Semi-Supervised Object Detection for Marine Monitoring using Temporal Information

Viljar Holm Elvevoll[1], Kim Tallaksen Halvorsen[2], and Ketil Malde[1,2]
[1] Department of Informatics, University of Bergen, Norway
[2] Institute of Marine Research, Bergen, Norway
e-mail: viljarhe00@hotmail.no, {kim.halvorsen | ketil.malde}@hi.no

***Abstract*—Accurate and sustainable monitoring of marine biodiversity is crucial for effective fisheries management and conservation. Traditional fish population assessments, relying on manual annotation and invasive techniques, are labor-intensive and potentially harmful to marine ecosystems. This work presents a Semi-Supervised Learning (SSL) approach that leverages extensive unlabeled underwater video data to significantly enhance object detection performance for fish species. By integrating the YOLOv8 object detector with Multi-Object Tracking (MOT) algorithms, specifically ByteTrack, a novel methodology is proposed to generate high-quality pseudolabels from temporal sequences. Iterative training incorporating these pseudolabels consistently improved model precision and recall, with the best-performing approach (ByteTrack with an extrapolated heuristic) demonstrating average precision of 90%, recall of 70%, mAP50 of 74%, and mAP50-95 of 59%. Notably, scores improved substantially over the baseline supervised model on all metrics. These results underscore the potential of temporally informed pseudolabeling in enhancing fish detection accuracy and robustness, reducing reliance on manual annotations and supporting sustainable marine monitoring practices.**

Keywords-Image classifcation; machine learning; species recognition; semi-supervised learning

I. Introduction

Traditionally, methods for monitoring marine ecosystems include trawling, netting, and manual visual surveys by divers, which are labor intensive, costly, and often disruptive to habitats or producing bycatch. Less invasive methods using underwater cameras, such as Baited Remote Underwater Video (BRUV) and Remote Underwater Video (RUV), are often an attractive alternative [1][2]. To process the large volumes of collected video data, it is necessary to use automated analysis tools, typically object detection models [3][4]. However, training such models requires large amounts of high-quality, labeled data, which is costly to produce.

To address this limitation, we here investigate semi-supervised learning [5], an automated method to iteratively generate training data sets using predictions from preliminary models (pseudolabels) that are considered sufficiently reliable. In contrast to earlier work, we selected the pseudolabeled data to use based on temporal information (*i.e.*, tracking) rather than more commonly used confidence scores. This is particularly advantageous in this setting, since an abundance of temporally contiguous video or image data can be produced, but expert annotation is time consuming and requires skilled curators.

The rest of the paper is structured as follows. In Section II, we describe the data set and the method for generating pseudolabels, as well as the training regime. In Section III, we present the results, and select the best performing method to investigate further. In Section IV, we discuss the results and their implications, and propose an explanation for the observations, before we conclude in Section V.

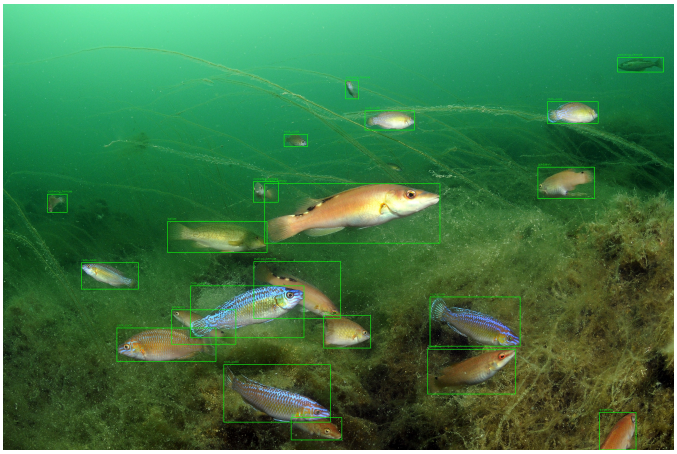

Figure 1. An example from the data set showing several annotated fish of various species (Photo by Erling Svendsen, used with permission).

