieving full coverage of all animals present. Each animal was labeled to the lowest taxonomic rank known to the annotating expert; as

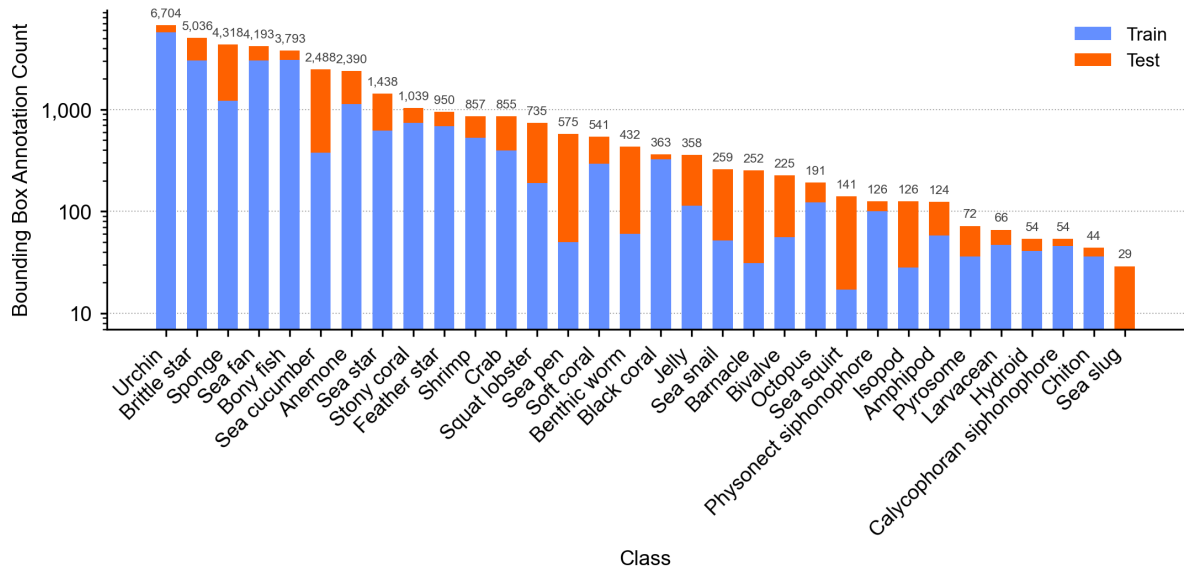


Figure 2: Annotation Count. Annotation count in the training and test sets for each of the 32 classes in the FathomNet CLEF 2026 Dataset on a log scale. Note that the sea slug training bar is nearly invisible ($n=7$) even on a log scale.



Figure 3: FathomNetCLEF 2026 classes. Exemplar images from the 32 classes included in this dataset. Note that these examples were selected as iconic representations of the classes. For the majority of the data, the animals are often more blurry, distant, or difficult to distinguish from the background.

described in 2.1, these fine-grained labels were then mapped to the 32 consolidated categories used for this challenge.

2.3. Timeline

The FathomNetCLEF 2026 competition launched on February 6, 2026, and was promoted through the LifeCLEF [4], Kaggle [5], and FGVC [6] challenge webpages. The competition ran for three months, with a final submission deadline of May 7, 2026.

For evaluation, the test set was split into a public and a private portion; in both cases, the underlying

solutions remain confidential. Participants were permitted to submit up to ten predictions per day via the Kaggle evaluation platform, where performance was assessed on the public test set, and rankings were visible throughout the competition on the public leaderboard. At the conclusion of the competition, each team could designate up to five submissions for evaluation on the private test set. Final rankings and the overall winners were determined by each team's highest-scoring submission on the private test set, revealed on the private leaderboard once the competition closed. The private and final leaderboard was based on 52% of the total test set.

