Frozen Taxonomy Projection and Conservative LTR method for FathomNetCLEF 2026

Qingbiao Hu[1], Zhongyuan Han*

[1]Foshan University, Foshan, China

Abstract

The FathomNetCLEF2026 evaluation task is a positive-unlabeled object detection benchmark for marine organisms whose training annotations provide positive instances but do not exhaustively mark all visible organisms. For this task, we propose a frozen-taxonomy projection and conservative reranking method. It retains the fine-grained 499-class outputs of a marine-domain MBARI-315k detector, projects them to the 32 challenge categories with positive-unlabeled-aware calibration, and improves predictions through bounded proposal reranking and multi-scale localization refinement. Our final submission ranked fifth with 0.2650 mAP on the completed private leaderboard, after ranking sixth with 0.2672 mAP on the public leaderboard.

Keywords

FathomNetCLEF2026, positive-unlabeled object detection, proposal reranking, localization refinement

1. Introduction

FathomNetCLEF2026 is a positive-unlabeled (PU) object detection task in marine imagery, organized in the LifeCLEF 2026 evaluation campaign [1, 2]. Participants must detect organisms from 32 consolidated categories, using sparsely annotated training images and a fully annotated hidden evaluation set [3]. The challenge images originate from the FathomNet ecosystem, an expert-annotated marine image database developed to enable machine learning for ocean observation [4].

The main difficulty addressed in this paper is incomplete annotation. A visible organism missing from the training annotations is not necessarily background, yet conventional supervised detector training penalizes detections at such unlabeled locations. In our preliminary experiments, a standard 32-class detector and a 32-class fine-tuned marine detector both performed poorly on the public evaluation, indicating that directly adapting a detection head to sparse task labels discards useful recognition signal and introduces false-negative supervision.

To address this problem, we propose a frozen-taxonomy projection and conservative reranking pipeline. We retain a marine-domain MBARI-315k YOLOv8 detector in its original 499-class label space [5, 6], project fine-class evidence to the 32 evaluation categories before non-maximum suppression, and learn only small PU-aware calibration and reranking modules. Multi-scale predictions are used as bounded support for score ordering and localization refinement rather than aggressive detection fusion.

2. Related Work

FathomNet provides expert-annotated underwater imagery and taxonomic metadata for training and evaluating marine visual recognition systems [4], and was used for the FathomNet2023 detection competition dataset [7]. Related underwater benchmarks address different visual settings: the Brackish dataset provides bounding-box annotated marine animals captured in brackish water [8], SUIM evaluates semantic segmentation of underwater imagery [9], and TrashCan targets marine debris segmentation [10]. FathomNetCLEF2026 is distinguished here by sparse positive labels for category-level organism detection.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

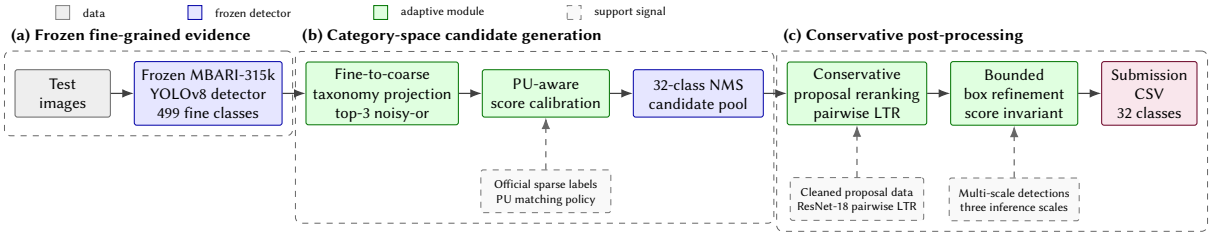

Figure 1: Overall architecture of the final FathomNetCLEF2026 system. The main horizontal path is the submission-time inference pipeline. Dashed regions group the frozen fine-grained detector, the category-space candidate generation stage and the conservative post-processing stage. Bottom signals are used only for calibration, reranking or localization support.

