Marine-Domain Semantic Transfer for Object Detection under Incomplete Annotations

Grzegorz Harańczyk[1]

[1]xhat, Krakow, Poland

Abstract

We describe our FathomNetCLEF2026 working-note solution for dense marine object detection under incomplete annotations, which ranked in the top 10 on the final private leaderboard. Many training images contain labeled organisms while visually similar unlabeled organisms remain present in the scene, making local validation behave like a positive-unlabeled setting. A competition-trained YOLOv8x detector initialized from a FathomNet Megalodon checkpoint reached a public leaderboard score of 0.0992 using five folds, test-time augmentation, and consensus ensembling. The largest improvement came from adding a pretrained MBARI-315k/FathomNet YOLOv8 detector as an inference-time marine-domain semantic donor. With conservative taxonomy mapping, donor/base score calibration, and class-wise merging, the private-best final submission reached public/private scores of 0.2527/0.2348. A slightly different margin-filtered variant achieved the highest public score, 0.2537, but did not improve the private score. A relatively small donor-related fraction of the final rows was associated with a much larger score gain, supporting the interpretation that semantic alignment mattered more than proposal volume. Our main finding is that calibrated marine-domain detector transfer was more effective in this challenge than our exploratory PU-adjacent training, artifact pruning, and additional competition-only ensembling branches.

Keywords

marine object detection, incomplete annotations, FathomNet, marine-domain semantic transfer, taxonomy mapping

1. Introduction

Marine image annotation is expensive, incomplete, and taxonomically difficult [1]. FathomNet-style image collections contain rich visual scenes where only a subset of visible organisms may be labeled. This makes the task positive-unlabeled (PU)-like: unlabeled regions cannot safely be treated as background, and recall-oriented detectors may be penalized by local validation labels even when they detect plausible organisms [2, 3, 4, 5]. The same setting also creates a domain-transfer problem for object detection: models must recognize fine-grained marine organisms, benthic growth forms, and acquisition artifacts that are poorly represented by generic detection vocabularies, motivating domain-specialized pretraining and transfer rather than relying only on the target annotations [1, 6, 7].

Our solution process began as a conventional competition object detection pipeline: build reliable folds, train stronger detectors, calibrate inference, and ensemble complementary models. This produced a useful but limited competition-trained detector. The largest score change occurred only after introducing an external marine-domain donor detector trained on the much larger MBARI/FathomNet label vocabulary. We therefore present the system as an empirical study of calibrated marine-domain semantic transfer under incomplete annotations, not as a new formal PU-learning algorithm.

The main contributions of this working note are:

- an empirical analysis of local-to-public validation gap under incomplete marine annotations;
- an empirical study of a disclosed donor-transfer pipeline that maps fine-grained MBARI labels into the 32 competition classes;
- submission-trace ablations showing that donor taxonomy and donor/base ranking dominated late-stage gains, while exploratory PU-adjacent and artifact-pruning branches were weaker in our experiments;

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

gharanczyk@gmail.com (G. Harańczyk)

0009-0004-6100-183X (G. Harańczyk)

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

- a transparent disclosure of external model usage and negative results.

2. Dataset and Task

The FathomNetCLEF 2026 task was part of LifeCLEF 2026 [8, 9]. The competition used 6,463 training images and 1,425 test images with 32 competition categories [1, 10]. The task is COCO-style object detection [11]. Public leaderboard scores are reported as mean average precision. Training labels are incomplete relative to visually dense marine scenes, while the official task description states that evaluation used a fully labeled test set [10]. This mismatch complicates validation: a detector that proposes extra biologically plausible organisms can look poor locally if those organisms are unlabeled, but useful on the public test set if they correspond to hidden annotations.

The image metadata also revealed a simple but important resolution mismatch. The train split was almost entirely composed of 1920×1080 and 3840×2160 frames, while the test split was mostly 1920×1080 but included a visible lower-resolution subset around 720×486 (Table 1). Thus, even input sizes such as 896 or 1280 substantially downsample many organisms, especially in the 4K training subset. This made small-object recall, thin structures, and resolution-dependent validation behavior recurring concerns during model selection.

Table 1
Dominant image resolutions in the official splits.

