

Improving Underwater Marine Species Classification through Imbalance-Aware Learning and Cross-Dataset Evaluation

Md Sazzad Hossain¹, Mosa. Farjana Eti¹, Md Saimul Haque Shanto², and Suzad Mohammad^{2,3}

¹ Department of Computer Science and Engineering, Institute of Science and Technology, Bangladesh

{m.sazzad.h3, etimourakkhiet}@gmail.com

² Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Bangladesh

{0424052002, 0423052030}@grad.cse.buet.ac.bd

³ Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Bangladesh

Abstract. Automated recognition of marine species from underwater imagery is an important tool for biodiversity monitoring, but training datasets are rarely balanced, and this imbalance is compounded by degraded visibility and inter-class similarity typical of underwater scenes. A model can therefore post high overall accuracy while failing on the classes that matter most, and strong performance on an internal test set does not guarantee reliable generalisation to independently collected images. This study presents a comprehensive evaluation of imbalance-aware classification on the 20-class AQUA20 dataset (8,171 images), comparing YOLO26 and ConvNeXt-Small under class balancing, Gentle Class-Balanced Focal Loss, high-resolution fine-tuning, test-time augmentation, and probability ensembling. The best balanced single model, a high-resolution ConvNeXt-Small trained with class-balanced focal loss, achieved **92.12% Top-1 accuracy** and **88.17% Macro F1**. On 9,738 shared-class images from a separately sourced external dataset, the frozen model retained meaningful discriminative ability without any adaptation, reaching **76.50% Top-1** and **75.28% Macro F1** — a noticeable drop that highlights the gap between within-dataset gains and cross-dataset generalisation. Overall, imbalance-aware fine-tuning improves within-dataset recognition, while robust generalisation to externally sourced marine imagery remains an open challenge.

Keywords: marine image classification; underwater species recognition; class imbalance; class-balanced focal loss; high-resolution fine-tuning; cross-dataset generalisation; external evaluation; explainable artificial intelligence.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

Cite as: Hossain, M.S., et al.: Improving Underwater Marine Species Classification through Imbalance-Aware Learning and Cross-Dataset Evaluation. In: Mahmud, M., et al. (eds.) Applied Intelligence and Informatics. AII 2026. CCIS, vol. 3206, pp. 325–339. Springer, Cham (2026).

1 Introduction

Covering roughly 71% of the Earth’s surface [11], the world’s oceans are vital to stabilizing global climate patterns, preserving ecosystems, and securing human welfare. Beyond their environmental function, oceans underpin trade, tourism, and food security for millions of people worldwide, making the systematic study and conservation of marine biodiversity a matter of both ecological and socioeconomic urgency. The AQUA20 dataset has 8,171 images across 20 aquatic categories, with a non-uniform class distribution [18,7]. As a result, a model can achieve strong overall performance while still underperforming on categories with fewer images, making per-category evaluation alongside the overall average necessary rather than optional. We use this dataset to evaluate several methods under a fixed protocol: train-only balancing, Gentle Class-Balanced Focal Loss, high-resolution continuation, test-time augmentation, and probability ensembling. Validation Macro F1 determines model selection, and the test set is kept separate from training and configuration decisions. The frozen primary model is further evaluated on an independent marine dataset without fine-tuning or adaptation, probing its generalisation beyond the source distribution.

We identify three principal factors underlying the difficulty of underwater species recognition: pronounced class imbalance across taxa, degraded visual fidelity intrinsic to underwater imagery, and the instability of model predictions under such conditions. This study combines imbalance-aware training with model interpretability to build an accurate and explainable system.

The main contributions are:

- a controlled comparison of data-level and loss-level imbalance strategies;
- an investigation of high-resolution continuation, test-time augmentation, and heterogeneous ensemble learning on classification performance;
- balanced reporting using Top- k , macro, weighted, and class-wise metrics;
- cross-dataset evaluation on an independent marine dataset;
- model interpretability analysis using Grad-CAM and LIME on selected AQUA20 test images.

Grad-CAM and LIME are used only to inspect prediction behaviour and representative errors, not as proof of model reasoning [20,15,1].

2 Related Work

Deep learning is now widely used for underwater fish recognition, but image degradation, variable backgrounds, and scarce labelled data remain persistent challenges [19,18]. Transfer learning has proved especially useful for small or imbalanced marine datasets [21]. Knausgård et al. [10] paired YOLO-based detection with a Squeeze-and-Excitation classifier, achieving strong source-dataset performance that dropped on a more variable temperate-fish dataset – underscoring how sensitive marine recognition systems are to shifts in data composition and acquisition conditions.

Transformer-based approaches offer another direction. Fish-TViT combined transfer learning, label smoothing, and Grad-CAM analysis for marine and freshwater fish classification [8]. FishAI used hierarchical vision-transformer classification and reported markedly weaker species-level than class-level performance [22], highlighting how hard fine-grained taxonomic recognition remains.

