SINERGI Vol. 30, No. 2, June 2026: 591-604
http://publikasi.mercubuana.ac.id/index.php/sinergi
http://doi.org/10.22441/sinergi.2026.2.022

Benchmarking YOLOv8 and vision transformers for intelligent fish monitoring in aquaponics and controlled aquarium environments

Tresna Dewi[1]*, Yurni Oktarina[1], Sri Rezki Artini[2], Gita Ayu Juliantika[1], Jhony Satria[3]
[1]Department of Electrical Engineering, Politeknik Negeri Sriwijaya, Indonesia
[2]Department of Civil Engineering, Politeknik Negeri Sriwijaya, Indonesia
[3]Koi Agro Farm, Indonesia

Abstract
Sustainable aquaculture requires reliable and accurate fish monitoring systems capable of operating across heterogeneous environmental conditions. Conventional monitoring approaches are labor-intensive and prone to human error, while recent advances in deep learning have enabled vision-based automation for aquatic environments. Convolutional object detectors such as YOLO and emerging Vision Transformer (ViT) models have demonstrated promising performance; however, most existing studies remain limited to single-environment evaluations and rarely address energy-constrained, real-world deployment. To bridge this gap, this study presents a systematic benchmark of YOLOv8 and ViT across two complementary settings: a controlled aquarium environment and a solar-powered, off-grid aquaponics system. The proposed framework integrates 1080p CCTV video acquisition, dataset annotation and augmentation, and standardized training and evaluation using COCO metrics. Experimental results show that ViT consistently outperforms YOLOv8 in detection accuracy and prediction stability across both environments. ViT achieves 99.73% accuracy in the controlled aquarium and ≥99.6% accuracy performance (99.68–99.73%) in aquaponics, while YOLOv8 records 87.90% accuracy in the aquarium and 93.92–97.92% across aquaponics fish classes, exhibiting higher sensitivity to background clutter. Statistical validation using McNemar's test ($p < 0.001$) confirms that these differences are statistically significant. Beyond accuracy, the results reveal a trade-off between robustness and computational efficiency. ViT provides superior resilience under occlusion and glare, whereas YOLOv8 offers faster inference suitable for real-time operation on resource-limited edge devices. End-to-end deployment on a solar-powered NVIDIA Jetson Xavier NX demonstrates the feasibility of continuous, off-grid aquaculture monitoring and provides practical guidance for context-aware model selection in intelligent aquaculture systems.

Keywords:
Aquaponics;
Deep Learning;
Fish Detection;
Vision Transformer;
YOLOv8;

Article History:
Received: August 15, 2026
Revised: January 1, 2026
Accepted: February 4, 2026
Published: June 8, 2026

Corresponding Author:
Tresna Dewi
Department of Electrical Engineering, Politeknik Negeri Sriwijaya, Indonesia
Email: tresna_dewi@polsri.ac.id

INTRODUCTION

The rising global demand for aquatic protein underscores the importance of advancing aquaculture systems capable of maintaining productivity while safeguarding environmental integrity [1, 2, 3, 4, 5]. Modern aquaculture is increasingly aligned with biodiversity protection, climate change mitigation, and circular economy strategies [2, 3, 4, 5]. Circular resource flows promote

wastewater reuse [3], while microbial-based interventions such as Bacillus cultures improve water quality and reduce environmental impact [6]. These measures contribute to sustainable resource management and ecosystem resilience [4].

Integrating renewable energy into aquaculture supports off-grid, low-carbon operations. Solar-powered aquaponics merges clean energy generation with integrated fish–plant production [7][8]. Hybrid solar–hydro energy management with deep learning forecasting optimizes resource use [7], while hybrid machine learning for photovoltaic (PV) prediction enhances energy reliability [8]. Agrivoltaic systems further apply AI to crop growth and harvest prediction [9]. For fish welfare, accurate dissolved oxygen (DO) prediction is critical; this has been advanced through long-term multivariate modeling [10] and transformer-based adaptive forecasting [11]. These capabilities support energy-autonomous aquaculture, though visual monitoring remains essential for biomass estimation, behavior analysis, and species identification.

Fish detection has evolved from geometry-based to deep learning approaches. Structured-light and multi-target tracking enabled precise spatial measurements but were constrained by environmental limitations [12]. CNN-based methods improved robustness, with YOLO frameworks enabling real-time detection in complex aquatic scenes [13, 14, 15]. Applications span deep-sea crawler imaging [14], morphometric classification with temporal modeling [16], and optimized CNN pipelines for high-speed recognition [17]. Transfer learning with SE-ResNet152 addressed small dataset limitations [18], while EfficientNet-based dual attention models enabled feeding behavior analysis [19].

YOLO adaptations have been deployed for fish passage monitoring [20], biomass estimation in turbid waters [21], morphometric measurement via attention-based keypoint detection [22], and multi-species counting in hydropower fishways [23]. Fisheries monitoring has also benefited from automated event detection [24]. Reviews consolidate deep learning best practices for fish habitat monitoring [12][13].

Segmentation approaches complement detection by enabling precise size estimation. Automatic segmentation for morphometric analysis has shown promise [25], while semi-supervised and weakly supervised strategies address data scarcity [26]. Underwater image enhancement improves detection reliability under turbidity and distortion [27]. Fully convolutional segmentation, such as SegNet, supports spatio-temporal identification in net cages [28].

Transformers have redefined detection methodologies. DETR introduced end-to-end transformer-based detection [29], followed by Vision Transformers (ViT) [30] for scalable recognition. Variants like Twins [31] improve efficiency and accuracy balance. In aquaculture, transformers have excelled in counting occluded or dense fish populations [32, 33, 34], drawing from crowd counting methods [35][36] and graph-modulated designs [37].

Instance segmentation advances, including BlendMask [38], Mask R-CNN [39], and underwater-specific WaterMask [40] deliver fine-grained localization. Foundation models such as Segment Anything (SAM) [33] and SAM-guided salient segmentation [35] offer label-efficient adaptation. Dense-scene modeling approaches, including TransCrowd [35] and Gramformer [37], further enhance fish school counting. Beyond aquaculture, YOLO variants in precision agriculture show transfer potential [41].

Despite this progress, three research gaps remain clear. First, most existing studies are restricted to single environments, limiting insights into conditions with distinct optical properties such as solar-powered aquaponics and controlled aquarium [13, 14, 15, 16, 17, 18, 19, 20]. Second, while YOLO has been widely applied, direct benchmarking of YOLO-based and Vision Transformer-based detectors is rare. Although recent ViT-related studies [31]–[33] demonstrate strong potential for handling dense and occluded fish populations, they do not systematically compare ViTs with YOLOv8 under heterogeneous aquaculture environments. Third, trade-offs between accuracy, inference speed, and energy efficiency remain underexplored, particularly for continuous monitoring in energy-autonomous systems [7, 8, 9, 10, 11].