Split	Resolution	Images
Train	1920×1080	3,415
Train	3840×2160	3,048
Test	1920×1080	1,184
Test	720×486	197
Test	other resolutions	44

3. Method

3.1. Competition-Trained Base Detector

Our strongest competition-trained base was a YOLOv8x detector [12] initialized from a FathomNet Megalodon checkpoint [6] and fine-tuned on the official competition train set. We trained five class-stratified folds at image size 896 for 50 epochs, used standard YOLO augmentations, ran low-threshold inference with test-time augmentation, and merged fold predictions by consensus scoring with class-wise NMS at IoU 0.55. Test-time augmentation used the Ultralytics inference-time augmentation option. The `sum_norm` consensus mode scored each same-class cluster by summing fold scores and dividing by the five fold sources, with scores capped at 1.0. This model family reached public score 0.0992.

3.2. MBARI-315k Marine-Domain Donor

The donor branch used a pretrained MBARI-315k/FathomNet YOLOv8 detector [7] as an inference-time donor. Donor inference used image size 1280, confidence threshold 0.05, donor NMS IoU 0.70, and maximum 300 detections before merging. The detector predicts hundreds of fine-grained marine labels. We mapped a conservative subset of these labels to the 32 competition categories. Examples include Ophiuroidea to brittle star, Strongylocentrotus fragilis to urchin, Peniagone/Psolus/Scotoplanes to sea cucumber, Porifera/Hexactinellida to sponge, and Chionoecetes tanneri to crab.

The promoted donor map was `expanded_safe_v2`, which contains 298 accepted donor label/name patterns covering all 32 competition categories. After calibration, donor scores were multiplied by 1.50 and capped at 1.0 before final merging in the private-best final submission.

3.3. Final Merge and Post-Processing

The final submission merged the Megalodon-init 5-fold base and the single pretrained MBARI donor with class-wise NMS. Among the final submitted variants, the private-best branch used merge IoU 0.60, donor score scale 1.50, and kept the top 100 detections per image. A separate public-best branch used donor score scale 1.625 followed by a deterministic black-margin filter that removes or clips boxes overlapping image margins. That branch was visually cleaner and achieved the highest public score, but it was not the best private submission. Figure 1 summarizes the promoted pipeline and separates the official-train fine-tuning path from the inference-time donor-transfer path.

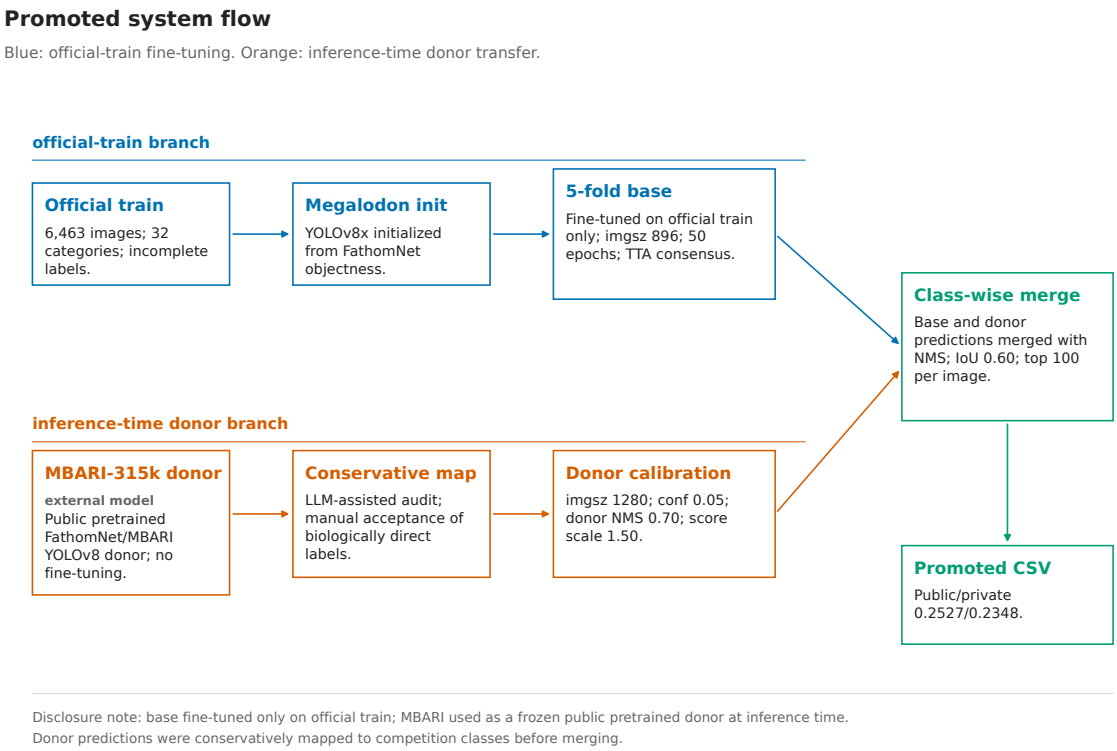

Figure 1: Promoted system flow. The Megalodon-initialized YOLOv8x branch was fine-tuned only on the official competition train split, while the MBARI-315k/FathomNet detector was used as an inference-time donor whose labels were conservatively mapped to competition categories.