Class imbalance is typically tackled through resampling, class weighting, focal loss, and effective-number-based class-balanced learning [12,3], with class-aware training also explored for minority fish classes [2]. But gains here depend heavily on the dataset, metric, and how sampling and loss strategies interact. Cross-dataset evaluation and prediction interpretation matter too: performance can drop with a change of image source, and plausible-looking saliency maps don't necessarily reflect faithful reasoning [5,1].

AQUA20 itself is a recent 20-class aquatic image classification benchmark [7]. AquaSelect, a peer-reviewed study, examined selective prediction under an abstention protocol [4], while two recent preprints explored frozen foundation-model embeddings and inference-path modification under substantially different protocols [17,16] – none directly comparable to the fixed train-validation-test setting used here.

This work complements that literature by evaluating data-level balancing, class-balanced focal learning, high-resolution continuation, test-time augmentation, probability ensembling, qualitative explainability, and source-only external evaluation within one consistent experimental campaign.

3 Problem Statement

Marine species recognition is complicated by poor illumination, turbidity, suspended material, colour distortion, occlusion, and visually similar categories [18,7]. AQUA20 adds a substantial class-distribution imbalance: several categories contain considerably fewer training images than the dominant classes [7]. Under these conditions, a model may achieve high overall accuracy while remaining unreliable for under-represented categories.

Evaluation on a held-out partition from the same image source also does not establish robustness under changes in camera, background, lighting, or water conditions. While post-hoc explanations might flag which image regions drive a prediction, but that doesn't mean the model is using genuine biological or causal logic [5,20,15,1]. The real-world challenge is building a system with high overall accuracy that doesn't sacrifice minority class performance, falls apart during source shifts, or overstates what its explanations actually prove

We formalise the classification task using

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N, \quad (1)$$

where x_i is an underwater image and $y_i \in \{1, \dots, 20\}$ is its AQUA20 class label. The objective is to learn a classifier f_θ that estimates $p_\theta(y = c \mid x)$, with prediction

$$\hat{y} = \arg \max_c p_\theta(y = c \mid x). \quad (2)$$

4 Methodology

4.1 Dataset and Training Distributions

AQUA20 contains 8,171 underwater images across 20 species categories [7], publicly available on Hugging Face [6]. A stratified split (seed 42) produced 5,247 training, 1,312 validation, and 1,612 test images; this split and class order were fixed throughout the study.

Two training distributions were compared to study class imbalance. The original distribution used all 5,247 training images. In the Aug-200 distribution, classes with fewer than 200 samples were augmented up to 200, while larger classes were left unchanged, giving 7,284 training images overall. Minority-class augmentation used random cropping (82–100% scale) followed by resizing, horizontal flipping (50% probability), rotation (-18° to $+18^\circ$), and mild brightness (0.82 – $1.18\times$), contrast (0.85 – $1.20\times$), colour (0.80 – $1.22\times$), and sharpness (0.85 – $1.25\times$) adjustments. Horizontal flipping was treated as label-preserving since AQUA20 labels encode organism/category identity rather than orientation. Balancing was applied only to the training partition; validation and test images were left untouched.

Table 1: AQUA20 train, validation, and test split counts.

Class	Train	Val.	Test	Class	Train	Val.	Test
coral	1,249	313	348	rayfish	305	77	95
crab	34	9	11	seaAnemone	712	178	221
diver	41	10	13	seaCucumber	28	7	10
eel	128	32	41	seaSlug	63	16	20
fish	1,760	440	538	seaUrchin	92	23	29
fishInGroups	218	54	72	shark	57	14	19
flatworm	40	10	13	shrimp	18	4	11
jellyfish	78	19	25	squid	19	5	10
marine dolphin	16	4	10	starfish	133	33	40
octopus	16	4	10	turtle	240	60	76

Initial experiments used 224×224 inputs; the selected ConvNeXt checkpoint was later continued at 384×384 . No balancing or synthetic expansion was applied to validation or test data.

4.2 Experimental Configurations

The campaign evaluated two pretrained YOLO26 classifiers [9] and several ConvNeXt-Small configurations [13]. Table 2 lists each configuration’s architecture, training distribution, role, and input resolution.

F01–F02 compare YOLO model scale under the same Aug-200 protocol. F04–F05 give the closest loss-focused comparison, though their training budgets differ. Comparisons involving F03, F06, and F08 change more than one factor at once and should be read as whole-configuration comparisons rather than strict ablations.

Table 2: Experimental configurations used in the final campaign.

ID	Architecture	Training data	Mechanism	Input
F01	YOLO26m-cls	Aug-200	Native YOLO classification training	224×224
F02	YOLO26s-cls	Aug-200	Native YOLO classification training	224×224
F03	ConvNeXt-Small	Original	Class-balanced focal loss ($\gamma = 1$)	224×224
F04	ConvNeXt-Small	Aug-200	Cross-entropy	224×224
F05	ConvNeXt-Small	Aug-200	Class-balanced focal loss ($\gamma = 1$)	224×224
F06	ConvNeXt-Small	Original	F03 checkpoint continuation at high resolution	384×384
F07	Frozen F06	–	Two-view test-time augmentation	384×384
F08	Frozen F06 and F01	–	Weighted probability ensemble	384×384 + 224×224

4.3 Class-Balanced Focal Objective

For a sample with true class y , cross-entropy is

$$\mathcal{L}_{\text{CE}} = -\log(p_y), \quad (3)$$

where p_y is the predicted probability of the correct class.