This study benchmarks YOLOv8 and ViT models for intelligent fish monitoring in two complementary environments: a controlled aquarium and a solar-powered real-world aquaponics system. The evaluation compares detection accuracy, inference latency, and energy-aware throughput under identical protocols. By analyzing model architecture, detection performance, and deployment feasibility, this work establishes a dual-environment benchmarking framework that extends recent ViT-based aquatic vision studies [31, 32, 33] through direct comparison with

YOLOv8, providing practical guidance for intelligent off-grid aquaculture monitoring.

The remainder of this paper is organized as follows. Section II discusses the proposed method. Section III presents the results and discussion. Section IV concludes the paper, highlighting the effectiveness of the proposed approach.

METHOD

This study developed and evaluated a real-world, off-grid aquaponics platform powered entirely by photovoltaic solar energy, as illustrated in Figure 1. The system integrates a solar panel, solar charge controller (SCC), and a 12 V DC pump connected to a deep-cycle battery, ensuring continuous recirculation of water between the aquaculture tank and hydroponic grow beds. A weatherproof CCTV camera captures continuous video streams of fish activity, which are processed by a computer implementing advanced deep learning architectures. Specifically, YOLOv8 and Vision Transformer (ViT) were deployed for automated fish detection and classification, with a controlled aquarium developed as a benchmarking counterpart. The design ensures energy-autonomous operation under fluctuating solar irradiance, varying turbidity levels, and diverse fish behavioral patterns.

The complete methodological framework is further illustrated in Figure 2, which outlines the data and model pipeline. The process begins with dataset collection from both aquaponics and aquarium environments, followed by preprocessing steps such as resizing and augmentation to improve model robustness. The next stage involves building and training models (YOLOv8 and ViT), after which validation is performed to assess detection accuracy and generalization. A decision stage evaluates whether the model performance is satisfactory; if not, the workflow loops back to the improvement stage for refinement and retraining. Once performance criteria are met, the results are visualized and the model is prepared for deployment on an edge AI platform (Jetson NX) integrated into the solar-powered aquaponics system.

To ensure reproducibility and clarify how optimal performance was achieved, the main hyperparameters used to train YOLOv8 and the Vision Transformer (ViT) are summarized in Table 1. Both models were trained for an equal number of epochs, while optimizers and learning rates were selected according to best practices for CNN- and transformer-based architectures.

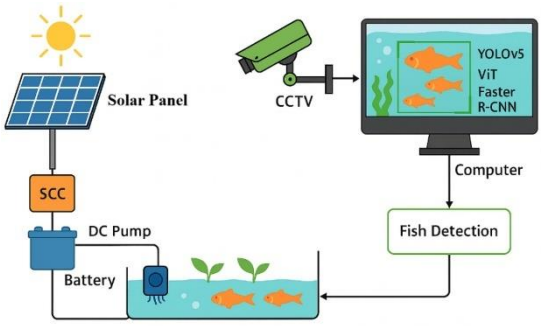


Figure 1. Solar-powered aquaponics setting is considered in this study

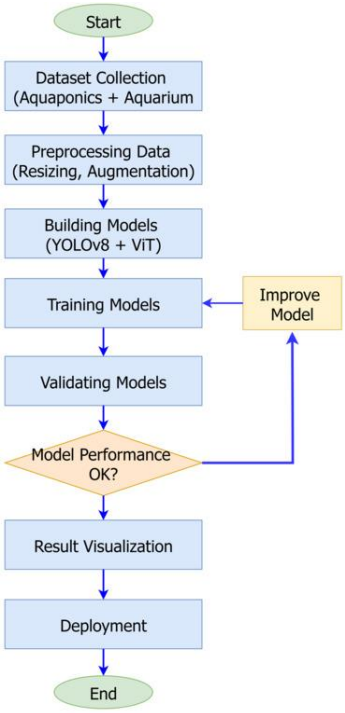


Figure 2. Methodology flowchart used in this study

Table 1. Training hyperparameters for YOLOv8 and Vision Transformer (ViT)

Hyperparameter	YOLOv8	Vision Transformer (ViT)
Batch size	16	32
Number of epochs	100	100
Optimizer	SGD	Adam
Learning rate	0.01	3×10^{-5}
Momentum	0.937	–
Weight decay	0.0005	0.0005
Dropout rate	–	0.1
Precision mode	Mixed precision	Mixed precision
Framework	PyTorch 1.13	TensorFlow/Keras 2.10

Material

The experimental platform was a fully autonomous aquaponics system (Figure 1) configured for continuous, off-grid operation. Electrical power was supplied by a 1300 Wp monocrystalline photovoltaic (PV) panel with a conversion efficiency of approximately 21.2%, selected for its high performance in both direct sunlight and partial shading conditions. The PV output was regulated via a maximum power point tracking (MPPT) solar charge controller, which ensured optimal energy harvesting across a range of solar irradiance levels. Energy storage was provided by a 12 V 100 Ah deep-cycle AGM battery, chosen for its deep-discharge tolerance, long cycle life, and maintenance-free operation. The stored energy powered a 12 V DC centrifugal pump responsible for recirculating water between the aquaculture tank and the hydroponic grow beds, maintaining both nutrient distribution and adequate dissolved oxygen levels. The fish considered in this study are Comet Goldfish (Komet), Manfish/Angelfish (Manfish/Malaikat), Goldfish (Mas Koki), Blue Gourami (Sepat Biru), dan Yellow Gourami (Sepat Kuning)

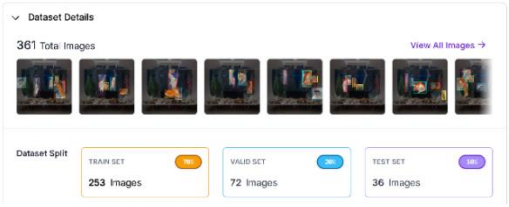
Visual data acquisition was achieved using a weatherproof IP CCTV camera with a resolution of 1080p and a frame rate of 30 frames per second, mounted above the fish tank. The installation height and angle were optimized to minimize glare and refraction from the water surface while ensuring maximum coverage of the fish activity zone. Continuous footage was captured under a variety of seasonal and lighting conditions, including clear skies, overcast weather, and periods of high turbidity.

(a) Aquarium dataset

(b) Dataset generated from Solar-powered aquaponics illustrated in Figure 1

Figure 3. Generated dataset

(a). Annotation process



(b). Dataset division

Figure 4. Labelling and annotation in Roboflow

The dataset used in this study was collected exclusively from a controlled aquarium environment and a real-world solar-powered aquaponics system. No external or publicly available datasets were used, as shown in Figure 3. All images underwent manual annotation using the Roboflow platform given in Figure 4, producing ground-truth bounding boxes and class labels in PyTorch format for YOLOv8 and COCO format for ViT.