Table 2
Core configuration of the private-best final-submission branch. Ultralytics defaults were used unless listed explicitly.

Component	Setting
Base detector	YOLOv8x initialized from FathomNet Megalodon checkpoint
Base training	5 class-stratified folds, 50 epochs, image size 896, batch 8, seed 42
Base optimizer/regularization	Ultralytics optimizer auto, lr0 0.01, lrf 0.01, weight decay 0.0005, AMP enabled
Base augmentation	mosaic 0.5 with close-mosaic 10, HSV jitter, translation 0.1, scale 0.5, horizontal flip 0.5
Base inference	low confidence threshold 0.0001, test-time augmentation, max 100 detections per fold/image
Base consensus	5-fold `sum_norm` consensus, class-wise merge IoU 0.55
Donor detector	pretrained MBARI-315k/FathomNet YOLOv8 used only at inference time
Donor inference	image size 1280, confidence 0.05, donor NMS IoU 0.70, max 300 detections before merging
Donor mapping/calibration	`expanded_safe_v2` taxonomy map, score scale 1.50, cap 1.0
Final merge	base plus donor, class-wise NMS IoU 0.60, top 100 detections per image

4. Experiments

4.1. Leaderboard Progression

Table 3 summarizes the public-score progression. The jump from 0.0992 to 0.2433 changed our interpretation of the task: the main missing ingredient was semantic coverage from a marine-domain donor, not another small competition-only detector adjustment.

Table 3
Main public leaderboard progression. All scores are official public leaderboard mAP.

Branch	Public score	Interpretation
Pre-Megalodon ensemble	0.0791	best early competition-trained ensemble
Megalodon-init YOLOv8x consensus	0.0992	strongest competition-trained base
Megalodon + MBARI mapped donor	0.2433	largest donor-transfer gain
Expanded safe taxonomy map	0.2445	safe map expansion helped
Expanded safe v2, donor scale 1.25	0.2504	donor priority improved
Expanded safe v2, donor scale 1.50, IoU 0.60	0.2527	private-best final-submission family
Expanded safe v2, scale 1.625, IoU 0.55	0.2536	best scale at IoU 0.55
Expanded safe v2, scale 1.625, IoU 0.60, margin filter	0.2537	public-best optional branch

4.2. Study Design and Submission-Trace Ablations

The final system was not selected from a single external-model trial. Development involved many public-leaderboard probes supported by generated submission artifacts, distribution summaries, source-contribution reports, visual reviews, and artifact-pruning diagnostics. We used these variants as a sequence of hypothesis tests: first to improve the competition-trained detector, then to diagnose validation shift and behavior under incomplete labels, and finally to calibrate the MBARI donor branch. The public leaderboard was an important optimization signal precisely because local labels were incomplete: unlike local folds, the fully labeled evaluation set was less likely to penalize recall-preserving detections of real but locally unlabeled organisms. We therefore treat the submission history as an engineering trace and place more weight on large, interpretable steps, artifact audits, and qualitative/source analyses than on tiny late-stage public differences. Late-stage calibration used public leaderboard feedback over donor-map, score-scale, confidence, merge-IoU, top-cap, and margin-filter variants. We therefore treat sub-0.001 public-score differences among endgame variants as engineering selection noise; the main conclusions rest on larger, interpretable transitions and supporting diagnostics. Table 4 summarizes the main submission-trace ablation families and the conclusions that shaped the final system.