3. Challenge Results

This was the inaugural edition of FathomNetCLEF within the LifeCLEF workshop. The competition saw 3,299 runs submitted by 102 teams, with six teams submitting working notes. Details regarding the top approaches are summarized in this paper [7]. Performance metrics for the top 10 teams on the private leaderboard are shown in Figure 5, which also includes public leaderboard scores. The highest private score was 0.3042, while the public score for that submission was 0.321. The top 10 submissions differed by 33.2%, indicating a significant gap. There was also limited correlation between the public and private leaderboard scores, implying possible overfitting to the public test set.

Figure 4 shows the mAP score averaged over the top 10 teams broken out by class. Here, the classes are sorted in descending order by their mAP score.

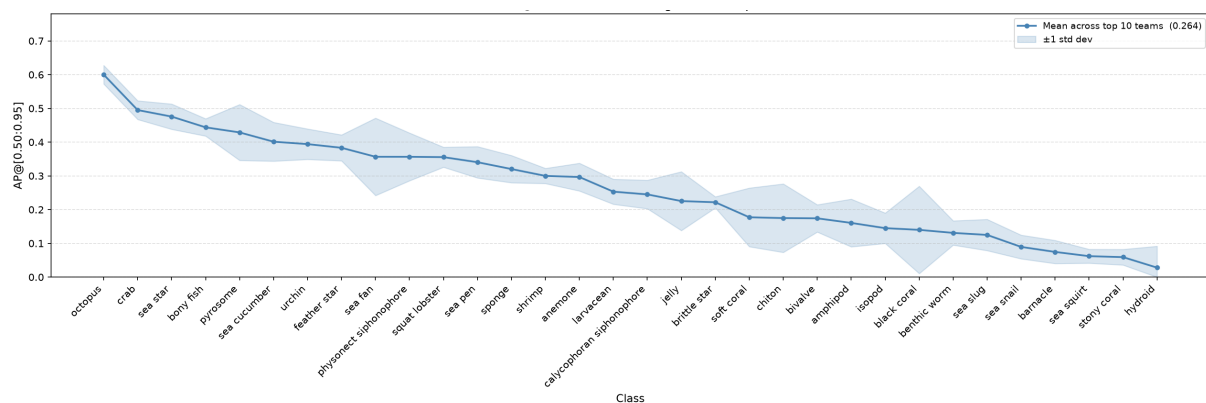


Figure 4: Per-Class Mean Average Precision (mAP) scores by Team. mAP scores broken down by each of the 32 classes across top 10 teams. Scores here are reflective of the private leaderboard.

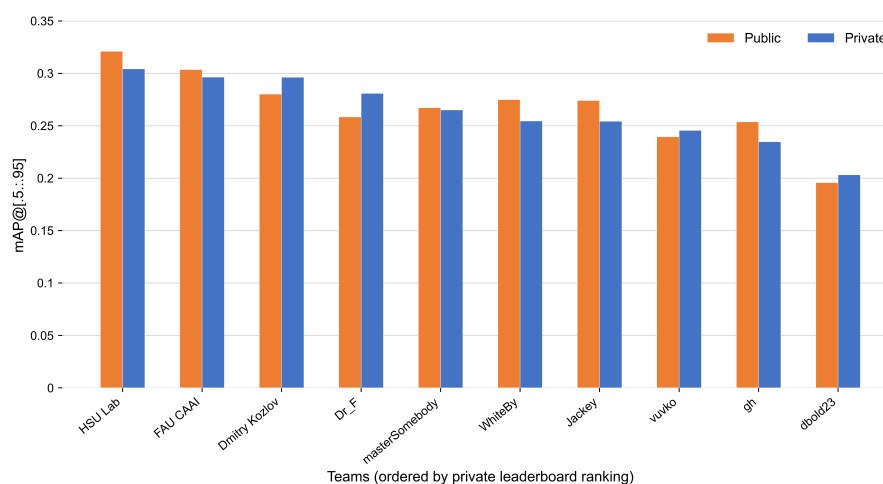


Figure 5: FathomNetCLEF 2026 results. The top 10 teams, ranked by their private leaderboard scores.