Computational experiments were conducted on a high-performance deep learning workstation equipped with an NVIDIA RTX 2050 GPU with 4 GB VRAM, an Intel Core i5-13420H CPU, 16 GB of DDR5 RAM, and running Windows 11. Model development employed PyTorch 1.13 for YOLOv8 and TensorFlow/Keras 2.10 for ViT, and supporting libraries such as OpenCV and Matplotlib for preprocessing, visualization, and evaluation.

Methods

The data acquisition process involved continuous recording from the IP CCTV camera, with representative frames extracted at fixed intervals to prevent temporal redundancy and maintain diversity in fish poses, orientations, and behaviors. This ensured that the dataset remained balanced across fish species, environmental conditions, and activity types.

Table 2 shows dataset composition and train/validation/test split for aquarium and greenhouse aquaponic environments, where per-class distribution is reported as ground-truth instance counts derived from the test-set confusion matrix, reflecting the object-level nature of the detection task.

Table 2. Dataset composition, data split, and test-set per-class instance distribution

Component	Statistic	Value
Total images	All images	**1,364**
Train/Val/Test split	Train images	**1,057** (77.49%)
	Validation images	**205** (15.03%)
	Test images	**102** (7.48%)
Per-class distribution (test set, from confusion matrix)	Comet goldfish	**148** (19.55% of fish instances)
	Angelfish	**151** (19.95%)
	Goldfish	**156** (20.61%)
	Blue gourami	**157** (20.74%)
	Yellow gourami	**145** (19.15%)
	Total fish instances (test set)	**757** (100%)
	Background (true "background" column)	**26**

Preprocessing included normalization and format conversion, while data augmentation techniques were extensively applied to improve model robustness. Geometric augmentations involved random rotations up to ±15°, horizontal and vertical flips, and scaling up to ±20%, while photometric adjustments altered brightness, contrast, and hue to mimic variable daylight conditions. Gaussian noise was added to simulate suspended particles in turbid water, and random erasing was applied to represent occlusions caused by plant roots or overlapping fish.

Deep learning architectures are given in Figure 5. YOLOv8 was implemented with a CSPDarknet53 backbone, PANet-based feature aggregation, and multi-scale detection heads optimized for real-time inference (Figure 5a). ViT employed a patch embedding mechanism and multi-head self-attention layers to capture long-range feature dependencies, making it particularly effective under low-contrast or complex background conditions (Figure 5b).

Training for YOLOv8 was conducted over 100 epochs using the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01 and a momentum of 0.937. ViT fine-tuning was performed for 100 epochs with the Adam optimizer (learning rate = 3×10^{-5}, weight decay = 0.0005, dropout rate = 0.1). Mixed-precision training was enabled to optimize GPU memory usage and reduce computation time.

The evaluation protocol followed COCO metrics, reporting mAP@0.5 for primary detection accuracy, mAP@0.5:0.95 for performance over multiple IoU thresholds, and class-specific precision, recall, and F1-scores.

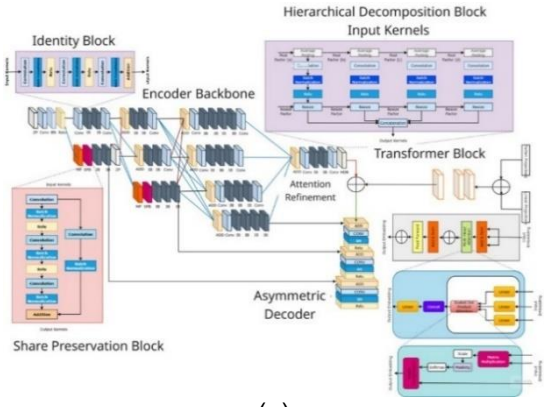
Inference latency, measured in milliseconds per frame, was also recorded to assess real-time viability. Testing was conducted under three controlled environmental conditions: clear water, moderate turbidity, and high turbidity, enabling assessment of model resilience to visual degradation.

Following benchmarking, the best-performing model was deployed on an NVIDIA Jetson Xavier NX edge AI device for real-world testing. The device processed live CCTV feeds locally, overlaying bounding boxes and class labels in real time and storing detection logs for long-term behavioral and population analysis. Field deployment confirmed the stability of the system under continuous operation, the adequacy of the solar power subsystem in meeting energy demands, and the robustness of detection performance under actual environmental conditions

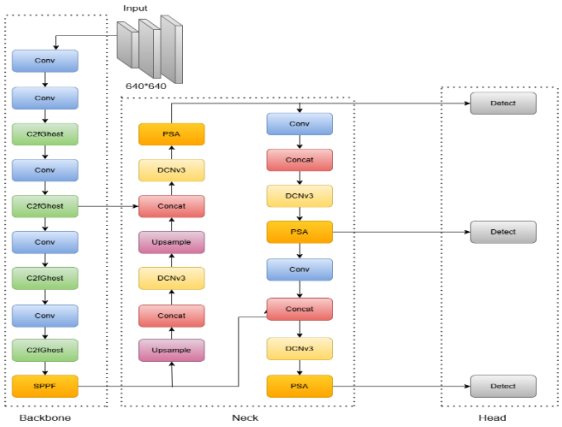
Mathematical Formulation

Let an input image be denoted as $I \in \mathbb{R}^{H\times W\times 3}$. The objective of the detection task is to predict a set of objects

$$\mathcal{O} = \{(b_i, c_i, s_i)\}_{i=1}^{N}, \quad (1)$$



(a)



(b)

Figure 5. ViT and YOLOv8 architecture implemented in this study

where $b_i$represents the bounding box coordinates, $c_i$the class label, and $s_i$the confidence score associated with the i-th object.

The Vision Transformer (ViT) processes the input image by partitioning it into non-overlapping patches of fixed size. Each patch is linearly embedded into a latent feature space and enriched with positional encoding before being passed through stacked transformer encoder layers based on multi-head self-attention and feed-forward networks. In this study, ViT is employed with a fixed configuration and adapted as a baseline for object recognition by learning global contextual representations. Model optimization follows a standard cross-entropy loss formulation over class predictions, without architectural or hyperparameter ablation.

YOLOv8 formulates object detection as a unified regression and classification problem and optimizes a composite loss function:

$$\mathcal{L}_{\text{YOLO}} = \lambda_{\text{box}}\mathcal{L}_{\text{CIoU}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}, \quad (2)$$

where $\mathcal{L}_{\text{CIoU}}$denotes the Complete Intersection over Union loss for bounding box regression, $\mathcal{L}_{\text{obj}}$represents the objectness loss, and $\mathcal{L}_{\text{cls}}$is the classification loss. The CIoU formulation jointly considers overlap area, center distance, and aspect ratio consistency, enabling more stable localization under cluttered aquatic backgrounds.