The imbalance-aware objective follows established focal-loss and effective-number formulations [12,3]. For a class with n_y training samples, the raw effective-number weight is

$$\tilde{w}_y = \frac{1 - \beta}{1 - \beta^{n_y}}. \quad (4)$$

Weights are normalised to a mean of one:

$$w_y = \frac{\tilde{w}_y}{\frac{1}{C} \sum_{c=1}^C \tilde{w}_c}, \quad (5)$$

with $C = 20$. This normalisation sits on top of the standard effective-number weighting of [3] purely to keep the loss scale comparable to unweighted cross-entropy – it is not a novel contribution. The resulting objective is

$$\mathcal{L}_{\text{CBF}} = -w_y(1 - p_y)^\gamma \log(p_y), \quad (6)$$

with $\beta = 0.999$ and $\gamma = 1.0$. “Gentle,” as used in the experiment archives, refers only to this moderate focal exponent; the loss components themselves are not new.

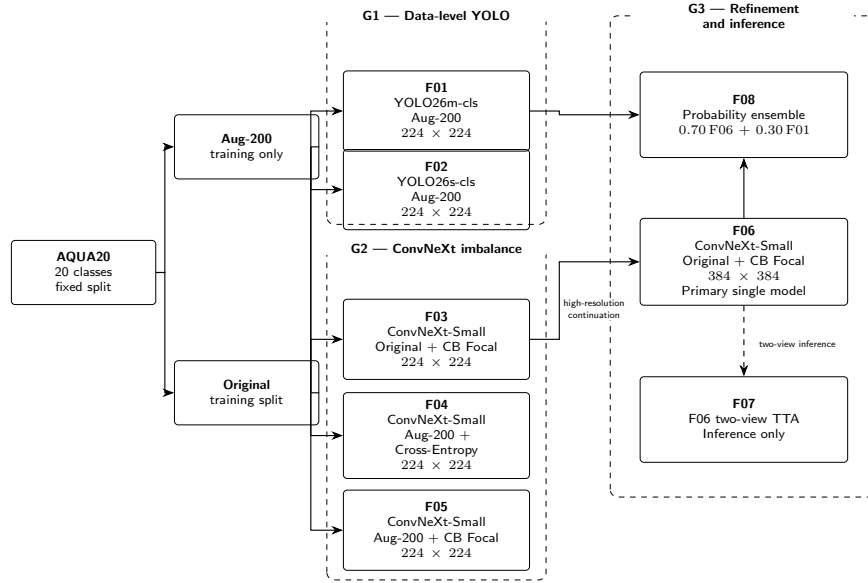


Fig. 1: Final experimental workflow.

4.4 Training and Model Selection

F01 and F02 used pretrained `yolo26m-cls.pt` and `yolo26s-cls.pt`, respectively, with batch size 64, up to 200 epochs, early-stopping patience of 30, initial learning rate 0.01, weight decay 5×10^{-4} , automatic mixed precision, and seed 42 (`optimizer=auto`; the resolved runtime optimiser was not logged).

F03 and F04 used AdamW (batch size 16, learning rate 1×10^{-5} , weight decay 1×10^{-4} , up to 40 epochs); F05 used the same settings with a 25-epoch cap. Cosine annealing was used throughout the ConvNeXt branch.

F06 continued from the validation-selected F03 checkpoint at 384×384 , batch size 8, learning rate 5×10^{-6} , weight decay 1×10^{-4} , up to 15 epochs.

Checkpoint selection used native Ultralytics fitness for F01/F02 and validation Macro F1 for F03–F06; the test split was touched only after the checkpoint or inference rule was locked. Run histories, selected checkpoints, best epochs, and integrity hashes were archived.

Computational resources and runtime. F06’s original training ran on a single NVIDIA Tesla P100-PCIE-16GB (Kaggle), taking 183.48 min over nine epochs (≈ 3.06 h, 20.39 min/epoch on average). For deployment context, the frozen F06 checkpoint was separately benchmarked on an NVIDIA Tesla T4 16 GB (FP32, batch size 1, 384×384): after 50 warm-up iterations, 200 timed passes gave a mean latency of 18.11 ms (median 18.17 ms, p95 18.36 ms), or ≈ 55.22 images/s. F06 has 49.47M parameters; peak GPU memory was 234.47 MB, with ≈ 34.75 MB incremental inference allocation. These figures cover forward-pass cost only, excluding I/O and data transfer.