3.1. Participants and Methods

We received six submissions, five of which were accepted into the proceedings. Details of these approaches are further developed in participants' working notes [8, 9, 10, 11, 12, 13]; of these, four representing distinct methodological approaches are summarized below. Note that the summarized teams do not necessarily correspond to the top-ranked leaderboard positions, as not all teams submitted a working note.

Team HSU Lab [11] (1st place) tackled the PU object detection challenge by leveraging an ensemble of pre-trained marine object detectors, specifically, MBARI/FathomNet YOLO models [14] [15], with weights kept frozen rather than fine-tuned on the competition training data. Their results clearly show the advantage of this approach, substantially outperforming the 0.03 mAP baseline achieved by directly fine-tuning YOLO. This aligns with a broader pattern across submissions: fine-tuning a detector on the official training split proved unreliable, since unlabeled objects can act as false negatives that penalize correct detections. Their approach then combines detections from the ensemble using cluster-level late fusion, wherein overlapping detections are merged and re-ranked by a lightweight PU-aware fusion network trained on verified positive annotations.

Team FAU CAAI [13] (2nd place) also used an ensemble of object detection models with distribution-aware techniques designed to mitigate train/test domain shift. Their largest performance boost came from augmenting the training data with additional publicly available images mined from the FathomNet database, incorporating frozen MBARI VARS-pretrained YOLO models [14] as zero-shot ensembles, and applying test-time augmentation (TTA) at resolutions that matched the test set's native bimodal distribution. Interestingly, they found that more external data was not always beneficial, concluding that the added domain shift it introduced could outweigh any benefit to labeling. The authors report extensive negative results and show, similarly to team HSU Lab, that keeping the MBARI pretrained models frozen outperformed fine-tuning.

Team gh [10] (9th place) combines a competition trained YOLOv8x detector with a frozen MBARI-315k/FathomNet detector used exclusively at inference time as a marine-domain semantic donor. Rather than developing a formal PU learning algorithm, they found that the largest performance gains came from conservatively mapping the donor's fine-grained taxonomy to the 32 competition classes, calibrating donor confidence scores, and merging donor and base predictions with class-wise non-maximum suppression (NMS). The authors note that this calibrated transfer of a marine-domain detector proved more effective in this challenge than their exploratory work on PU-adjacent training, artifact pruning, and additional competition-only ensembling branches. Interestingly, the donor contributed only 11% of the final detectors, yet accounted for a nearly two-fold increase in performance, leading the authors to conclude that the gains were better explained by the semantic value the donor added rather than simply the volume of proposals it contributed. This is a particularly compelling insight in the application space as machine learning designers gauge whether to prioritize quality or quantity of additional class examples, especially in domains where each additional annotation is costly, either from acquisition of the observation itself and/or expert verification.

Team Robert0921's [9] (11th place) key insight was to decouple detection from classification, motivated by the observation that end-to-end detection plateaued around 0.12 mAP on the public leaderboard and could not overcome the positive-unlabeled setting. Their pipeline runs a frozen Megalodon YOLOv8x detector to generate proposals, a fine-tuned DINOv3 classifier to label each crop, and fuses both confidence scores for final ranking. They also used tiled inference, cropping each image into overlapping patches before running the detector, to prevent small organisms from shrinking below detectable size when the full image is resized. They found that validation metrics computed on the training set, along with detector-only metrics, were unreliable for model selection given the train/test domain shift. Instead, the authors built their own external proxy dataset from publicly available FathomNet-sourced annotations matching the test set's distribution, and combined this with public leaderboard feedback for model selection; they found that preserving proposal recall and improving downstream ranking consistently outperformed fine-tuning the detector or training on noisy pseudo-labels.