To evaluate deployment feasibility, inference efficiency is quantified using latency, throughput, and energy consumption. Let Tdenote the average inference time per frame (seconds) and $P$the average power consumption (watts). Energy per inference is defined as:

$$E = P \times T, \quad (3)$$

while frames per second (FPS) is computed as:

$$\text{FPS} = \frac{1}{T}. \quad (4)$$

RESULTS AND DISCUSSION

The experimental results evaluate YOLOv8 and ViT across both controlled datasets and real-world deployment in a solar-powered aquaponics system (Figure 6), comparing accuracy, precision, recall, and inference speed under varying water clarity and turbidity. As a novel contribution, this discussion presents the dual-environment benchmark of YOLOv8 and ViT. demonstrating how architectural differences between convolutional and self-attention models directly influence detection accuracy, robustness, and deployment suitability.

Table 3. Energy and Inference Performance Comparison

Model	InRes	Lat	FPS	P (W)	Energy (J)
YOLO	640 × 640	11	90.9	68	0.748
ViT	640 × 640	10	100	68	0.68

Figure 6. Solar-powered aquaponics experimental test-bed considered in this study

Table 3 shows the energy and inference performance considered in this study where InRes is input resolution, Lat is latecy (ms/frame), P is power (W), and Energy is Energy per Inference (J).

Experiment 1: Controlled Aquarium Evaluation

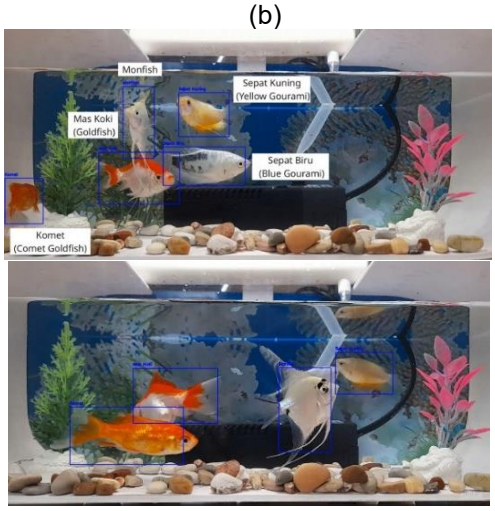
The initial experiment was conducted in a controlled aquarium environment to evaluate the performance of the proposed model under stable and well-regulated conditions, with the detection results for YOLOv8 presented in Figure 7a and those for ViT shown in Figure 7b.

The experiment on fish classification and object detection in a controlled aquarium environment was conducted to evaluate the performance of the Transformer model in recognizing five previously trained fish species. The testing process involved a camera directed at a mini aquarium, under stable lighting conditions and a relatively complex visual background.

The performance of Vision Transformer (ViT) and YOLOv8 was evaluated for multi-class fish recognition in a controlled aquarium environment across five target categories: Comet Goldfish, Angelfish, Goldfish, Blue Gourami, and Yellow Gourami. ViT achieved 99.73% overall accuracy, surpassing YOLOv8's 87.90%, reflecting its superior ability to extract discriminative features and capture global contextual dependencies. Figure 8 shows the confusion matrix results for fish-detection performance analysis in a controlled environment.

(a) YOLOv8

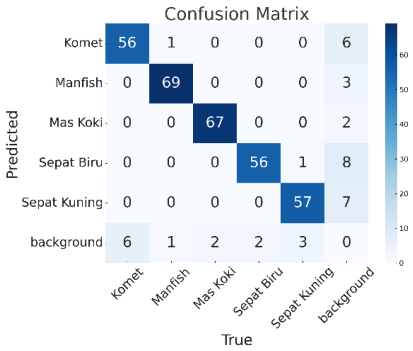
(b)



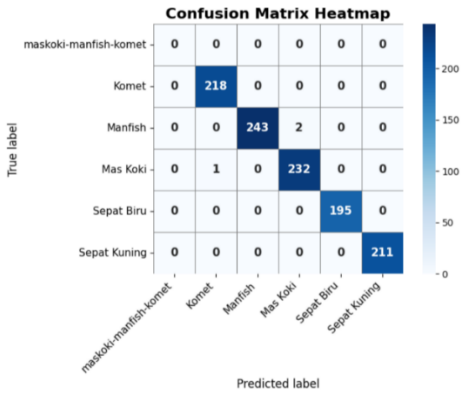
(c) ViT

Figure 7. Fish detection with ViT and YOLOv8 in controlled aquarium

Class-wise, ViT attained ≥0.99 precision for all predicted classes except the goldfish–angelfish–comet combined class (no predictions), with Comet Goldfish (0.995 precision, 1.000 recall), Angelfish (1.000, 0.992), and Goldfish (0.991, 0.996) among the strongest. YOLOv8 showed competitive results for Comet Goldfish (0.972, 0.958) and Angelfish (0.971, 0.971) but lower performance for Goldfish (0.966, 0.862) and Blue Gourami (0.934, 0.891), and completely failed to detect Yellow Gourami (0.000 across all metrics).



(a) YOLOv8



(b) ViT

Figure 8. Confusion matrix results for fish-detection performance analysis in a controlled environment

ViT's F1-scores exceeded 0.99 across all predicted classes, while YOLOv8's F1-scores reflected its precision–recall imbalance, with notable declines for Goldfish (0.911) and Blue Gourami (0.912). McNemar's test ($p < 0.001$) confirmed that the performance difference is statistically significant.

ViT's high accuracy and stability make it well suited for high-precision aquaculture monitoring in controlled environments. However, its higher computational demand may constrain real-time deployment without further optimization. In contrast, YOLOv8 remains advantageous for latency-sensitive, edge-based inference but requires targeted retraining, particularly for Yellow Gourami, to approach ViT's level of performance.

Experiment 2: Solar-Powered Aquaponics Evaluation

Following the controlled aquarium evaluation in Experiment 1, Experiment 2 assessed model performance in a real-world solar-powered aquaponics system. This setting introduced dynamic variables—fluctuating natural light, water turbidity changes, plant shadows, and surface reflections—absent in laboratory

conditions, thereby better representing the complexities of operational aquaculture monitoring.