A related FathomNet 2025 competition considered hierarchical image classification rather than PU object detection. The organizers' results report states that its winning solution used organism crops, environmental context and hierarchy-aware classification [11]. The 2026 task instead requires bounding-box localization in the presence of missing annotations.

PU learning addresses training from labeled positives and unlabeled examples [12, 13, 14]. Missing annotations in detection have specifically been treated as a PU problem by Yang et al. [15]. Our method uses this task constraint operationally: the domain detector is frozen, and sparse labels are used for conservative calibration and proposal ranking rather than for training a new detection head.

The operations used downstream of the frozen detector are established components rather than new model families. Temperature scaling is a standard calibration approach [16]; Soft-NMS and weighted boxes fusion are representative detector post-processing alternatives [17, 18]. In our ablations, bounded updates were retained because aggressive fusion increased false positives and reduced evaluation mAP.

3. Method

Figure 1 summarizes the final system. The detector itself remains frozen; all task-specific adaptation is performed after MBARI detection through pre-NMS taxonomy projection and positive-unlabeled calibration, conservative proposal reranking, gated proposal addition, and localization-only box refinement.

The paper uses the PU detection setting of Yang et al. [15] as a problem formulation, but it does not introduce a new PU risk estimator or train an end-to-end detector with a PU loss. The methodological contribution is the combination of four conservative adaptations around a frozen marine detector: manually audited fine-to-coarse taxonomy projection before NMS, small PU-aware score calibration, bounded proposal reranking, and localization-only refinement. These choices are intended to preserve useful fine-grained detector evidence while avoiding false-negative supervision from unlabeled organisms. Algorithm 1 gives the corresponding submission-time procedure.

Algorithm 1. Inference procedure used by the final system.

Input: test image x, frozen MBARI-315k detector D, audited mapping M from MBARI fine classes to challenge categories, calibrated projection parameters, pairwise LTR reranker, and multi-scale support detections.
Output: submission rows (b, c, q) for the 32 challenge categories.

1. Run D at the target image scale and keep the raw 499-class candidate scores before final category suppression.
2. For each challenge category c, collect mapped fine classes $G_c = M^{-1}(c)$, apply fine-class bias terms, and aggregate the top three mapped responses with noisy-or pooling.
3. Apply category-specific bias and temperature calibration learned from positive, ignored and low-risk-negative proposals, then run NMS in the 32-class category space.
4. Rerank retained proposals using cross-scale support, sequence context and the pairwise LTR reranker; add new proposals only when strict support and count limits are satisfied.
5. Refine box geometry toward same-class multi-scale support when the movement constraints are satisfied; keep category labels and scores unchanged.

3.1. Task Formulation

Let $\mathcal{C}$ denote the set of 32 challenge categories. Each training image x_i is accompanied by an observed annotation set A_i, while the complete set of visible instances $A_i^\star$ is generally unknown and may satisfy $A_i \subset A_i^\star$. The goal is to predict a set of detections $\hat{A}(x) = \{(b_k, c_k, q_k)\}$ for a test image, where b_k is a bounding box, $c_k \in \mathcal{C}$ and q_k is a confidence score. Evaluation is carried out on fully annotated hidden test images using mean Average Precision.

The key training constraint is that a proposal not matched to A_i cannot automatically be assigned a negative label. In our calibration and ranking stages, proposals matched to same-class annotations are positives; ambiguous unmatched proposals, including high-confidence detections or proposals near another annotated object, are ignored; only unmatched proposals far from any annotation are used as low-risk negatives.

3.2. Proposed Method

3.2.1. Frozen Detector and Taxonomy Projection

We use the MBARI-315k YOLOv8 model as a frozen detector with 499 fine-grained marine classes. Let $s_j(b)$ be its confidence for fine class j on box candidate b, and let G_c contain the fine classes mapped to challenge category c. The mapping was constructed from explicit MBARI model class names to the 32 official category names. Fine classes were mapped when their taxonomic or concept label was unambiguous for a challenge category, including species, genera and higher taxa that were direct members of that category. Broad parent or ambiguous classes, such as *Animalia*, *Cnidaria*, *Decapoda*, *Siphonophorae*, *Tunicata* and *Hexacorallia*, were intentionally left unmapped. The conservative mapping covered 357 of the 499 MBARI classes. Its effect was evaluated by separating direct post-hoc mapping (E12), pre-NMS aggregation (E14) and trainable fine-class biasing (E22) in the ablation table, and by testing a larger mapping overlay that did not improve the final pipeline.