4. Discussion and Conclusion

Several key takeaways emerge from the top-performing approaches in this inaugural edition of FathomNetCLEF. First, and perhaps unsurprisingly, fine-tuning a detector directly on the official training split proved unreliable in the positive-unlabeled setting. Unlabeled organisms act as false negatives, penalizing correct detections and decreasing overall performance. Several teams addressed this by leveraging ensemble approaches, combining FathomNet and MBARI baseline models for pseudo-labeling or to train more robust proposal scorers.

The strongest-performing approaches tackled the positive-unlabeled problem indirectly rather than through explicit PU methods; pseudo-labeling, self-training, and fine-tuning consistently hurt performance. Fine-tuning MBARI models was particularly damaging, as it altered the in-distribution pretraining that made them valuable in the first place. Successful strategies instead relied on incorporating external data, exploiting model diversity through ensembles rather than searching for a single optimal model, and optimizing scoring and ranking to account for distribution shift.

Building on this need for external signal, foundation models such as BioCLIP 2.5 and DINOv3 remain promising directions for incorporating knowledge beyond the training set. However, post-hoc uses of these models proved unstable in practice. Directly editing detector labels, replacing output scores, or aggressively re-ranking detections after inference using these models consistently hurt leaderboard performance. Rather than reflecting a limitation of these models' underlying capability, these results suggest the challenge lies in how to integrate them more effectively [8] [9].

These modeling lessons are best interpreted alongside the limitations of the test data itself. Figure 6 shows ground truth and predicted annotations for a top-performing class, octopus (left panel), alongside the brittle star class, which did not perform as well overall (right panel). Notably, the HSU Lab model appears to detect brittle stars not captured in the ground truth - a reminder that this test set, despite being meticulously curated by expert marine taxonomists over hundreds of hours, will never achieve perfect annotation coverage. The annotation approach can also vary in ways that meaningfully affect evaluation; this is particularly true for corals, where bounding an individual organism presents an inherently difficult labeling task. Annotation incompleteness is ultimately unavoidable and in some capacity will be present in all test data.

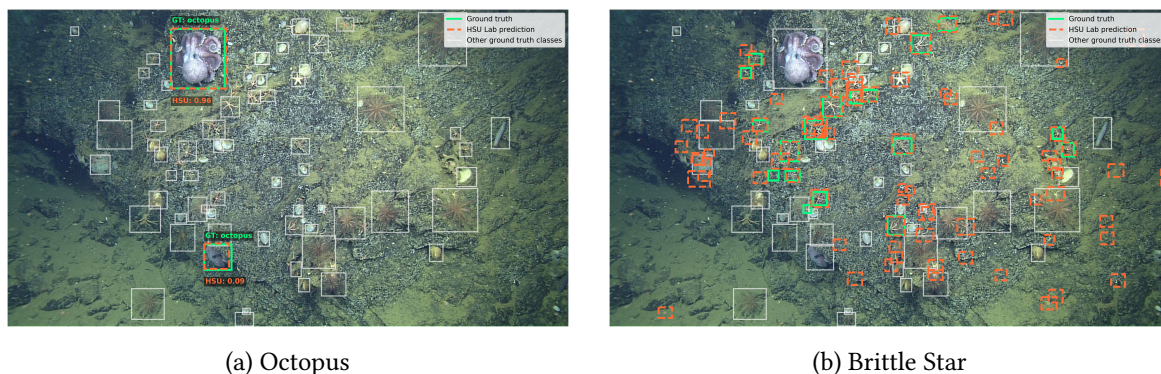


Figure 6: Example test images illustrating challenge results of team HSU Lab across classes. (a) An octopus detection example in which ground truth annotations are shown in solid green and predictions from the top-performing team (HSU Lab) are shown in dashed orange. (b) A brittle star detection example, representing a more challenging class in this challenge. White bounding boxes represent other concepts besides octopuses and brittle stars, respectively.