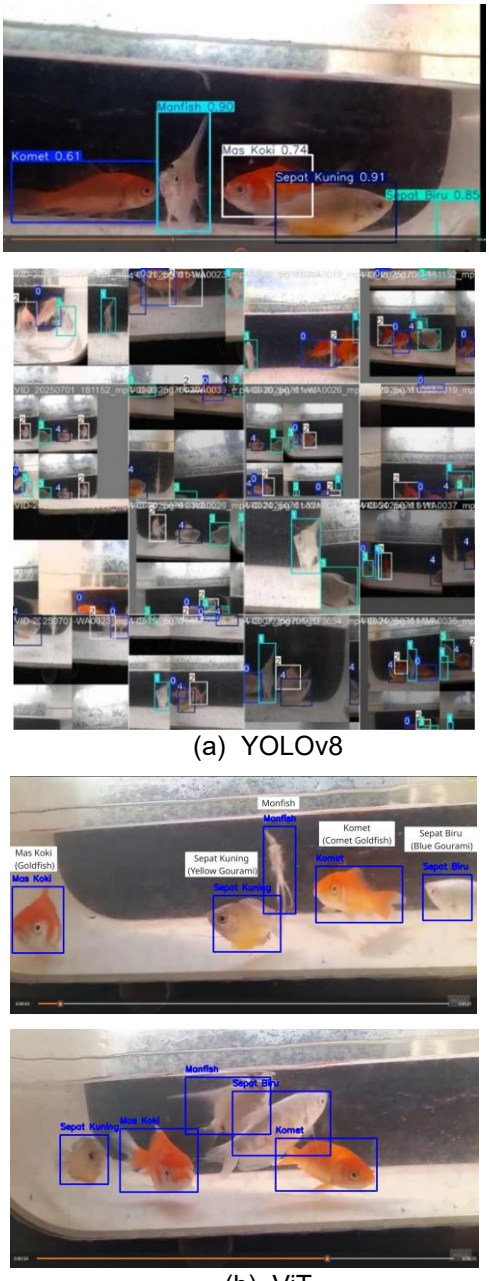
The system employed photovoltaic (PV) modules as the sole power source for autonomous, off-grid operation. Continuous fish detection and classification were performed using live video streams from overhead surveillance cameras above the aquaponics tanks. The objective was to examine how Vision Transformer (ViT) and YOLOv8, previously benchmarked in controlled conditions, adapt to the increased visual variability of outdoor environments.

This transition from a controlled to a renewable-energy-driven operational context provides a basis for evaluating robustness, scalability, and deployment feasibility in sustainable aquaculture. The results offer insights into trade-offs among accuracy, inference stability, and energy efficiency when applying advanced deep learning models in off-grid agricultural systems.

Figure 9 compares the performance of YOLOv8 (a) and Vision Transformer (ViT) (b) in a solar-powered aquaponics system for five target species: Comet Goldfish, Angelfish, Goldfish, Blue Gourami, and Yellow Gourami. In Figure 9a, YOLOv8 detects Comet Goldfish (0.61), Angelfish (0.90), Goldfish (0.74), Yellow Gourami (0.91), and Blue Gourami (0.85), but exhibits variable bounding box precision and misalignment under occlusion or reflective glare. The montage illustrates detection fluctuations across consecutive frames.

In Figure 9b, ViT detects Goldfish (1.00), Yellow Gourami (1.00), Angelfish (1.00), Blue Gourami (1.00), and Comet Goldfish (1.00), with consistent bounding box alignment and high stability despite occlusion and inter-species proximity. These results indicate that YOLOv8 is more sensitive to environmental noise, whereas ViT maintains superior spatial precision and classification consistency in real-world aquaponics deployment.

Figure 10 presents a comparative evaluation of YOLOv8 and Vision Transformer (ViT) classification performance in the solar-powered aquaponics environment across six categories: Comet Goldfish, Angelfish, Fancy Goldfish, Blue Gourami, Yellow Gourami, and background. As shown in Figure 10a, YOLOv8 achieves true positive counts of 139 (Comet Goldfish), 139 (Angelfish), 150 (Fancy Goldfish), 148 (Blue Gourami), and 141 (Yellow Gourami), corresponding to class-wise accuracies of 93.92%, 92.05%, 96.15%, 93.67%, and 97.92%, respectively.



(a) YOLOv8

(b) ViT

Figure 9. Fish detection with ViT and YOLOv8 in real-aquaponics-deployment

Misclassifications are predominantly assigned to the background class, resulting in frequent false background predictions when fish are partially occluded or affected by reflective glare. Error rates range from 2.08% for Yellow Gourami to 7.95% for Angelfish, indicating consistent but imperfect detection performance.

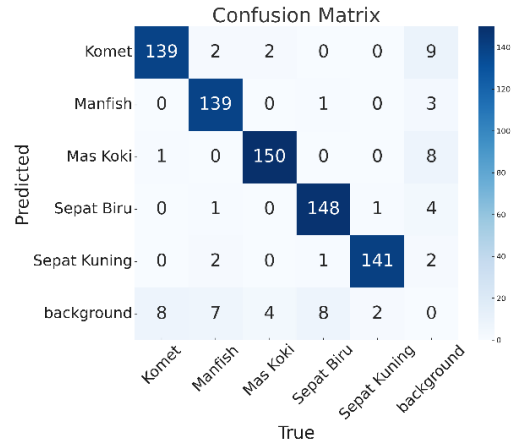
In contrast, Figure 10b demonstrates that ViT achieves substantially higher classification consistency across all evaluated species. True positive counts reach 1618 for Comet Goldfish, 1580 for Angelfish, 1562 for Fancy Goldfish, 1637

for Blue Gourami, and 1583 for Yellow Gourami. Class-wise accuracy reaches 100% for all species except Blue Gourami, which attains 99.68% accuracy with a marginal error rate of 0.32%, attributable to five Yellow Gourami instances misclassified as Blue Gourami.

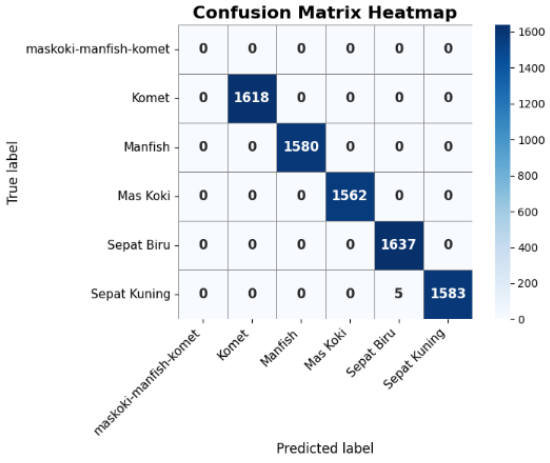
No background misclassification was observed in the test set under the evaluated experimental conditions. Overall, ViT maintains stable performance in the range of 99.68%–99.73%, indicating highly reliable classification with negligible inter-class confusion.

Table 4 summarizes the ablation analysis across model architecture, domain sensitivity, and error characteristics. The results indicate that detector architecture strongly influences performance, particularly under strict localization metrics. YOLO-based detectors achieve higher precision, recall, and mAP, reflecting superior bounding-box localization consistency, whereas Transformer-based models exhibit lower mAP values but maintain stable classification behavior across domains. Domain-shift analysis shows that YOLOv8 and YOLOv11 benefit from greenhouse deployment through improved precision and recall, while ViT performance remains comparatively stable with limited variation. Class-wise recall analysis confirms strong inter-species separability, with Angelfish and Fancy Goldfish achieving the highest recall due to distinctive morphology. In contrast, Blue Gourami and Yellow Gourami are more susceptible to background interference, consistent with their slender shape and narrow object-scale distribution.

