Positive-Unlabeled Marine Object Detection with Class Mapping and CatBoost Reranking

Junwei Peng[1], Xingrun Wang[1] and Zhongyuan Han[1,*]

[1]Foshan University, Guangdong, China

Abstract

The FathomNet 2026 evaluation task aims to detect and localize marine organisms in underwater images under a positive-unlabeled setting, where unlabeled objects may still be present in the images. For this task, we propose an inference-oriented marine object detection and reranking pipeline based on an MBARI-315k YOLOv8 detector. The method reads the detector's pre-NMS 499-class output, maps the MBARI class scores to the 32 competition categories with a manual class-mapping table, and applies NMS on the resulting competition-class output. We further improve candidate generation and score ordering through multi-scale inference, weighted boxes fusion, and CatBoostRanker-based confidence reranking. On the FathomNet 2026 evaluation data, the final multi-scale WBF and CatBoostRanker method achieved a Public score of 0.2749 with rank 4 and a Private score of 0.2546 with rank 6.

Keywords

Positive-Unlabeled Object Detection, Marine Images, CatBoost Reranking

1. Introduction

The FathomNet 2026 challenge, organized as part of LifeCLEF 2026, focuses on object detection in underwater images under a positive-unlabeled setting [1, 2]. The task requires systems to identify and localize marine organisms while handling the fact that unlabeled objects may still be present in the images. This setting reflects real marine visual archives, where expert annotations are costly and often incomplete.

A major difficulty in this task is reliable candidate scoring under incomplete annotations. Marine organisms can show similar appearances across different underwater environments, making it difficult to distinguish true detections from visually plausible false positives. In addition, since the official metric is mAP, the confidence assigned to each candidate directly affects the ordering of true and false detections.

In our initial experiments, we tried to fine-tune the detector directly with the competition training set, but the early Public scores consistently remained below 0.1. We hypothesize that this was mainly caused by the positive-unlabeled annotation setting: direct fine-tuning implicitly treats unlabeled but present objects as background, which can weaken the fine-grained marine recognition ability already learned by the MBARI pretrained detector. Therefore, we abandoned the direct fine-tuning route and shifted subsequent experiments toward inference-time adaptation without updating the detector parameters.

To address this problem, we propose an inference-oriented detection and reranking pipeline based on the MBARI-315k YOLOv8 detector. Our method reads the detector's pre-NMS 499-class output, maps the MBARI class scores to the 32 competition categories with a manual class-mapping table, and applies NMS directly on the mapped output. We further use multi-scale inference, weighted boxes fusion (WBF), and CatBoostRanker-based confidence reranking to improve candidate generation and score ordering. The final submitted system achieved a Public score of 0.2749, ranking fourth, and a Private score of 0.2546, ranking sixth.

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

*Corresponding author.

841099774@qq.com (J. Peng); wangxingrun@fosu.edu.cn (X. Wang); hanzhongyuan@gmail.com (Z. Han)

0009-0006-3925-8158 (J. Peng); 0000-0002-7364-198X (X. Wang); 0000-0001-8960-9872 (Z. Han)

2. Related Work

Marine image analysis has increasingly relied on large curated visual archives and deep object detectors. FathomNet was introduced as an open-source image database that standardizes and aggregates expertly curated ocean imagery for machine learning, while also emphasizing that manual annotation of complex marine visual data is difficult to scale [3]. Recent surveys of underwater object detection similarly note that underwater imagery suffers from low visibility, low contrast, blur, color deviation, and diverse acquisition conditions, so many systems adapt general-purpose detectors with image enhancement, feature fusion, multi-scale processing, or domain adaptation [4]. These observations are consistent with the FathomNet 2026 setting, where robust candidate generation and score calibration are important because marine objects can be small, visually similar, and embedded in heterogeneous backgrounds.