4.3. Offline Versus Public Behavior

The local/offline metrics reveal a substantial validation gap (Table 6). Class-stratified validation for the Megalodon-init base was very high, but the public score was much lower. A D2/source-holdout stress split collapsed to the public-score regime and became a better disaster gate, but still did not predict the MBARI donor-transfer gain. The mapped donor was first evaluated locally as a standalone diagnostic; after the final selection process, a standalone top-100 donor export was also scored on the public leaderboard. We use this donor-only export only to separate donor signal from merge effects; it was not a promoted branch because the merged donor/base system was stronger on both D2 source-holdout mAP and public leaderboard score. We derived source groups from the official image URLs. For images hosted at `hurlimage.soest.hawaii.edu`, the path segment after `Localization_Scenes/` identifies the acquisition source; for example, `D2-EX2301-09`. The D2/source-holdout diagnostic assigned all 534 training images from `hurlimage.soest.hawaii.edu/D2-EX2301-09` (1,023 annotations) to validation fold 0 and the remaining 5,929 training images to fold 1. The source-holdout Megalodon baseline used the same recipe as the base detector but was trained on the non-D2 fold and evaluated

Table 4
Major submission-trace ablation families considered during development.

Family	Variants and observation
Competition-only detectors	YOLO/RT-DETR ensembles, Megalodon-init folds, TTA → Megalodon-init 5-fold consensus was the best base but saturated near 0.0992.
Validation/source stress tests	class-stratified CV, grouped folds, D2 source-holdout → local metrics were useful as gates but poor final rankers for donor variants.
PU-adjacent training and suppression	pseudo-labeling, hard negatives, class thresholds → exploratory branches were plausible locally, but weaker than recall-preserving donor/base ranking in our experiments.
Taxonomy mapping	safe MBARI map, expanded map, Zoantharia probe → conservative biological mapping mattered more than broad label inclusion.
Donor/base calibration	donor scale sweep, confidence sweep, per-class multipliers → scale 1.50 with merge IoU 0.60 was the private-best final submission, while scale 1.625 was public-best among submitted endgame variants.
Post-processing and allocation	artifact pruning, black-margin filter, top-300, refill, WBF → visual cleanup helped interpretation; broad geometry/ranking changes were flat or risky.

Table 5
Public/private status of selected final submissions.

Submission	Public score	Private score	Final status
v2 scale 1.50, IoU 0.60	0.2527	0.2348	private-best among final submitted variants; top 10 final leaderboard
v2 scale 1.625, IoU 0.60, margin filter	0.2537	0.2346	public-best optional post-processing branch
v2 scale 1.625, IoU 0.55	0.2536	0.2344	close scale-1.625 reference

on the held-out D2 fold. Donor diagnostics used the frozen MBARI detector on the same held-out D2 images; the base-plus-donor diagnostic merged those D2-fold predictions and evaluated them against the D2 annotations. We used this split as a source-shift stress test because class-stratified folds can mix visually correlated frames from the same expedition or dive across train and validation.

Table 6
Offline metrics and public scores for key milestones.

Milestone	Offline protocol	mAP	AP50	AP75	AP_S	AP_M	AP_L	Public
Early YOLO baseline	legacy 5-fold CV	0.1956	0.3286	0.2119	0.0490	0.0571	0.2267	0.0384
Early YOLO baseline	date/time grouped CV	0.0720	0.1289	0.0698	0.0097	0.0284	0.0854	0.0299
Megalodon base	class-stratified 5-fold + TTA	0.4102	0.5259	0.4496	0.0254	0.0998	0.4816	0.0992
Megalodon base	D2 source-holdout	0.0718	0.1871	0.0411	0.0044	0.0531	0.1366	0.0992
MBARI mapped donor	D2 source-holdout; standalone top-100 export	0.0566	0.1093	0.0476	0.0051	0.0342	0.1053	0.2369
Megalodon + MBARI mapped	D2 source-holdout	0.0761	0.1421	0.0723	0.0108	0.0513	0.1300	0.2433
Selected private-best (v2 scale 1.50, IoU 0.60)	public/private leaderboard	–	–	–	–	–	–	0.2527

Two observations are central. First, class-stratified validation overestimated transfer: local mAP 0.4102 corresponded to public score 0.0992. Second, adding the MBARI donor changed D2 mAP only modestly, from 0.0718 to 0.0761, while public score increased from 0.0992 to 0.2433. This is consistent with the donor addressing a hidden-test semantic coverage problem that competition-only validation labels did not fully measure.

4.4. Negative and Flat Probes

Several branches were informative despite weak public results. Donor downweighting to scale 0.75 scored 0.2380, and increasing donor confidence to 0.07 scored 0.2413; both were worse than the mapped-donor baseline. Artifact pruning and black-margin cleanup improved visual quality but were public-flat or private-flat. Top-300 public probing scored the same as the top-100 version, suggesting the evaluator effectively used the normal top detections. Weighted box fusion and simple per-class calibration were also flat or weaker than the global donor-scale branch. Simple PU pseudo-labeling did not become a

winning path.