Table 5 contrasts the controlled aquarium and solar-powered aquaponics environments. The aquarium offers stable indoor conditions with regulated lighting and minimal background variation, whereas the aquaponics system operates under variable natural illumination with additional challenges from turbidity, plant shadows, and surface reflections. While both environments involve the same five fish species, performance differences are evident.



(a) YOLOv8



(b) ViT

Figure 10. Model performance evaluation in real-aquaponics deployment

Table 4. Ablation Analysis across Model Architecture, Domain Sensitivity, and Error Characteristics

Factor	Configuration	Precision	Recall	mAP
Architecture	YOLOv8 (Aquarium)	0.94	**0.962**	0.968
	YOLOv8 (Greenhouse)	**0.959**	0.945	**0.977**
	YOLOv11 (Aquarium)	0.933	0.902	0.953
	YOLOv11 (Greenhouse)	0.947	0.945	0.973
	Transformer (Aquarium)	0.815	0.818	0.816
	Transformer (Greenhouse)	0.808	0.825	0.816
Domain Shift	Aquarium → Greenhouse (YOLOv8)	0.019	−0.017	0.009
	Aquarium → Greenhouse (YOLOv11)	0.014	0.043	0.02
	Aquarium → Greenhouse (Transformer)	−0.007	0.007	≈0
Class Sensitivity	Goldfish	—	**0.99**	—
	Angelfish	—	0.96	—
	Blue Gourami	—	0.97	—
	Yellow Gourami	—	0.93	—

Table 5. Comparison of controlled aquarium and solar-powered aquaponics implementation summary

Aspect	Controlled Aquarium	Solar-Powered Aquaponics
Environment	Indoor mini aquarium with stable, regulated conditions	Outdoor aquaponics system integrated with hydroponics
Lighting Conditions	Stable artificial lighting	Variable natural lighting
Visual Challenges	Minimal environmental noise; controlled background	Fluctuating turbidity, plant shadows, surface reflections
Power Source	Grid-powered laboratory setup	1300 Wp photovoltaic system with MPPT controller and AGM battery
Data Acquisition	HD camera mounted above aquarium; stable viewpoint	Weatherproof IP CCTV (1080p, 30 fps) mounted above tanks
Target Species	Comet Goldfish, Angelfish, Goldfish, Blue Gourami, and Yellow Gourami	Comet Goldfish, Angelfish, Goldfish, Blue Gourami, and Yellow Gourami
YOLOv8 Accuracy	87.90%	93.92–97.92% per-class accuracy (except background)
ViT Accuracy	99.73%	≥99.6% per-class accuracy (99.68–99.73%) (except Blue Gourami at 99.68%)
YOLOv8 Notable Limitations	Failed to detect Yellow Gourami, with lower precision and recall for Goldfish and Blue Gourami.	High background misclassification; reduced bounding box precision under glare
ViT Notable Strengths	Near-perfect precision and recall for all predicted classes	Stable detection despite occlusion and proximity; minimal inter-species confusion

In the controlled aquarium, YOLOv8 achieves 87.90% accuracy compared to 99.73% for ViT. In aquaponics, YOLOv8 reaches 93.92%–97.92% per-class accuracy (excluding background), whereas ViT attains ≥99.6% accuracy (99.68%–99.73%), with Blue Gourami exhibiting the lowest value. YOLOv8 performance is primarily constrained by missed detections and background confusion, whereas ViT maintains stable classification across both settings.

The observed performance differences can be attributed to architectural characteristics. ViT demonstrates strong robustness under glare, partial occlusion, and dense fish proximity due to its global self-attention mechanism, which captures long-range contextual relationships beyond local appearance cues. This enables more effective suppression of illumination artifacts and improved discrimination of closely spaced or overlapping fish, contributing to consistently high accuracy across environments.

Conversely, YOLOv8 relies on localized convolutional features and anchor-based detection, making it more sensitive to background noise from reflections, shadows, and water-surface disturbances, particularly in outdoor aquaponics scenarios. While YOLOv8 offers faster inference and lower computational overhead suitable for resource-constrained edge devices, ViT prioritizes robustness and accuracy at higher computational cost.

Overall, the dual-environment evaluation confirms that architectural design plays a decisive role in real-world aquaculture monitoring. Rather than a universally optimal solution, effective deployment depends on operational priorities, including accuracy requirements, hardware constraints, and energy availability. By linking architectural behavior with empirical performance under energy-autonomous conditions, this study

provides practical guidance for AI model selection in sustainable aquaculture systems.

Despite the strong performance observed, this study is limited by the number of evaluated fish species and data collection from a single controlled aquarium and one aquaponics site, which may constrain statistical generalizability. Extreme operating conditions such as nighttime monitoring and heavy rainfall were not included. Consequently, the results should be interpreted within the studied scope, motivating future multi-species, multi-location, and all-weather evaluations.

CONCLUSION

This study presented a comprehensive benchmark of Vision Transformer (ViT) and YOLOv8 models for intelligent fish monitoring across two operational settings: a controlled aquarium environment and a real-world, solar-powered aquaponics system. Experimental results consistently demonstrated that ViT outperformed YOLOv8 in detection accuracy and prediction stability across both environments. ViT achieved 99.73% accuracy in the controlled aquarium and ≥99.6% accuracy performance (99.68–99.73%) in the aquaponics system. In contrast, YOLOv8 attained 87.90% accuracy in the aquarium and 93.92%–97.92% across aquaponics fish classes, while exhibiting higher sensitivity to background misclassification under visually complex conditions. Statistical validation using McNemar's test ($p < 0.001$) confirmed that the observed performance differences were statistically significant. Beyond accuracy, this work emphasized deployment-oriented evaluation by incorporating inference latency, throughput, and energy consumption metrics. Deployment on a Jetson Xavier NX powered entirely by solar energy demonstrated that continuous, off-grid aquaculture monitoring is feasible under realistic operational constraints. These results highlight a fundamental trade-off: while YOLOv8 offers higher inference efficiency, ViT provides superior robustness and accuracy in complex aquatic scenes, particularly in scenarios involving reflections, occlusions, and background clutter. The key contributions of this study are threefold. First, it establishes a unified dual-environment benchmarking framework that bridges controlled indoor aquariums and outdoor aquaponics systems under a consistent experimental protocol. Second, it extends conventional accuracy-based evaluation by integrating energy-aware and latency-sensitive performance indicators, directly addressing edge deployment feasibility. Third, it provides a statistically grounded comparative baseline between ViT and YOLOv8, offering practical guidance for context-aware model selection in aquaculture monitoring applications. Future work will focus on improving computational efficiency through lightweight transformer architectures such as MobileViT and TinyViT, as well as hybrid CNN–Transformer backbones. Expanding the dataset to include additional fish species, varying turbidity levels, occlusion patterns, and nighttime conditions will further enhance generalizability. Energy-optimized edge AI deployment using platforms such as Jetson Orin Nano or Coral TPU, combined with pruning and quantization techniques, will improve long-term sustainability. Integrating multimodal IoT sensors for dissolved oxygen, pH, temperature, and water flow will enable closed-loop, data-driven aquaculture management. Extensions toward fish tracking, counting, and behavior analysis with continual learning will further strengthen intelligent aquaculture systems. Overall, this study provides a robust and deployment-validated foundation for accurate, efficient, and sustainable AI-driven monitoring in modern aquaculture.