Incomplete annotation is a central issue in this task. Yang et al. explicitly formulated object detection as a positive-unlabeled problem and argued that standard positive-negative training treats unlabeled object regions as background, which can produce misleading supervision when annotations are missing [5]. Later work on sparsely annotated detection followed related ideas: SparseDet separates labeled and unlabeled proposals with pseudo-positive mining to reduce the effect of missing boxes [6], while Calibrated Teacher uses calibrated pseudo labels and a focal IoU weighting term to handle sparse annotations more robustly [7]. These methods mainly address training under incomplete labels.

Beyond training strategies for incomplete annotations, post-processing and reranking are also important for improving detection submissions without retraining the backbone detector. Weighted boxes fusion (WBF) was proposed to ensemble detections by averaging overlapping boxes with confidence-based weights, preserving localization evidence from multiple detectors or inference variants rather than simply suppressing all but one box [8]. For score calibration, CatBoost provides an ordered boosting mechanism [9]. In this work, CatBoostRanker is used as a candidate-level reranker over score, normalized box geometry, category and source-domain identity, image-level rank statistics, score percentiles, and mapping-derived category statistics. Our final pipeline follows this practical direction by retaining the MBARI detector, mapping its 499-class output to 32 competition classes before NMS, and then applying multi-scale WBF and CatBoostRanker to improve the ordering of candidate detections.

3. Method

Our method is an inference-oriented object detection and reranking pipeline for positive-unlabeled marine imagery. Figure 1 gives an overview of the full system. The pipeline keeps the original MBARI-315k YOLOv8 detector unchanged, converts its pre-NMS 499-class output into the 32 competition categories, and then improves candidate generation and ordering with multi-scale inference, weighted boxes fusion, and CatBoostRanker-based reranking.

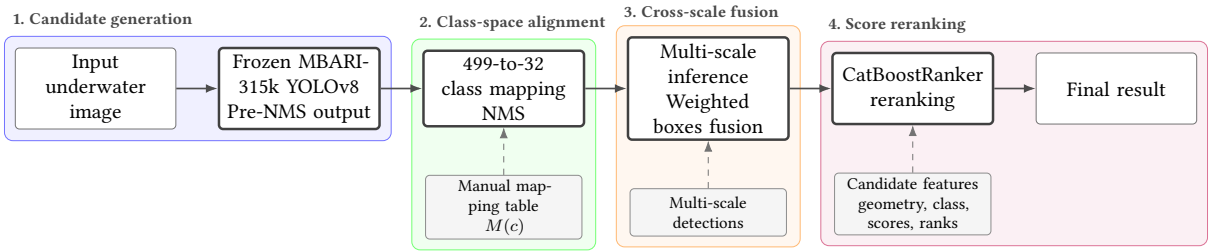

Figure 1: Overview of the proposed detection and reranking framework. Colored regions mark the main stages, while dashed arrows indicate auxiliary information used by the corresponding module.

3.1. Task Formalization

The task is object detection in underwater imagery under a positive-unlabeled learning regime. Given an image, the system outputs a set of candidate detections, each containing an image identifier, a

category identifier, bounding-box coordinates, and a confidence score. A valid submission contains one row per predicted object, including `annotation_id`, `image_id`, `category_id`, `bbox_x`, `bbox_y`, `bbox_width`, `bbox_height`, and `score`. The practical objective is to produce high-quality candidate boxes and rank them so that correct same-class detections appear ahead of false positives.

3.2. Proposed Method

3.2.1. Pre-NMS 499-to-32 Class-Score Mapping

In the initial stage, we ran inference on the test set with the MBARI-315k YOLOv8 detector, but did not directly use the 499-class detections produced by the model's default post-processing. We kept the original 499-class YOLO weights unchanged, read the bounding-box coordinates, objectness confidence, and 499-class classification scores before YOLO NMS, and used a manual mapping table to aggregate the fine-grained MBARI categories into the 32 categories required by the competition. Specifically, for each candidate box, the MBARI class scores belonging to the same competition category were first merged into one competition-class score according to the mapping table. This produced a new 32-class output tensor, on which YOLO NMS was then applied directly. The purpose of this design was to preserve as much of the MBARI detector's fine-grained marine-category recognition ability as possible while aligning the output category space with the competition submission format.