4.5 Multi-Seed Robustness Evaluation

To gauge sensitivity to training randomness, additional runs used three seeds (42, 123, 2026) on the same fixed train/val/test split (split seed 42 was not regenerated).

For the original-distribution comparison, both cross-entropy and Gentle Class-Balanced Focal were run across all three seeds; the existing F03 result was reused as the seed-42 focal run. The 384×384 configuration was likewise run across all three seeds, reusing F06 as the seed-42 result, while the seed-123 and seed-2026 runs were initialised from their respective validation-selected 224×224 focal checkpoints.

All robustness runs selected checkpoints by validation Macro F1 before touching the test split. Results are reported as mean $\pm$ sample standard deviation ($n - 1$) across the three seeds.

4.6 Inference Refinements

F07 applied deterministic two-view inference to the frozen F06 model, averaging probabilities from the original image and its horizontal flip:

$$\mathbf{p}_{F07} = \frac{\mathbf{p}(x) + \mathbf{p}(\text{flip}(x))}{2}. \quad (7)$$

F08 combined the class-aligned probabilities of F06 and F01:

$$\mathbf{p}_{F08} = 0.70 \mathbf{p}_{F06} + 0.30 \mathbf{p}_{F01}. \quad (8)$$

Candidate weights were screened on the validation split; the final 0.70/0.30 rule was chosen by validation Macro F1 (with Macro Recall and Top-1 as tie-breakers) and locked before test evaluation.

4.7 External Evaluation

The frozen F06 model was evaluated on the Sea Animals Image Dataset [14] (13,711 images, 23 folders), decoded in RGB, resized to 384×384 , and normalised with ImageNet statistics. A predefined semantic mapping identified 15 classes shared with AQUA20 (9,738 images); Table 3 gives the full correspondence and Table 4 lists unmatched classes.

Class-Mapping Criteria External classes entered the strict shared subset only with a predefined AQUA20 counterpart. Exact or near-lexical matches were mapped directly; non-identical names were mapped only where the correspondence was clearly specific (e.g. *seaSlug*–*Nudibranchs*, *rayfish*–*Sea Rays*, *turtle*–*Turtle_Tortoise*). *fishInGroups* was kept separate from *Fish* since it encodes group context rather than the same class. Unmatched classes were excluded from the strict shared subset.

Table 3: External-to-AQUA20 class mapping criteria.

AQUA20 class	External class	Mapping basis
coral	Corals	Exact lexical match
crab	Crabs	Exact lexical match
eel	Eel	Exact lexical match
fish	Fish	Lexical match (singleton fish only)
jellyfish	Jelly Fish	Lexical / spacing variant
marine_dolphin	Dolphin	Terminology variant, same taxon
octopus	Octopus	Exact lexical match
rayfish	Sea Rays	Terminology variant, same taxon
seaUrchin	Sea Urchins	Lexical / plural variant
shark	Sharks	Lexical / plural variant
shrimp	Shrimp	Exact lexical match
squid	Squid	Exact lexical match
starfish	Starfish	Exact lexical match
turtle	Turtle_Tortoise	Broad mapping (tortoise + turtle folder)
seaSlug	Nudibranchs	Taxonomic-synonym mapping

Table 4: Classes excluded from the strict shared-class subset.

AQUA20-only (no external match)	External-only (no AQUA20 match)
diver, fishInGroups, flatworm, seaAnemone, seaCucumber	Clams, Lobster, Otter, Penguin, Puffers, Seahorse, Seal, Whale

An overlap audit compared all 8,171 AQUA20 images against the 13,711 external images via SHA-256 and perceptual hashing: no exact duplicates were found, and 34 candidate near-duplicate pairs were flagged and documented separately.

E01A used the original 20-class F06 output space as the primary external evaluation; E01B restricted and renormalised probabilities to the 15 shared classes as a diagnostic assuming known taxonomy overlap; E01C compared confidence and entropy behaviour between shared and external-only images. No external training, fine-tuning, or model selection was performed.

5 Results and Discussion

5.1 Internal AQUA20 Results

All internal results come from the unchanged 1,612-image AQUA20 test split, evaluated with Top-1/Top-5 accuracy, Macro Precision/Recall/F1, Weighted F1, class-wise scores, and confusion matrices. Macro F1 is prioritised for its equal weighting across classes.

ConvNeXt configurations outperformed the evaluated YOLO26 classifiers in balanced performance. Among the 224×224 ConvNeXt models, F03 gave the best Macro Precision and Macro F1, though it and F04 used different training distributions, so this is not a clean loss-only comparison.

Table 5: AQUA20 test performance (%).