Table 7
Selected baselines and negative or flat probes. Public results are official public leaderboard mAP unless private is explicitly shown; approximate rows summarize development probes.

Probe	Public result	Interpretation
Initial YOLO/RT-DETR ensemble baseline	0.0791	useful pre-Megalodon competition-trained baseline
Initial ensemble added to Megalodon base	0.0995–0.0998	little residual diversity after Megalodon-init
Megalodon proposals + ROI classifier	about 0.04–0.052	domain pretraining helped as initialization, not as a frozen proposal pipeline
Donor scale 0.75	0.2380	donor was under-ranked
Donor confidence 0.07	0.2413	stricter donor filtering removed useful detections
Expanded map + Zoantharia	0.2444	broad taxonomy addition was slightly worse than safer map
Artifact pruning / donor artifact pruning	about 0.2435	visually plausible cleanup did not explain main gain
Per-class donor calibration v1	0.2529	weaker than one global donor scale
Top-300 public probe	0.2536	no gain over normal capped submission
Public-best margin filter	0.2537 public / 0.2346 private	visually clean, not private-best
Donor image size 1536	0.2335	higher input resolution hurt transfer in this branch
WBF / coordinate fusion	not promoted	geometry changes looked poorly justified for this source pair

5. Analysis

5.1. Donor Contribution

The donor changed only a minority of final rows. In the initial 0.2433 branch, 89.5% of final rows were base-only, 6.0% overlapped donor and base, and 4.5% were donor-only. In the public-best 0.2537 branch, 122,391 final rows were base-only, 8,807 overlapped donor and base, 6,617 were donor-only, and 66 were unmatched by the simple attribution heuristic. Donor-related rows were therefore only about 11.2% of the final submission, yet the donor branch coincided with the largest score jump. This supports a semantic-value interpretation rather than a simple proposal-volume interpretation.

The most donor-influenced competition categories were biologically coherent: sponge, sea cucumber, brittle star, urchin, anemone, sea fan, sea star, sea pen, soft coral, squat lobster, crab, benthic worm, and sea squirt. These are exactly the types of taxa where an external MBARI/FathomNet vocabulary is expected to be more informative than a 32-class competition detector. Figure 2 visualizes this contrast between the large public score step and the modest donor-related row share.

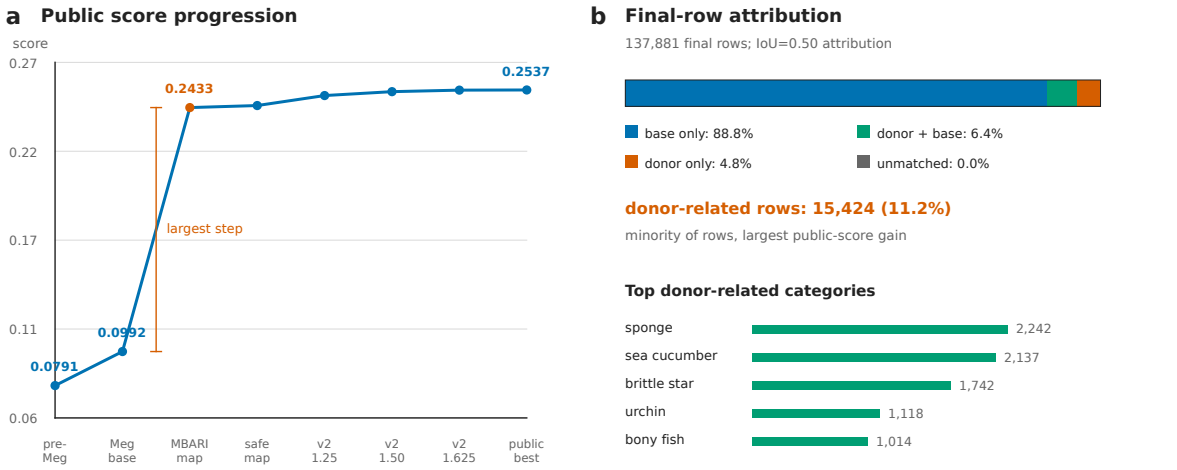

Figure 2: Public-score progression and final-row source attribution for the public-best margin-filtered branch. The donor branch accounts for only about 11.2% of final rows, but coincides with the largest public-score jump, supporting a semantic-coverage interpretation rather than a pure proposal-volume interpretation.