The manual mapping table was constructed as a fixed class-space prior before final inference. We started from the 499 MBARI class names used by the frozen detector and the 32 target categories defined by the competition. Each MBARI class was reviewed according to its taxonomic or semantic meaning. If a class clearly matched one competition category, it was assigned to that category; if it was too broad, ambiguous, outside the competition taxonomy, or visually inconsistent with a single target category, it was left unmapped and did not contribute directly to any submitted class.

To check the mapping, we also used a data-assisted review based on the competition training images. The frozen detector was run on the training set, and its candidate detections were matched with the available annotations to summarize which competition categories, or unmatched regions, each MBARI class tended to overlap with. This step was used to identify ambiguous and conflicting classes, but it was not used as an automatic mapping rule. Since the training data follow a positive-unlabeled setting, an unmatched candidate does not necessarily indicate a false detection; it may also correspond to a real but unlabeled organism.

Ambiguous cases were therefore resolved by jointly considering class-name semantics and the data-assisted summary. MBARI classes with clear taxonomic or semantic correspondence to a competition category were kept in the mapping table, whereas classes that were too broad, semantically non-unique, or difficult to associate with a single target category were left unmapped. The resulting manual mapping table was fixed during inference and used to convert the detector's pre-NMS 499-class scores into the 32 competition categories.

For each competition class, the mapped MBARI class scores were aggregated with max pooling:

$$s_c = \max_{m \in M(c)} s_m,$$

where $M(c)$ is the set of MBARI classes manually mapped to competition class c, and s_m is the detector score for MBARI class m in the pre-NMS output. The resulting 32-class tensor was passed directly to YOLO non-maximum suppression.

We fixed the class aggregation function to max pooling in this work. This is because, in the pre-NMS setting, a competition category may correspond to several fine-grained MBARI classes, but for a candidate box, only one or a few of them usually provide the main support. Max pooling therefore preserves the strongest fine-grained evidence when projecting the frozen detector's 499-class output to the 32 competition categories, while keeping the aggregated score scale close to the original detector output. The goal of this choice was to reduce unnecessary disturbance to candidate ranking under incomplete annotation. We did not perform a systematic comparison against alternative aggregation rules such as mean or sum-based pooling in this work.

Table 1
Dataset metadata used in this evaluation.

Split	Images	Image ID range	Annotations	Categories
Train	6,463	1–6,569	22,225	32
Test	1,425	1–1,460	0	32

3.2.2. Multi-scale Inference and Weighted Boxes Fusion

Single-scale inference can miss small organisms or produce unstable localization for targets with substantial scale variation. To increase candidate diversity, we ran inference at multiple image scales and fused the resulting detections with weighted boxes fusion (WBF). WBF combines overlapping detections by averaging their coordinates with confidence-based weights, preserving localization evidence from multiple inference variants instead of suppressing all but one box.

3.2.3. CatBoostRanker Score Reranking

The CatBoostRanker model was used only to recalibrate the confidence scores of existing candidate detections, without changing their bounding boxes or predicted category IDs. We trained the ranker with the `YetiRankPairwise` objective and grouped candidates by `category_source`, so that detections from the same competition category and source domain were ranked together.

The supervision labels were constructed from the best same-class IoU between each candidate and the available annotations. Candidates with IoU ≥ 0.5 were assigned label 2, candidates with $0.35 \leq \text{IoU} < 0.5$ were assigned label 1, and the remaining candidates were assigned label 0. This formulation encouraged the model to learn a relative ordering that favors better localized candidates under incomplete annotation.

The feature vector combined five types of information: (1) score features, including the original score and its log and logit transforms; (2) normalized box geometry, including box position, size, center, area, aspect ratio, distance to image borders, and whether the box touches an image edge; (3) within-image ranking statistics, such as candidate count, category-specific count, and rank and rank percentile within the image and within the image-category subset; (4) category and source descriptors, including the predicted competition category, source domain, category-source pair, and score percentiles within the category and within the source-category subset; and (5) mapping-derived statistics from the manual class-alignment table, including mapped-class count, support totals, estimated "other" tendency, conflict ratios, unmapped ratio, and tier indicators.