Compared with prior studies, this work demonstrates improved robustness and deployment relevance for AI-based aquaculture monitoring. YOLO-based methods have shown strong real-time performance [13][14], but remain sensitive to turbidity, lighting variation, and background noise, while CNN-based approaches achieve high accuracy mainly in controlled settings and degrade under domain shifts [15, 16, 17]. Enhancements such as semi-supervised learning and image preprocessing [24][25] partially address these issues but still rely on local feature extraction. In contrast, transformer-based models [30, 31, 32] enable global contextual representation, offering greater resilience to occlusion and environmental complexity, consistent with recent aquaculture prediction studies [10][11]. The results of this study align with these trends, where ViT outperforms YOLOv8 in both controlled and real-world aquaponics environments. Unlike existing works that are largely limited to single-domain evaluations [18][19], this study introduces a dual-environment, solar-powered benchmarking framework, integrating detection accuracy with energy-aware deployment. Therefore, this work not only improves detection performance but also advances practical implementation toward sustainable, intelligent aquaculture systems.

ACKNOWLEDGMENT

The authors would like to acknowledge the Electrical Engineering Department, Politeknik Negeri Sriwijaya, and the Directorate General of Research and Development; Director of Research

and Community Service (DPRM), for funding support through Contract No. 187/C3/DT.05.00/PL-BATCH II/2025, dated 14 July 2025, under the Fundamental Research Scheme.

REFERENCES

[1] A. Ahmad, S. R. S. Abdullah, H. A. Hasan, A. R. Othman, and N. Ismail, "Aquaculture industry: Supply and demand, best practices, effluent and its current issues and treatment technology," *J. Environ. Manage*., vol. 287, p. 112271, 2021, doi: 10.1016/j.jenvman.2021.112271.

[2] K. Chary, A.-J. van Riel, A. Muscat, A. Wilfart, S. Harchaoui, M. Verdegem, and I. J. de Boer, "Transforming sustainable aquaculture by applying circularity principles," *Rev. Aquac.*, vol. 16, no. 2, pp. 656–673, 2024, doi: 10.1111/raq.12860.

[3] S. K. Das, B. Mondal, U. K. Sarkar, B. K. Das, and S. Borah, "Understanding and approaches towards circular bio-economy of wastewater reuse in fisheries and aquaculture in India: An overview," *Rev. Aquac*., vol. 15, no. 3, pp. 1100–1114, 2023, doi: 10.1111/raq.12758.

[4] D. D. Mizuta, H. E. Froehlich, and J. R. Wilson, "The changing role and definitions of aquaculture for environmental purposes," *Rev. Aquac*., vol. 15, no. 1, pp. 130–141, 2023, doi: 10.1111/raq.12706.

[5] E. Sala, J. Mayorga, D. Bradley, R. B. Cabral, T. B. Atwood, A. Auber, et al., "Protecting the global ocean for biodiversity, food and climate," *Nature*, vol. 592, no. 7854, pp. 397–402, 2021, doi: 10.1038/s41586-021-03371-z.

[6] V. Hlordzi, F. K. Kuebutornye, G. Afriyie, E. D. Abarike, Y. Lu, S. Chi, and M. A. Anokyewaa, "The use of Bacillus species in maintenance of water quality in aquaculture: A review," *Aquac. Rep*., vol. 18, p. 100503, 2020, doi: 10.1016/j.aqrep.2020.100503.

[7] T. Dewi, P. Risma, Y. Oktarina, S. Dwijayanti, E. N. Mardiyati, A. Br Sianipar, D. R. Hibrizi, M. S. Azhar, and D. Linarti, "Smart integrated aquaponics system: Hybrid solar-hydro energy with deep learning forecasting for optimized energy management in aquaculture and hydroponics," *Energy Sustain. Dev*., vol. 85, p. 101683, Apr. 2025, doi: 10.1016/j.esd.2024.101683.

[8] T. Dewi, E. N. Mardiyati, P. Risma, and Y. Oktarina, "Hybrid machine learning models for PV output prediction: Harnessing Random Forest and LSTM-RNN for sustainable energy management in aquaponic system," *Energy Convers. Manag*., vol. 330, p. 119663, Apr. 2025, doi: 10.1016/j.enconman.2024.119663.

[9] Y. Oktarina, Z. Nawawi, B. Y. Suprapto, and T. Dewi, "Towards ecological sustainability: Harvest prediction in agrivoltaic chili farming with CNN transfer learning," *Iraqi J. Agric. Sci*., vol. 55, no. 6, pp. 1910–1926, Dec. 2024.

[10] J. Hu, P. Wang, D. Li, and S. Liu, "A long-term multivariate time series prediction model for dissolved oxygen," *Ecol. Inform*., p. 102695, 2024, doi: 10.1016/j.ecoinf.2024.102695.

[11] D. Li, J. Hu, M. Li, and S. Zhao, "A long-term dissolved oxygen prediction model in aquaculture using transformer with a dynamic adaptive mechanism," *Expert Syst. Appl*., p. 125258, 2024, doi: 10.1016/j.eswa.2024.125258.

[12] Y. Mei, B. Sun, D. Li, H. Yu, H. Qin, H. Liu, N. Yan, and Y. Chen, "Recent advances of target tracking applications in aquaculture with emphasis on fish," *Comput. Electron. Agric*., vol. 201, p. 107335, Oct. 2022, doi: 10.1016/j.compag.2022.107335.

[13] A. Al Muksit, F. Hasan, M. F. H. B. Emon, M. R. Haque, A. R. Anwary, and S. Shatabda, "YOLO-Fish: A robust fish detection model to detect fish in realistic underwater environment," *Ecol. Inform*., vol. 72, p. 101847, 2022, doi: 10.1016/j.ecoinf.2022.101847.

[14] L. Ortenzi, J. Aguzzi, C. Costa, S. Marini, D. D'Agostino, L. Thomsen, et al., "Automated species classification and counting by deep-sea mobile crawler platforms using YOLO," *Ecol. Inform*., vol. 82, p. 102788, 2024, doi: 10.1016/j.ecoinf.2024.102788.

[15] X. Xu, W. Li, and Q. Duan, "Transfer learning and SE-ResNet152 networks-based for small-scale unbalanced fish species identification," *Comput. Electron. Agric*., vol. 180, p. 105878, 2021, doi: 10.1016/j.compag.2020.105878.

[16] S. K. Aruna, N. Deepa, and T. Devi, "Underwater fish identification in real-time using convolutional neural network," in *Proc. 7th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Madurai, India, 2023, pp. 586–591, doi: 10.1109/ICICCS56967.2023.10142531.