5. Acknowledgments

Seed funding for FathomNet was provided by National Geographic Society (518018 to KK), National Oceanic and Atmospheric Administration (NA18OAR4170105 to KCB), and the Monterey Bay Aquarium

Research Institute through generous support from the David and Lucile Packard Foundation (to KK). Additional funding support has been provided by National Geographic Society (NGS-86951T-21 to KCB), the National Science Foundation (OTIC 1812535 and Convergence Accelerator 2137977; to KK), and the Monterey Bay Aquarium Research Institute (to KK). Additional individuals whose contributions enriched FathomNet include contributors to the full coverage test dataset (Linda Kuhn, Rob Sherlock, Megan Bassett, and Lonny Lundsten), members of MBARI's Video Lab (Nancy Jacobsen Stout, Kyra Schlining, Susan von Thun, Kristine Walz, Larissa Lemon), Bioinspiration Lab (Joost Daniels, Paul Roberts, Krish Mehta), and Alexandra Lapidus. National Geographic Society contributions were facilitated by Denley Delaney and Alan Turchik.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checks and ChatGPT for improving clarity and rewording sentences. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] K. Katija, E. Orenstein, B. Schlining, L. Lundsten, K. Barnard, G. Sainz, O. Boulais, M. Cromwell, E. Butler, B. Woodward, et al., Fathomnet: A global image database for enabling artificial intelligence in the ocean, *Scientific Reports* 12 (2022) 15914. doi:10.1038/s41598-022-19939-2.
- [2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014*, Springer International Publishing, Cham, 2014, pp. 740–755.
- [3] G. Patterson, J. Daniels, B. Woodward, K. Barnard, G. Sainz, L. Lundsten, K. Katija, Fathomverse: A community science dataset for ocean animal discovery, 2024. URL: <https://arxiv.org/abs/2412.01701>. arXiv: 2412.01701.
- [4] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: Ai challenges for biodiversity understanding and ecosystem management, in: *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, Springer, 2026.
- [5] L. Chrobak, K. Barnard, Fathomnetclef2026, <https://kaggle.com/competitions/fathomnet-2026>, 2026.
- [6] F. Organizers, 13th fine-grained visual categorization workshop (fgvc13), <https://sites.google.com/view/fgvc13>, 2026.
- [7] L. Chrobak, K. Barnard, K. Katija, Overview of FathomNetCLEF 2026: Positive-unlabeled object detection in marine images, in: *Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum*, 2026.
- [8] C. Bo-An, P.-H. Shen, Y.-Y. Hsu, Frozen-detector cluster-level late fusion for positive-unlabeled marine object detection, in: *Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum*, 2026.
- [9] R. Brock, S. Krupa, J. Kahei Tam, Decoupled pipeline with proposal reranking and score fusion for positive-unlabeled marine species detection, in: *Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum*, 2026.
- [10] G. Haranczyk, Marine-domain semantic transfer for object detection under incomplete annotations, in: *Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum*, 2026.
- [11] Q. Hu, Z. Han, Frozen taxonomy projection and conservative ltr method for fathomnetclef 2026, in: *Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum*, 2026.
- [12] J. Peng, X. Wang, Z. Han, Positive-unlabeled marine object detection with class mapping and

catboost reranking, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

- [13] D. Zimmerman, S. Shukla, G. Sklivanitis, Caai at fathomnetclef 2026: Multi-architecture ensembling with distribution-aware test-time augmentation, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.
- [14] FathomNet, Mbari models, <https://huggingface.co/collections/FathomNet/mbari>, 2026. Hugging Face collection, trained by the Monterey Bay Aquarium Research Institute using FathomNet and VARS data. Accessed: 2026-07-06.
- [15] FathomNet, Fathomnet baseline models, <https://huggingface.co/collections/FathomNet/fathomnet-baseline-models>, 2026. Hugging Face collection. Accessed: 2026-07-06.