At inference time, the rank predictions were standardized and converted into a bounded multiplicative score update. Specifically, the final score was computed by multiplying the original score with $\exp(\alpha z)$, where z is the standardized rank prediction, and then clipping the multiplier to a fixed range. In the final saved configuration, $\alpha = 0.1$ and the multiplier was clipped to $[0.85,\ 1.15]$. This design preserved the detector's original confidence scale while refining the ordering of true and false detections under the mAP metric.

4. Experiment

4.1. Experimental Setup

4.1.1. Experimental Data

The training metadata file contains image records, bounding-box annotations, and 32 competition categories. The test metadata file contains the target test images and the same category definitions. The image and annotation counts in Table 1 are the numbers of records in the metadata `images` and `annotations` arrays, respectively.

4.1.2. Parameter Settings

The initial detection pipeline used single-scale inference. The input image size was set to 1280, the confidence threshold to 0.001, the NMS IoU threshold to 0.7, and the maximum number of detections per image to 300. The low confidence threshold was used to preserve more potential object candidates and avoid filtering out plausible true detections too early during inference. The final pipeline added multi-scale inference, WBF, and CatBoostRanker reranking to improve cross-scale candidate stability and score ordering.

4.1.3. Evaluation Metric

The official metric is mean Average Precision (mAP) computed on a fully labeled evaluation set. The Public leaderboard score is computed on approximately 48% of the test data and provides feedback during the competition, while the Private leaderboard score is computed on approximately 52% of the test data and is used for the final ranking.

4.1.4. Baselines

As a preliminary detector-adaptation route, we first fine-tuned MBARI YOLOv8 detectors with the official training annotations. To mitigate missing annotations, we further trained pseudo-label detector variants. B1 used high-confidence teacher predictions as pseudo labels, while B2 used a more conservative pseudo-label threshold and stricter per-image and per-class limits. We also trained a low-augmentation detector B3 to increase model diversity, and tested WBF ensembles of these detectors.

We then compared the proposed frozen-detector pipeline with several main variants. The manual pre-NMS 499-to-32 mapping detector is the initial fixed-detector baseline. We also tested whether the conversion from the original MBARI 499-class output space to the 32 competition categories could be learned from the training set. The first trainable variant used a $\mathrm{Linear}(499, 32)$ mapping layer, and the second used a two-layer $499 \to 256 \to 32$ MLP mapper. We also compared CatBoostRanker-only reranking with the full multi-scale WBF and CatBoostRanker pipeline.

4.2. Experimental Results and Analysis

Table 2 summarizes the Public leaderboard scores of the preliminary detector fine-tuning and pseudo-label experiments.

Table 2
Preliminary detector fine-tuning and pseudo-label experiments on the Public leaderboard.

Method variant	Public score
MBARI YOLOv8 detector fine-tuned with official annotations (A)	0.0450
Pseudo-label detector (B1)	0.0485
Conservative pseudo-label detector (B2)	0.0491
Low-augmentation detector (B3)	0.0393
WBF ensemble of A+B1+B2	0.0613
WBF ensemble of A+B1+B2+B3	0.0618

The preliminary detector fine-tuning and pseudo-label route consistently remained below 0.1 on the Public leaderboard. Among individual detectors, the B1 and B2 pseudo-label detectors slightly improved over the baseline detector fine-tuned with official annotations, but the best single-detector score was only 0.0491. The low-augmentation detector B3 achieved 0.0393 as a single model. With WBF, the A+B1+B2 ensemble improved to 0.0613, and adding B3 further increased the score slightly to 0.0618. Although these results indicate some complementarity from pseudo-label training and model fusion, the overall performance remained poor. This result motivated us to stop further detector fine-tuning.