One representative donor-heavy review pattern helps explain the mechanism behind the donor gain. The base branch was not empty: it could already fill the top-100 cap and produced many brittle-star predictions. After donor/base merging, however, the final allocation shifted almost entirely toward donor-related rows: donor-only rows and same-class donor/base overlaps dominated the final brittle-star allocation, with most rows mapped from the MBARI label *Ophiuroidea*. This pattern illustrates both effects of the donor branch: it added detections not matched by the base and re-ranked or confirmed detections where the two models overlapped.

5.2. Taxonomy Audit

The conservative taxonomy map was important. We rejected broad or ambiguous labels such as equipment, sampler, trash, and several labels whose visual meaning could span multiple competition categories. A specific example was Zoantharia: adding it mostly increased anemone-like predictions and scored slightly below the safer map. The practical rule was to add labels only when the biological mapping to one competition class was direct and conservative. A map learned from non-D2 official-train overlaps was weaker than the hand-reviewed biological mapping, so the final system kept the manual taxonomy-audit route rather than treating observed overlaps as sufficient. We used large language model (LLM) assistance to draft and audit candidate donor-to-competition taxonomy mappings, but accepted mappings were manually reviewed and kept only when the biological correspondence was conservative.

5.3. Visual Inspection

Because local labels were incomplete, visual inspection was an important complement to offline metrics. We manually reviewed fixed image sets, prediction overlays, donor-heavy examples, and artifact-focused samples to understand whether public-score changes reflected plausible organism detections or obvious failure modes.

We also experimented with VLM-assisted, agentic visual-review tooling during development. The tooling prepared per-image dossiers, overlays, and candidate failure summaries to help direct human attention toward possible error modes, such as dense donor-heavy scenes, margin artifacts, infrastructure/HUD regions, and plausible below-cap organisms. These summaries helped prioritize manual review of donor-heavy fixed images such as dense brittle-star-like scenes, grouped visually similar acquisition artifacts for the later margin audit, and flagged risky taxonomy-expansion hypotheses such as broad anemone-like additions for separate leaderboard probes. We did not use VLM outputs as labels or as an automatic correction mechanism for the final submission. Instead, the VLM-assisted review acted as a triage aid for manual inspection and hypothesis generation.

Manual inspection confirmed both the value and the limitations of the donor. The donor often found plausible organisms missed by the base model, especially in dense benthic scenes. Persistent errors remained: missed nearby similar organisms, red camera artifacts, black margins, HUD/text overlays, and large habitat or structure boxes. Broad pruning was risky because dense marine scenes contain many real but low-score organisms.

The fixed-review examples included dense brittle-star fields, low-contrast soft-coral and substrate regions, homogeneous sea-cucumber scenes, mixed sponge/sea-pen/debris scenes, and colonized infrastructure. They motivated the qualitative failure taxonomy above, but were used only for triage and hypothesis generation, not as extra labels or as a replacement for leaderboard evaluation.

A black-margin artifact—horizontal black bands at the top or bottom of some test frames—was one concrete source-shift diagnostic. A margin audit found 189 of 1,425 test images with top or bottom blackish margins, while a train audit found no true examples among 6,463 images after manual review of the four detected candidates. On the scale-1.625 IoU 0.60 branch, the deterministic filter removed 1,211 rows and clipped 904 rows. The resulting submission reached public/private scores of 0.2537/0.2346, compared with 0.2527/0.2348 for the private-best scale-1.50 branch. We therefore treat margin filtering as a visually clean acquisition-artifact fix rather than as part of the main donor gain.

6. Positive-Unlabeled Discussion

Local validation and visual review were difficult to interpret because visible organisms in dense marine scenes can remain unlabeled in the official training annotations. Conceptually, a training image can contain one annotated organism and several nearby same-class organisms that are visually plausible to a human reviewer but absent from the label file. A detector proposal on those unlabeled organisms then has ambiguous status: it may be a biologically plausible positive but appears as a false positive under local labels. This phenomenon is the practical reason we treat the task as PU-like.

This framing follows positive-unlabeled learning [2, 3] and related object-detection studies on missing or sparse annotations [4, 5]. This affected both training and validation. Our selected branch does not modify the supervised detector loss or implement a formal PU risk estimator; instead, it addresses the practical effects of incomplete annotations indirectly through low-threshold inference, consensus merging, conservative top-100 ranking, marine-domain initialization, and an external marine-domain donor. These choices preserved plausible object candidates rather than aggressively suppressing them.