We learn a bounded fine-class logit bias α_j:

$$s'_j(b) = \sigma(\mathrm{logit}(s_j(b)) + \alpha_j),$$

and aggregate the top three mapped responses before NMS:

$$p_c(b) = 1 - \prod_{j \in T_3(G_c, b)} (1 - s'_j(b)).$$

Here $T_3(G_c, b)$ contains the three highest-scoring fine classes in G_c for candidate b. Pre-NMS aggregation preserves coarse-category evidence that would otherwise be discarded by a competing fine-class maximum.

3.2.2. PU-Aware Calibration

We calibrate projected category scores while the detector remains frozen. For category-specific bias β_c and temperature τ_c, the calibrated score is

$$\hat{p}_c(b) = \sigma\left(\frac{\mathrm{logit}(p_c(b)) + \beta_c}{\tau_c}\right).$$

The parameters α_j, β_c, and τ_c are optimized using the PU-aware labeling policy above, with clipping and L2 regularization to prevent large score shifts under sparse labels. Calibrated 32-class responses are then filtered with category-space NMS. The temperature parameterization follows standard confidence calibration [16]; the task-specific choice is how proposals are selected for these bounded adjustments.

3.2.3. Conservative Reranking and Refinement

Starting from the projected 960-pixel candidates, we use independent multi-scale outputs as support rather than applying full weighted box fusion [18]. Scale-consistent boxes and boxes supported by sequence context receive bounded score updates; new proposals are admitted only under strict support and count caps. A supervised ResNet-18 [19] crop reranker is then trained from official annotations and cleaned FathomNet images. Its PU-aware learning-to-rank variant uses a pairwise score-ordering objective in the learning-to-rank family [20]. It operates within per-image, per-category proposal groups and applies bounded logit adjustments using crop embeddings, detector confidence, box geometry, overlap context and rank features. Relevance labels are derived from same-class IoU to observed annotations; ambiguous proposals are excluded from pair construction rather than forced to be negatives. Pairwise training therefore changes only the relative ordering of proposals that can be compared with lower PU risk.

Finally, the localization refinement module changes neither categories nor scores. For each retained prediction, it identifies same-class support boxes at 768, 960 and 1152 pixels and moves its geometry only slightly toward a weighted support target. Minimum overlap and maximum movement constraints ensure that refinement remains local and does not behave like unconstrained box fusion. In the final refinement family, a box required same-class multi-scale support and was rejected if the refined geometry moved too far from the original prediction; this explains the sharp drop observed for the less constrained E40 variant.

4. Experiments

4.1. Experimental Setup

4.1.1. Experimental Data

The official data are provided in COCO format [21]. They contain 6,463 training images with 22,225 observed boxes and 1,425 hidden evaluation images. We split the official training data into 5,494 training images and 969 validation images, containing 18,766 and 3,459 labeled boxes, respectively. For the supervised crop reranker, we additionally used cleaned images collected from the public FathomNet Database [4].

4.1.2. Data Processing

We converted the official annotations to YOLO format for early baselines. For the final method, raw outputs from the frozen MBARI model were cached at 768, 960 and 1152 pixels. Taxonomic mapping and calibration were applied before category-space NMS. To construct training instances for the calibration and reranking modules, same-class ground-truth matches were positive, ambiguous unmatched proposals were ignored and distant unmatched proposals were treated as low-risk negatives. The final refinement module used multi-scale predictions only as geometric support.

4.1.3. Evaluation Metric

We report mean Average Precision (mAP) returned by the FathomNetCLEF2026 evaluation platform. The public evaluation split contains approximately 48% of the test data and is used to report the ablation results. The final competition result is calculated on the remaining private split of approximately 52% of the test data. Higher mAP indicates better combined classification and localization. Before submission, we also audited row counts, score quantiles, category growth, invalid boxes and the maximum number of predictions per image; these audits were screening criteria rather than evaluation metrics.