Configuration	Top-1	Top-5	Macro P	Macro R	Macro F1	Weighted F1
F01: YOLO26m-cls + Aug-200	84.68	98.95	83.31	73.99	75.99	84.42
F02: YOLO26s-cls + Aug-200	83.44	97.95	80.91	65.26	70.39	82.84
F03: ConvNeXt-Small + CB focal	91.63	99.69	90.79	86.93	87.96	91.57
F04: ConvNeXt-Small + Aug-200	91.38	99.63	89.32	84.32	86.00	91.26
F05: ConvNeXt-Small + Aug-200 + CB focal	91.13	99.75	89.75	82.65	84.35	90.98
F06: ConvNeXt-Small HR384	92.12	99.75	90.43	87.44	88.17	92.08
F07: F06 + two-view TTA	92.12	99.69	89.88	86.53	87.40	92.08
F08: F06–F01 ensemble	92.25	99.88	90.67	87.02	87.94	92.18

F04 and F05 offer the closer loss comparison, despite differing training budgets. F05 lowered Macro Recall (84.32% $\rightarrow$ 82.65%) and Macro F1 (86.00% $\rightarrow$ 84.35%) – a configuration-level effect, not evidence against class-balanced focal loss generally.

F06 was the strongest balanced single model: 92.12% Top-1 and 88.17% Macro F1 – 0.50 and 0.21 points above F03, respectively. Since F06 also changed checkpoint continuation, learning rate, batch size, and training budget, the gain cannot be attributed to resolution alone.

F06 was further tested with two inference-level refinements. Two-view TTA (F07) did not improve over F06, while the F06–F01 probability ensemble (F08) achieved the highest Top-1 accuracy (92.25%) but did not improve Macro F1. Therefore, F06 remained the primary balanced model, with F08 reported as a complementary ensemble.

5.2 Multi-Seed Robustness Analysis

Across three seeds, cross-entropy and Gentle Class-Balanced Focal gave near-identical mean test Macro F1 ($87.22 \pm 1.62\%$ vs. $87.21 \pm 0.84\%$). Thus, the focal objective did not improve mean test Macro F1, but showed lower seed-to-seed variability.

Relative to the 224×224 Gentle Class-Balanced Focal configuration, the 384×384 continuation increased mean Macro Recall from 85.71% to 86.92% and mean Macro F1 from 87.21% to 87.87%, corresponding to gains of 1.21 and 0.66 percentage points, respectively. The Macro-F1 standard deviation decreased from 0.84 to 0.49 percentage points; this reduction is specific to Macro F1 and does not imply greater stability for every metric.

Validation results differed more clearly: cross-entropy achieved $87.39 \pm 0.87\%$ mean validation Macro F1, compared with $89.48 \pm 0.55\%$ for Gentle Class-Balanced Focal. Thus, although mean test Macro F1 was essentially unchanged, the focal configuration showed stronger balanced-class validation performance and lower validation seed-to-seed variability.

Table 6: Multi-seed ConvNeXt-Small results (% , mean $\pm$ SD).

Configuration	Top-1	Top-5	Macro P	Macro R	Macro F1	Weighted F1
ConvNeXt-Small + CE	91.52 $\pm$ 0.66	98.90 $\pm$ 0.40	91.64 $\pm$ 1.14	85.29 $\pm$ 1.56	87.22 $\pm$ 1.62	91.38 $\pm$ 0.69
ConvNeXt-Small + Gentle CB Focal	91.69 $\pm$ 0.16	99.69 $\pm$ 0.00	90.38 $\pm$ 0.83	85.71 $\pm$ 1.10	87.21 $\pm$ 0.84	91.61 $\pm$ 0.15
ConvNeXt-Small HR384 + Gentle CB Focal	91.79 $\pm$ 0.38	99.81 $\pm$ 0.06	90.06 $\pm$ 0.73	86.92 $\pm$ 1.11	87.87 $\pm$ 0.49	91.73 $\pm$ 0.37

5.3 Calibration Analysis

The frozen F06 model was assessed on the test set with 15 equal-width confidence bins, yielding ECE = 0.0175, NLL = 0.2511, and Brier score = 0.1200. Mean top-1 confidence (93.68%) slightly exceeded top-1 accuracy (92.12%), indicating mild over-confidence of approximately 1.56 percentage points. As Fig. 2 shows, most predictions cluster in the high-confidence region where confidence and accuracy align well; a few sparse low-confidence bins deviate more and warrant caution. No post-hoc calibration was applied – these are descriptive results for the frozen model.

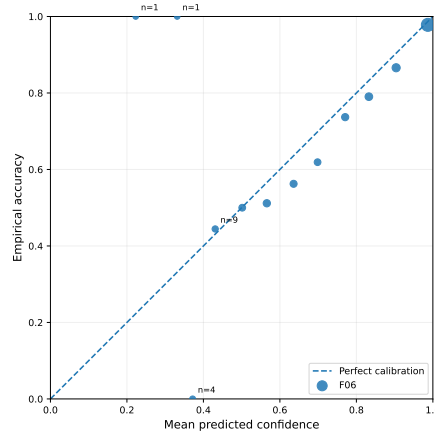


Fig. 2: Reliability diagram of F06 on AQUA20 test data (15 bins). ECE=0.0175.