Our exploratory PU-adjacent experiments were less successful. Simple pseudo-labeling, hard-negative filtering, ROI-style rejection, and broad artifact pruning did not produce the main gain in this setting. Early pseudo-label branches improved some local reads but did not transfer to public submissions: representative PU v1/v2 submissions remained around 0.054–0.062 public score. Masking-based teacher-ignore training generated broad ignore regions but degraded the fold-0 baseline, and a baseline-matched broad ignore-loss rerun reached only 0.0865 local mAP versus 0.1183 for the corresponding supervised fold. A later consensus-PU student looked mildly useful locally but was public-negative. Infrastructure and hard-negative helper branches also produced either small split-sensitive gains or negative public checks. These results do not show that PU learning is inferior in general; they show that our simple PU-adjacent branches were less effective than domain-specific semantic transfer and donor/base calibration. More principled PU methods with higher-precision ignore regions or uncertainty-aware losses could still be useful in future work, especially if combined with the donor branch rather than replacing it.

7. External Resources and Disclosure

Following organizer guidance that public pretrained models and external resources should be transparently disclosed [10], we used two external FathomNet-related models (Table 8). The FathomNet Megalodon YOLOv8x checkpoint initialized the competition-trained base, which was then fine-tuned only on the official training images. The public FathomNet/MBARI-315k YOLOv8 detector was used as a frozen inference-time donor. We did not fine-tune this donor, download external image archives, add external labels to the official train split, pseudo-label external unlabeled images, or use hidden-test labels. Its predictions were mapped into the 32 competition categories and merged with the competition-trained base.

Because the donor model and challenge data both come from the broader FathomNet/MBARI ecosystem, we treat the reported gain as a competition-system result rather than as a controlled disjoint-source generalization study. Our source and resolution audits helped characterize the official challenge splits, but the MBARI-315k model card did not provide a complete training-image manifest for a test-set near-duplicate audit. This does not affect the disclosed use of the donor model, but it limits claims about transfer to fully independent marine image collections, since we could not fully audit near-duplicates.

Table 8
External resources used in the promoted system.

Resource	Role	Task-specific training by us	Inference-time use
FathomNet Megalodon YOLOv8x checkpoint	initialization for base detector	fine-tuned only on official train	yes, through the 5-fold base
FathomNet/MBARI-315k YOLOv8 detector	marine-domain donor	none	yes, mapped and merged with base

8. Limitations and Future Work

The taxonomy map was rule-based and expert-reviewed rather than learned from data; this improved interpretability but may miss safe labels. We did not implement a full principled PU detection algorithm. Visual analysis was targeted rather than exhaustive: the fixed review sets identified recurring failure families, but they do not provide a quantitative taxonomy of all error modes. Public leaderboard feedback was used during development, especially for late-stage donor/base calibration; as a result, small endgame score differences should be interpreted more cautiously than the large, interpretable transition from the competition-trained base to the donor-augmented system. The private-best final submission scored 0.2348 on the private leaderboard and ranked 9th.

Future work should explore systematic expansion of the donor-to-competition taxonomy map, better donor-class calibration, alternative marine-domain donors, small-object and top-100 allocation diagnostics, and principled PU training with ignore regions or uncertainty-aware losses. The visual-review process also suggests concrete follow-ups: a curated failure-case benchmark for dense donor-heavy scenes, margin-aware cropped inference for images with acquisition borders, quantitative measurement of VLM triage usefulness against human review, and donor-aware PU training that keeps semantic transfer while reducing loose structure boxes and artifact detections.

9. Conclusion

Our strongest result is the size and character of the MBARI donor gain. A carefully tuned competition-trained detector reached 0.0992 public score. Adding a pretrained MBARI-315k detector through conservative taxonomy mapping immediately reached 0.2433, and calibration raised the selected branch to 0.2527 public / 0.2348 private, ranking in the top 10 on the final private leaderboard. The PU framing remains relevant because it explains why local validation was misleading and why suppression was risky. The strongest path we found worked around this issue at inference time: it matched the detector's semantic vocabulary to the ocean and calibrated which donor and base detections survived the final cap. The evidence is therefore triangulated rather than based on one score alone: incomplete labels helped explain unreliable local validation, source and margin audits exposed train/test acquisition differences, the MBARI donor supplied missing marine semantics, and conservative taxonomy mapping determined which external categories transferred cleanly. This suggests that calibrated marine-domain detector transfer is a strong practical strategy for FathomNet-style object detection under incomplete annotations.