4.1.4. Baseline Methods

We compared the proposed pipeline with four representative baselines: (1) a conventional 32-class YOLO26n [22] trained on official annotations (E1); (2) a marine-domain MBARI model fine-tuned as a 32-class detector (E10); (3) a frozen MBARI detector followed by direct post-hoc 499-to-32 mapping (E12); and (4) a frozen MBARI detector with pre-NMS max-score taxonomy aggregation (E14). Additional ablations examine calibration, reranking, proposal addition and localization refinement.

4.2. Results and Analysis

The final E39 submission achieved 0.2672 mAP and ranked sixth on the public evaluation split. On the private split used for the final standings, it achieved 0.2650 mAP and ranked fifth. Table 1 reports the ablation results on the public split, while the private score is reported as the final competition result.

Table 1
Key ablations on the public evaluation split.

ID	Method	Public mAP
E1	Standard 32-class YOLO26n, image size 768	0.0283
E10	Fine-tuned MBARI-315k YOLOv8 as a 32-class detector	0.0356
E12	Frozen MBARI, direct 499-to-32 taxonomy mapping	0.2565
E14	Pre-NMS 499-to-32 max aggregation	0.2592
E18-lite	Top-3 noisy-or with PU-aware calibration	0.2616
E22	Trainable fine-class projection bias	0.2621
E24	Strict 960-pixel inference	0.2649
E25	Safe two-scale reranking	0.2661
E33	Crop-based PU-aware learning-to-rank reranker	0.2667
E39	Multi-scale box refinement	**0.2672**

4.2.1. Analysis

Table 1 shows that preserving the frozen MBARI taxonomy was the primary source of improvement. Replacing the 499-class head with a newly trained 32-class head achieved only 0.0356 mAP (E10), whereas direct mapping of the original frozen detector reached 0.2565 mAP (E12). Pre-NMS projection and PU-aware calibration then improved this base without exposing the detection head to uncertain negative labels. Mapping coverage alone was not sufficient: a later direct-mapping branch with 369 mapped MBARI classes scored 0.2603 mAP, and mapping-driven proposal addition reached only 0.2660 mAP. This supported the final choice of conservative mapping plus bounded calibration and reranking rather than simply adding more fine classes.

The later ablations indicate that conservative post-processing was preferable to aggressive proposal expansion. Safe multi-scale reranking, contextual gating, pairwise LTR reranking and bounded refinement produced incremental gains from E24 to E39, reaching 0.2672 mAP. In contrast, representative aggressive variants regressed: early weighted box fusion reached only 0.0098 mAP, student fusion reached 0.2603 mAP, and mapping-driven proposal addition reached 0.2660 mAP. These outcomes are consistent with increased false positives under sparse labels.

Localization refinement was beneficial only inside a narrow trust region. E39 improved the retained proposal set by moving boxes toward same-class multi-scale support, while the less constrained E40 variant fell to 0.2427 mAP. This result supports the proposed bounded refinement design.

5. Limitations and Generalizability

Although developed for marine imagery, the framework can transfer to other PU detection tasks when a reliable fine-grained source detector and a meaningful source-to-target taxonomy mapping are available.

Its applicability mainly depends on the alignment between the source and target label spaces. Future work could automate the mapping and calibration stages using ontology metadata and validation evidence, reducing the manual effort required for a new domain.

6. Conclusions

We presented a frozen-taxonomy projection and conservative reranking method for PU marine object detection in FathomNetCLEF2026. Instead of retraining a 32-class detection head from incomplete labels, the method retains the MBARI-315k fine-grained detector and applies pre-NMS taxonomy projection, PU-aware calibration, bounded reranking and multi-scale localization refinement. The final submission achieved 0.2672 public mAP, demonstrating the value of preserving marine-domain taxonomy while limiting updates driven by uncertain negatives. In the completed evaluation, the same submission achieved 0.2650 private mAP and placed fifth in the final standings.