[17] A. Jalal, A. Salman, A. Mian, M. Shortis, and F. Shafait, "Fish detection and species classification in underwater environments using deep learning with temporal information," *Ecol. Inform*., vol. 57, p. 101088, 2020, doi: 10.1016/j.ecoinf.2020.101088.

[18] V. Kandimalla, M. Richard, F. Smith, J. Quirion, L. Torgo, and C. Whidden,

"Automated detection, classification and counting of fish in fish passages with deep learning," *Front. Mar. Sci*., vol. 8, p. 2049, 2022, doi: 10.3389/fmars.2021.778835.

[19] K. M. Knausgård, A. Wiklund, T. K. Sørdalen, K. T. Halvorsen, A. R. Kleiven, L. Jiao, and M. Goodwin, "Temperate fish detection and classification: a deep learning based approach," *Appl. Intell*., pp. 1–14, 2022, doi: 10.1007/s10489-022-03437-y.

[20] S. J. Rani, I. Ioannou, R. Swetha, R. D. Lakshmi, and V. Vassiliou, "A novel automated approach for fish biomass estimation in turbid environments through deep learning, object detection, and regression," *Ecol. Inform*., p. 102663, 2024, doi: 10.1016/j.ecoinf.2024.102663.

[21] M. Saqib, M. R. Khokher, X. Yuan, B. Yan, D. Bearham, C. Devine, C. Untiedt, T. Cannard, K. Maguire, G. N. Tuck, et al., "Fishing event detection and species classification using computer vision and artificial intelligence for electronic monitoring," *Fish. Res*., vol. 280, p. 107141, 2024, doi: 10.1016/j.fishres.2024.107141.

[22] D. Cao, C. Guo, M. Shi, Y. Liu, Y. Fang, H. Yang, Y. Cheng, W. Zhang, Y. Wang, Y. Li, et al., "A method for custom measurement of fish dimensions using the improved YOLOv5-keypoint framework with multi-attention mechanisms," *Water Biol. Secur*., vol. 3, no. 4, p. 100293, 2024, doi: 10.1016/j.watbs.2024.100293.

[23] R. Garcia, R. Prados, J. Quintana, A. Tempelaar, N. Gracias, S. Rosen, H. Vågstøl, and K. Løvall, "Automatic segmentation of fish using deep learning with application to fish size measurement," *ICES J. Mar. Sci*., vol. 77, no. 4, pp. 1354–1366, 2020, doi: 10.1093/icesjms/fsaa018.

[24] M. Jahanbakht, M. R. Azghadi, and N. J. Waltham, "Semi-supervised and weakly-supervised deep neural networks and dataset for fish detection in turbid underwater videos," *Ecol. Inform*., vol. 78, p. 102303, 2023, doi: 10.1016/j.ecoinf.2023.102303.

[25] H. Sun, J. Yue, and H. Li, "An image enhancement approach for coral reef fish detection in underwater videos," *Ecol. Inform*., vol. 72, p. 101862, 2022, doi: 10.1016/j.ecoinf.2022.101862.

[26] S. Abe, T. Takagi, S. Torisawa, K. Abe, H. Habe, N. Iguchi, K. Takehara, S. Masuma, H. Yagi, T. Yamaguchi, and S. Asaumi, "Development of fish spatio-temporal identifying technology using SegNet in aquaculture net cages," *Aquac. Eng*., vol. 93, p. 102146, 2021, doi: 10.1016/j.aquaeng.2021.102146.

[27] L. Yang, H. Yu, Y. Cheng, S. Mei, Y. Duan, D. Li, and Y. Chen, "A dual attention network based on EfficientNet-B2 for short-term fish school feeding behavior analysis in aquaculture," *Comput. Electron. Agric*., vol. 187, p. 106316, 2021, doi: 10.1016/j.compag.2021.106316.

[28] X. Yu, Y. Wang, D. An, and Y. Wei, "Counting method for cultured fishes based on multi-modules and attention mechanism," *Aquac. Eng*., vol. 96, p. 102215, 2022, doi: 10.1016/j.aquaeng.2021.102215.

[29] J. Li, C. Liu, L. Wang, Y. Liu, R. Li, X. Lu, et al., "Multi-species identification and number counting of fish passing through fishway at hydropower stations with LigTraNet," *Ecol. Inform*., vol. 82, p. 102704, 2024, doi: 10.1016/j.ecoinf.2024.102704.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," arXiv preprint arXiv:2005.12872, 2020, doi: 10.48550/arXiv.2005.12872.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020, doi: 10.48550/arXiv.2010.11929.

[32] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Adv. Neural Inf. Process. Syst*., vol. 34, pp. 9355–9366, 2021, doi: 10.48550/arXiv.2104.13840.

[33] G. Sun, Z. An, Y. Liu, C. Liu, C. Sakaridis, D.-P. Fan, and L. Van Gool, "Indiscernible object counting in underwater scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 13791–13801, doi: 10.1109/CVPR52729.2023.01325.

[34] D. Liang, X. Chen, W. Xu, Y. Zhou, and X. Bai, "TransCrowd: Weakly-supervised crowd counting with transformers," *Sci. China Inf. Sci.*, vol. 65, no. 6, p. 160104, 2022, doi: 10.1007/s11432-021-3445-y.

[35] H. Lin, Z. Ma, X. Hong, Q. Shangguan, and D. Meng, "Gramformer: Learning crowd counting via graph-modulated transformer," in *Proc. AAAI Conf. Artif. Intell*., 2024, pp. 3395–3403, doi: 10.1609/aaai.v38i4.28126.

[36] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "BlendMask: Top-down meets bottom-up for instance segmentation," in Proc. *IEEE/CVF Conf. Comput. Vis. Pattern*

Recognit. (CVPR), 2020, pp. 8573–8581, doi: 10.1109/CVPR42600.2020.00860.

[37] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, 2020, doi: 10.1109/TPAMI.2018.2844175.

[38] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, S. Kwong, et al., "WaterMask: Instance segmentation for underwater imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 1305–1315, doi: 10.1109/ICCV51070.2023.00126.

[39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, et al., "Segment Anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV),* 2023, pp. 3992–4003, doi: 10.1109/ICCV51070.2023.00371.

[40] S. Lian, Z. Zhang, H. Li, W. Li, L. T. Yang, S. Kwong, et al., "Diving into underwater: Segment Anything Model guided underwater salient instance segmentation and a large-scale dataset," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024, Paper 235, doi: 10.48550/arXiv.2406.06039.

[41] B. Clinton, A. Amperawan, and T. Dewi, "Object detection approach using YOLOv5 for plant species identification," *J. Elektronika dan Telekomunikasi (JET)*, vol. 24, no. 2, pp. 120–128, Dec. 2024, doi: 10.55981/jet.643.