5.4 External and Known–Unknown Evaluation

The frozen F06 model, without retraining or adaptation, was tested on 9,738 images across 15 classes shared between AQUA20 and the Sea Animals Image Dataset [14].

Raw 20-class evaluation gave 76.50% Top-1 and 75.28% Macro F1, down 15.62 and 12.89 percentage points from internal F06. Restricting outputs to the 15 shared classes raised these to 78.68% and 76.02%, but this assumes prior knowledge of the shared taxonomy and is therefore treated only as a diagnostic.

The 3,973 external-only images showed lower mean confidence (0.642 vs. 0.847) and higher entropy (0.403 vs. 0.184) than shared images. Entropy-based unknown

Table 7: Source-only external evaluation (%).

Setting	Top-1	Top-5	Macro P	Macro R	Macro F1	Weighted F1
E01A: raw 20-class output	76.50	95.75	81.36	74.16	75.28	78.36
E01B: restricted 15-class output	78.68	96.77	80.49	76.71	76.02	79.01

detection was only moderate (AUROC 0.720, unknown-AUPR 0.569, FPR 0.830 at 95% recall). F06 should therefore be treated as a closed-set classifier, not a validated unknown-rejection system. Typical confusions included puffers to fish, lobster to shrimp, and whale or seal images to marine dolphin.

5.5 Context within AQUA20 Literature

Table 8 situates these results among reported AQUA20 studies. Differences in validation, supervision, coverage, and model-selection procedures rule out a strict cross-study ranking.

Table 8: Contextual AQUA20 literature comparison.

Study	Model/protocol	Acc.	F1	Comparison note
Fuad et al. [7]	ConvNeXt, 32×32 ; 6,559 train, 1,612 test	90.69	88.92	No separate validation split
Daga et al. [4]	AquaSelect fusion at 80% coverage	94.8	88.6 Macro	20% abstained or referred
Rost [17]	Frozen DINOv3 embeddings with logistic regression	90.5 ± 0.2	88.5 ± 0.5	Preprint; different protocol
Rost [16]	Frozen DINOv3 with circuit duplication	—	87.5 Macro	Preprint; semi-supervised protocol
Current F06	ConvNeXt-Small, CB focal, HR384 continuation	92.12	88.17 Macro	Direct comparison only with F08
Current F08	0.70 F06 + 0.30 F01	92.25	87.94 Macro	Direct comparison only with F06

AquaSelect’s headline result uses 80% coverage, while the Rost studies are preprints evaluated under different protocols. Accordingly, Table 8 is contextual only. Within the present fixed protocol, F08 leads on Top-1 while F06 retains the higher Macro F1.

5.6 Class-Wise Performance and Error Structure

Macro scores trailing weighted scores indicate uneven per-class performance. Figure 3 shows F06’s row-normalised confusion matrix and per-class F1 scores.

Most classes retained strong diagonal responses, with errors concentrated among visually similar categories: marine dolphin was confused with shark or fish-in-groups, while octopus was confused with eel, crab, or fish. Starfish, turtle, diver, fish, and rayfish achieved comparatively strong scores; marine dolphin performed worst, followed by octopus, squid, and shark. These weaker classes contain relatively few test images, so small changes in prediction counts can substantially affect their scores.

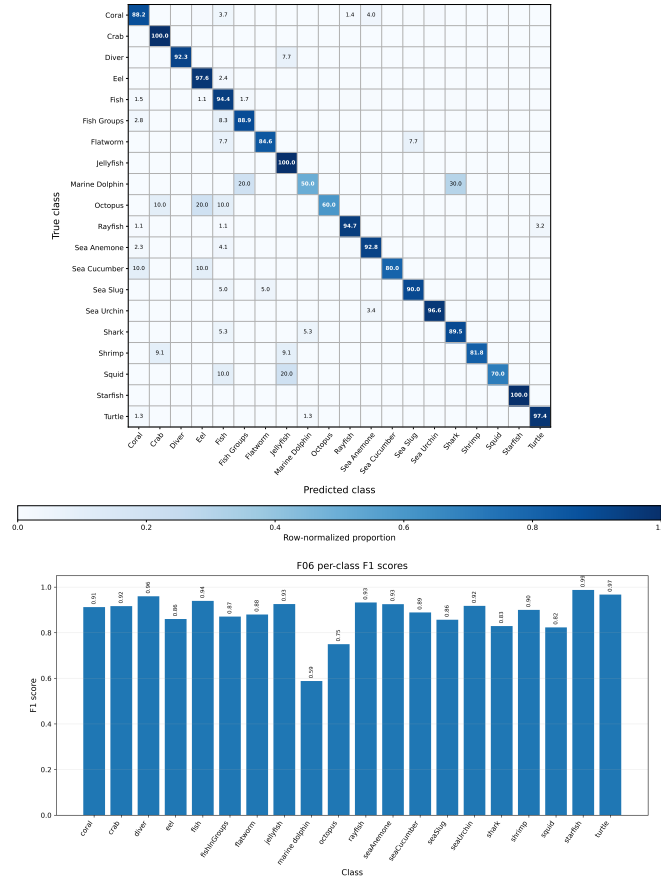


Fig. 3: Class-wise F06 performance: confusion matrix and per-class F1 scores.