Acknowledgments

This work is supported by the Guangdong Province Basic and Applied Basic Research Fund (Guangdong-Foshan Joint Funds, 2024A1515140026)

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI Codex for grammar and spelling check, paraphrase and reword, improve writing style, citation management, and formatting assistance, including LaTeX and TikZ source editing. The tool was not used to generate experimental data, reported metrics, or scientific conclusions. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] 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.

[2] 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.

[3] K. Barnard, L. Chrobak, FathomNetCLEF2026 @ LifeCLEF & CVPR-FGVC, https://kaggle.com/competitions/fathomnet-2026, 2026. Unpublished. Kaggle.

[4] K. Katija, E. Orenstein, B. Schlining, L. Lundsten, K. Barnard, G. Sainz, O. Boulais, M. Cromwell, E. Butler, B. Woodward, K. L. C. Bell, FathomNet: A Global Image Database for Enabling Artificial Intelligence in the Ocean, Scientific Reports 12 (2022) 15914. doi:10.1038/s41598-022-19939-2.

[5] FathomNet, MBARI-315k-yolov8, https://huggingface.co/FathomNet/MBARI-315k-yolov8, 2023. Hugging Face model repository.

[6] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, Software release, https://github.com/ultralytics/ultralytics, 2023. Version 8.0.0; model documentation: https://docs.ultralytics.com/models/yolov8/.

[7] E. C. Orenstein, et al., The FathomNet2023 Competition Dataset, arXiv preprint arXiv:2307.08781, 2023. URL: https://arxiv.org/abs/2307.08781.

[8] M. Pedersen, J. Bruslund Haurum, R. Gade, T. B. Moeslund, Detection of Marine Animals in a New Underwater Dataset with Varying Visibility, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 18–26.

[9] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, J. Sattar, Semantic Segmentation of Underwater Imagery: Dataset and Benchmark, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 1761–1768. doi:10.1109/IROS45743.2020.9340821.

[10] J. Hong, M. Fulton, J. Sattar, TrashCan: A Semantically-Segmented Dataset towards Visual Detection of Marine Debris, arXiv preprint arXiv:2007.08097, 2020. URL: https://arxiv.org/abs/2007.08097.

[11] L. Chrobak, Looking at the Bigger Picture: Results from FathomNet's 2025 Kaggle Competition, https://www.fathomnet.org/news/looking-at-the-bigger-picture%3A-results-from-fathomnet%E2%80%99s-2025-kaggle-competition, 2025. FathomNet, August 11, 2025.

[12] C. Elkan, K. Noto, Learning Classifiers from Only Positive and Unlabeled Data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 213–220. doi:10.1145/1401890.1401920.

[13] M. C. du Plessis, G. Niu, M. Sugiyama, Convex Formulation for Learning from Positive and Unlabeled Data, in: Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 1386–1394.

[14] R. Kiryo, G. Niu, M. C. du Plessis, M. Sugiyama, Positive-Unlabeled Learning with Non-Negative Risk Estimator, in: Advances in Neural Information Processing Systems, volume 30, 2017, pp. 1675–1685.

[15] Y. Yang, K. J. Liang, L. Carin, Object Detection as a Positive-Unlabeled Problem, in: Proceedings of the British Machine Vision Conference, 2020, pp. 1–12.

[16] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On Calibration of Modern Neural Networks, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1321–1330.

[17] N. Bodla, B. Singh, R. Chellappa, L. S. Davis, Soft-NMS – Improving Object Detection with One Line of Code, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5561–5569. doi:10.1109/ICCV.2017.593.

[18] R. Solovyev, W. Wang, T. Gabruseva, Weighted Boxes Fusion: Ensembling Boxes from Different Object Detection Models, Image and Vision Computing 107 (2021) 104117. doi:10.1016/j.imavis.2021.104117.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[20] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, Learning to Rank Using Gradient Descent, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 89–96. doi:10.1145/1102351.1102363.

[21] 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: Computer Vision – ECCV 2014, Springer, 2014, pp. 740–755.

[22] G. Jocher, J. Qiu, Ultralytics YOLO26, Software release, https://github.com/ultralytics/ultralytics, 2026. Version 26.0.0; model documentation: https://docs.ultralytics.com/models/yolo26/.