Table 3 summarizes the Public leaderboard scores of the main methodology variants after this design change. The manual pre-NMS 499-to-32 mapping detector achieved a Public score of 0.2740. The trainable linear and MLP mapping layers performed substantially worse, with Public scores of 0.0974 and 0.0671, respectively. These results suggest that learning the class-space conversion directly from the available training annotations was not reliable in this positive-unlabeled setting. The poor performance of the trainable mapping layers indicates that the main issue may be the unreliable supervision used to learn the class-space conversion. Under the positive-unlabeled setting, many real marine organisms may be absent from the annotations. As a result, detections from biologically meaningful MBARI classes may be treated as background or unmatched candidates during training, producing misleading gradients for the 499-to-32 mapping. CatBoostRanker reranking improved the Public score to 0.2749, showing that candidate-level score calibration was useful. The final multi-scale WBF plus CatBoostRanker pipeline also achieved 0.2749 on the Public leaderboard, which did not further improve over CatBoostRanker-only reranking.

Table 3
Public leaderboard scores of the main methodology variants after the design change.

Method variant	Public score
Manual pre-NMS 499-to-32 mapping detector	0.2740
Trainable linear 499-to-32 mapping layer	0.0974
Trainable two-layer MLP mapping layer	0.0671
CatBoostRanker score reranking	0.2749
Multi-scale WBF + CatBoostRanker	0.2749

The official leaderboard reports separate Public and Private scores. Tables 4 and 5 summarize the displayed Public and Private leaderboard results.

Table 4
Public leaderboard results.

Team	Rank	Score
HSU Lab	1	0.3210
FAU CAAI	2	0.3035
Dmitry Kozlov	3	0.2801
Ours(WhiteBy)	4	0.2749

Table 5
Private leaderboard results.

Team	Rank	Score
HSU Lab	1	0.3042
FAU CAAI	2	0.2964
Dmitry Kozlov	3	0.2962
Dr_F	4	0.2808
masterSomebody	5	0.2650
Ours(WhiteBy)	6	0.2546

On the Public leaderboard, our team obtained a score of 0.2749 and ranked fourth among the displayed submissions. On the final Private leaderboard, the same submission obtained a score of 0.2546 and ranked sixth.

5. Conclusions

This paper presented an inference-oriented pipeline for positive-unlabeled marine object detection. The method retains the MBARI-315k YOLOv8 detector, maps its 499-class pre-NMS output to the 32 competition categories, and refines candidate ordering with multi-scale WBF and CatBoostRanker-based reranking. The final submission achieved a Public score of 0.2749 with rank 4 and a Private score of 0.2546 with rank 6. Future work should focus on reducing class-level errors, improving candidate-score calibration under incomplete annotations, and exploring pseudo-labeling to make use of high-confidence unlabeled objects.

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 authors used GPT-5.4 to translate the text into English and to improve the visual presentation of the method figure. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] Chrobak, L., Barnard, K., Katija, K., 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.

[2] 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.

[3] 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.

[4] M. Jian, N. Yang, C. Tao, H. Zhi, H. Luo, Underwater object detection and datasets: a survey, Intelligent Marine Technology and Systems 2 (2024) 9.

[5] Y. Yang, K. J. Liang, L. Carin, Object detection as a positive-unlabeled problem, arXiv preprint arXiv:2002.04672 (2020).

[6] S. Suri, S. Rambhatla, R. Chellappa, A. Shrivastava, Sparsedet: Improving sparsely annotated object detection with pseudo-positive mining, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 6770–6781.

[7] H. Wang, L. Liu, B. Zhang, J. Zhang, W. Zhang, Z. Gan, Y. Wang, C. Wang, H. Wang, Calibrated teacher for sparsely annotated object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 2519–2527.

[8] R. Solovyev, W. Wang, T. Gabruseva, Weighted boxes fusion: Ensembling boxes from different object detection models, Image and Vision Computing 107 (2021) 104117.

[9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical features, Advances in neural information processing systems 31 (2018).