5.7 Qualitative Explainability Analysis

Four cases in Fig. 4 span correct and incorrect predictions at high and low confidence. Case A (id 614, fish, confidence 0.999) shows Grad-CAM and LIME focusing on the fish body. Case B (id 1605, turtle, confidence 0.223) is correctly classified despite low confidence, with activation localised to the head and flippers rather than the surrounding water.

For Case C (id 577, true fish, predicted coral, confidence 0.995), activation spreads over the coral background rather than the camouflaged fish. For Case D (id 1061, true marine dolphin, predicted fishInGroups, confidence 0.622), activation falls on co-occurring fish rather than the dolphin, consistent with a bias toward the majority class. These patterns support using Grad-CAM and LIME as diagnostic tools rather than proof of correct model reasoning [20,15,1].

5.8 Discussion

The experiments show that imbalance-aware learning improves balanced recognition within the AQUA20 distribution, but internal gains do not directly translate

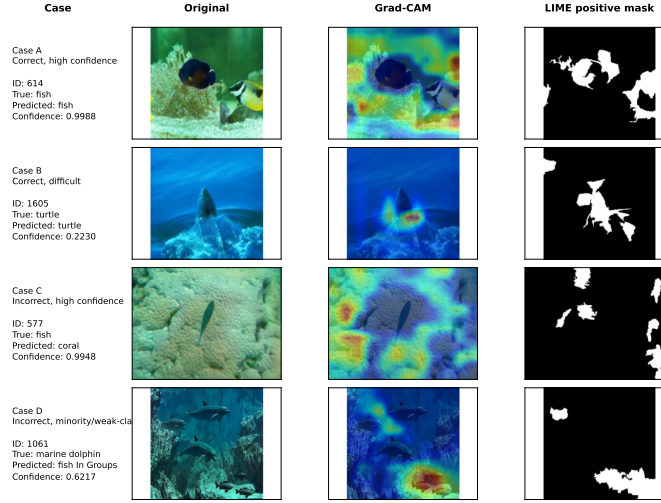


Fig. 4: Representative correct and incorrect Grad-CAM and LIME explanations for the frozen F06 model.

to external robustness. Multi-seed analysis indicates that Gentle Class-Balanced Focal Loss mainly improves validation behaviour and reduces Macro-F1 variability rather than increasing mean test Macro F1. High-resolution continuation provides the strongest balanced performance, while the ensemble gives only a marginal and statistically unsupported Top-1 improvement. External evaluation further highlights the remaining challenges of distribution shift and closed-set assumptions. Therefore, F06 is retained as the primary balanced model, with F08 reported as a complementary ensemble.

6 Conclusion and Future Work

This study evaluated imbalance-aware marine image classification on AQUA20 under a fixed train-validation-test protocol, comparing train-only balancing, Gentle Class-Balanced Focal Loss, high-resolution continuation, test-time augmentation, and probability ensembling.

Among the 224×224 ConvNeXt configurations, F03 achieved the strongest single-run Macro F1 without Aug-200 balancing. The subsequent multi-seed robustness analysis showed that Gentle Class-Balanced Focal Loss did not increase mean test Macro F1 over the matched cross-entropy baseline ($87.21 \pm 0.84\%$ versus $87.22 \pm 1.62\%$), although it reduced seed-to-seed Macro-F1 variability. The 384×384 high-resolution configuration achieved $91.79 \pm 0.38\%$ mean Top-1 accuracy and $87.87 \pm 0.49\%$ mean Macro F1 across three seeds, while the frozen F06 seed-42 run remained the strongest balanced single-model result at 92.12% Top-1 and 88.17% Macro F1. The F08 ensemble pushed single-run Top-1 accuracy slightly higher to 92.25%, though its Macro F1 stayed below F06, and two-view test-time augmentation did not improve on the selected model.

Source-only external evaluation showed a substantial drop from the internal AQUA20 result: 76.50% Top-1 accuracy and 75.28% Macro F1 on the raw 20-class output space. The known-unknown analysis reinforced this, showing the model remains a closed-set classifier with limited ability to separate unseen categories. Grad-CAM and LIME offered useful qualitative diagnostics without establishing causal reasoning.

Future work should extend the multi-seed robustness analysis to remaining configurations and alternative partitions, conduct a duplicate-image audit between datasets, and validate Grad-CAM and LIME explanations against localisation annotations or human judgment. Additional priorities include open-set recognition, distribution-shift calibration, ensemble stability analysis, multi-source training, and minority-class enrichment

Data and Code Availability. The AQUA20 dataset used in this study is publicly available through the official Hugging Face repository [6]. The Sea Animals Image Dataset used for external evaluation is publicly available through its original Kaggle repository [14]. The source code, fixed dataset split definitions, experiment configurations, class-order metadata, external taxonomy mapping, statistical-analysis scripts, and preserved prediction files are publicly available in the <https://github.com/1sazzad/AQUA20-Marine-Classification> project GitHub repository .