Code, Model, and Artifact Availability

The accompanying public artifact package is available at https://github.com/gharanczyk/fathomnet-cvpr-fgvc2026. It contains the training, inference, taxonomy-mapping, calibration, merging, and diagnostic scripts needed to reproduce the main submissions from repository-relative paths. The private-best CSV is included under `artifacts/final_submission/`, together with the `expanded_safe_v2` donor map, source-holdout fold files, promoted merge configuration, optional margin-filter reports, distribution summaries, and source-contribution reports used in this paper. The initial release does not redistribute trained checkpoint weights; full inference expects locally placed or externally downloaded checkpoints. For artifact-level reproduction, the package provides scripts to regenerate the final CSV from stored predictions, verify the released outputs, and compute file hashes. The package documents checkpoint provenance for the FathomNet Megalodon YOLOv8x and MBARI-315k YOLOv8 models, exact merge/export scripts, final submission artifacts, and the local path conventions needed to reproduce the full pipeline when weights are available.

Declaration on Generative AI

During preparation of this work, the authors used ChatGPT/Codex for drafting assistance, language editing, LaTeX formatting, and organization of experimental notes and artifacts. LLM assistance was also used to organize candidate taxonomy mappings; final mapping decisions were manually reviewed by the authors. Prototype VLM-assisted visual-review tooling was used only to triage image dossiers and candidate failure summaries for human inspection, not to create labels or automatically modify final predictions. All scientific claims, experiments, analyses, and final text were reviewed and edited by the authors, who take full responsibility for the content.

References

[1] 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. URL: https://doi.org/10.1038/s41598-022-19939-2. doi:10.1038/s41598-022-19939-2.

[2] J. Bekker, J. Davis, Learning from positive and unlabeled data: A survey, Machine Learning 109 (2020) 719–760. URL: https://doi.org/10.1007/s10994-020-05877-5. doi:10.1007/s10994-020-05877-5.

[3] Y. Yang, K. J. Liang, L. Carin, Object detection as a positive-unlabeled problem, CoRR abs/2002.04672 (2020). URL: https://arxiv.org/abs/2002.04672. arXiv:2002.04672.

[4] M. Xu, Y. Bai, B. Ghanem, Missing labels in object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 1–10. URL: https://openaccess.thecvf.com/content_CVPRW_2019/html/Weakly_Supervised_Learning_for_RealWorld_Computer_Vision_Applications/Xu_Missing_Labels_in_Object_Detection_CVPRW_2019_paper.html.

[5] Y. Niitani, T. Akiba, T. Kerola, T. Ogawa, S. Sano, S. Suzuki, Sampling techniques for large-scale object detection from sparsely annotated objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6510–6518. URL: https://openaccess.thecvf.com/content_CVPR_2019/html/Niitani_Sampling_Techniques_for_Large-Scale_Object_Detection_From_Sparsely_Annotated_Objects_CVPR_2019_paper.html.

[6] FathomNet, FathomNet Megalodon Detector, Hugging Face model card, https://huggingface.co/FathomNet/megalodon, 2025. URL: https://huggingface.co/FathomNet/megalodon, Ultralytics YOLOv8x object detector trained on public FathomNet localizations, accessed 12 May 2026.

[7] FathomNet, MBARI Monterey Bay 315k YOLOv8, Hugging Face model card, https://huggingface.co/FathomNet/MBARI-315k-yolov8, 2024. URL: https://huggingface.co/FathomNet/MBARI-315k-yolov8, Object detection model from the MBARI collection, accessed 12 May 2026.

[8] 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.

[9] L. Chrobak, K. Barnard, 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.

[10] K. Barnard, L. Chrobak, FathomNetCLEF2026 @ LifeCLEF & CVPR-FGVC, https://kaggle.com/competitions/fathomnet-2026, 2026. Kaggle competition page, accessed 12 May 2026.

[11] 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, volume 8693 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 740–755. URL: https://doi.org/10.1007/978-3-319-10602-1_48. doi:10.1007/978-3-319-10602-1_48.

[12] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, Software, https://github.com/ultralytics/ultralytics, 2023. URL: https://github.com/ultralytics/ultralytics, AGPL-3.0 license; release package specifies Ultralytics 8.3.0 or newer.