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems*. vol. 31, pp. 9505–9515 (2018)
2. Alaba, S.Y., Nabi, M.M., Shah, C., Prior, J., Campbell, M.D., Wallace, F., Ball, J.E., Moorhead, R.: Class-aware fish species recognition using deep learning for an imbalanced dataset. *Sensors* **22**(21), 8268 (2022). <https://doi.org/10.3390/s22218268>
3. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9268–9277 (2019). <https://doi.org/10.1109/CVPR.2019.00949>
4. Daga, M., Naole, S., Ramu, S.P.: AquaSelect: Learning when to abstain via score fusion for reliable underwater species classification. *Array* **30**, 100890 (2026). <https://doi.org/10.1016/j.array.2026.100890>
5. Folkman, L., Pitt, K.A., Stantic, B.: A data-centric framework for combating domain shift in underwater object detection with image enhancement. *Applied Intelligence* **55**, 272 (2025). <https://doi.org/10.1007/s10489-024-06224-0>
6. Fuad, T.R., Ahmed, S., Ivan, S.: AQUA20: A benchmark dataset for underwater species classification under challenging conditions. <https://huggingface.co/datasets/taufiktrf/AQUA20> (2025), dataset; accessed 15 July 2026
7. Fuad, T.R., Ahmed, S., Ivan, S.: AQUA20: A benchmark dataset for underwater species classification under challenging conditions. *Arabian Journal of Science and Engineering* (2026). <https://doi.org/10.1007/s13369-026-11231-z>

8. Gong, B., Dai, K., Shao, J., Jing, L., Chen, Y.: Fish-TViT: A novel fish species classification method in multi-water areas based on transfer learning and vision transformer. *Heliyon* **9**(6), e16761 (2023). <https://doi.org/10.1016/j.heliyon.2023.e16761>
9. Jocher, G., Qiu, J., Liu, M., Lyu, S., Akyon, F.C., Kalfaoglu, M.E.: Ultralytics YOLO26: Unified real-time end-to-end vision models (2026). <https://doi.org/10.48550/arXiv.2606.03748>, preprint
10. Knausgård, K.M., Wiklund, A., Sjørdalen, T.K., Halvorsen, K.T., Kleiven, A.R., Jiao, L., Goodwin, M.: Temperate fish detection and classification: A deep learning-based approach. *Applied Intelligence* **52**, 6988–7001 (2022). <https://doi.org/10.1007/s10489-020-02154-9>
11. Kossinna, E.: Die tiefen des weltmeeres. Tech. Rep. 9, Institut für Meereskunde, Berlin (1921), e. S. Mittler & Son, 70 pp.
12. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017). <https://doi.org/10.1109/ICCV.2017.324>
13. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022). <https://doi.org/10.1109/CVPR52688.2022.01167>
14. Raijin: Sea animals image dataset. <https://www.kaggle.com/datasets/vencerlanz09/sea-animals-image-dataset> (2023), kaggle dataset; accessed 15 July 2026
15. Ribeiro, M.T., Singh, S., Guestrin, C.: “Why Should I Trust You?”: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016). <https://doi.org/10.1145/2939672.2939778>
16. Rost, T.M.: Inference-path optimization via circuit duplication in frozen visual transformers for marine species classification (2026). <https://doi.org/10.48550/arXiv.2604.03428>, preprint
17. Rost, T.M.: Label-efficient underwater species classification with logistic regression on frozen foundation model embeddings (2026). <https://doi.org/10.48550/arXiv.2604.00313>, preprint, version 2
18. Saleh, A., Laradji, I.H., Konovalov, D.A., Bradley, M., Vazquez, D., Sheaves, M.: A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports* **10**, 14671 (2020). <https://doi.org/10.1038/s41598-020-71639-x>
19. Saleh, A., Sheaves, M., Rahimi Azghadi, M.: Computer vision and deep learning for fish classification in underwater habitats: A survey. *Fish and Fisheries* **23**(4), 977–999 (2022). <https://doi.org/10.1111/faf.12666>
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017). <https://doi.org/10.1109/ICCV.2017.74>
21. Xu, X., Li, W., Duan, Q.: Transfer learning and SE-ResNet152 networks-based for small-scale unbalanced fish species identification. *Computers and Electronics in Agriculture* **180**, 105878 (2021). <https://doi.org/10.1016/j.compag.2020.105878>
22. Yang, C., Zhou, P., Wang, C.S., Niu, Z., Zhu, L., Yuan, Y., Shen, H.B., Fu, G.Y., Xu, X.W., Pan, X.: FishAI: Automated hierarchical marine fish image classification with vision transformer. *Engineering Reports* **6**(12), e12992 (2024). <https://doi.org/10.1002/eng2.12992>