II. Methods

For this study, we used a data set consisting of 1248 images from a combination of sources (RUVs, photos by divers) under different conditions and with variable resolutions (see Figure 1 for an example). The annotation by experts from the Institute of Marine Research include 10 categories (Figure 2): corkwing wrasse (*Symphodus melops*; male and female), two-spotted goby (*Pomatoschistus flavescens*), goldsinny (*Ctenolabrus rupestris*, rock cook (*Centrolabrus exoletus*, cuckoo wrasse (*Labrus mixtus* male and female), pollack (*Pollachius pollachius*), ballan wrasse (*Labrus bergylta*), and *unknown* fish that could not be labeled to the species level due to low visibility or by being too distant from the camera. In addition, six unannotated videos from similar habitats were used as sources of pseudolabeled frames for the semi-supervised training. As some species were not present or very scarce in the unlabeled data, the semi-supervised method is only trained on five of the classes: male and female corkwing, two-spotted goby, goldsinny, and ballan.

The object detection model used was YOLOv8 (Ultralytics, 2023), a state-of-the-art single pass object detector [6], combined with two advanced tracking algorithms ByteTrack [7] and DeepSort [8] to construct an iterative pipeline:

1. Base Model Training: An initial YOLOv8 model (YOLOv8x, the largest variant) was trained using a smaller,

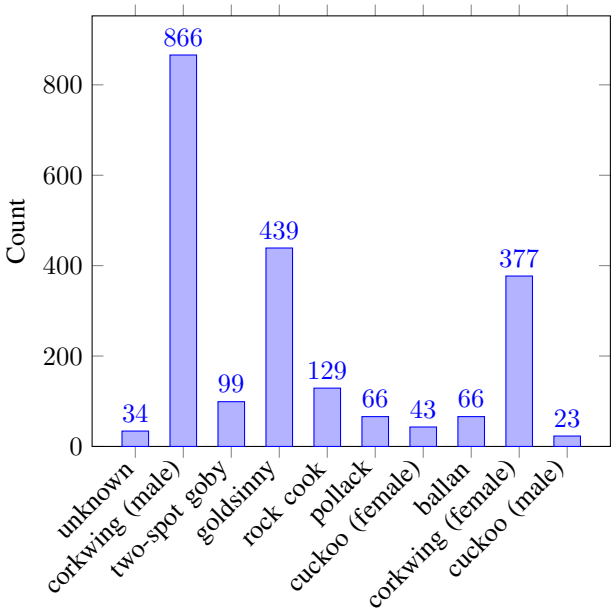

Figure 2. The distribution of the classes of annotated objects in the data set.

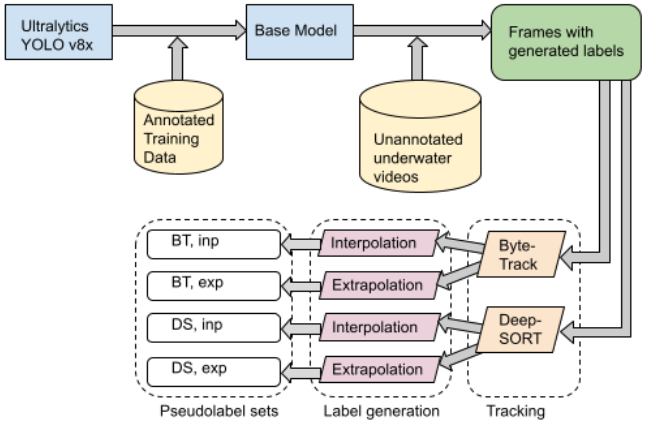

Figure 3. The process for generating the base model and the four pseudolabeled data sets (step 1 and step 2), using automated labeling of video data, tracking, and label interpolation and extrapolation.

TABLE I. PERFORMANCE METRICS AFTER 100 EPOCHS OF TRAINING.

Model	Prec	Recall	mAP50	mAP50-95
DS, inp	0.74988	0.56890	0.61023	0.47017
DS, exp	0.66251	0.56209	0.57969	0.45028
BT, inp	0.70776	0.53875	0.58142	0.44878
BT, exp	0.88792	0.62665	0.69120	0.52972

manually annotated dataset. This provides the baseline for performance evaluation.

2. Pseudolabel Generation using Temporal Information: The trained model is then used to process unlabeled underwater video recordings and extract pseudolabeled data as illustrated in Figure 3. By integrating Multi-Object Tracking (MOT) algorithms (ByteTrack and DeepSORT), frames with objects that are missed or given low score by the detector can still be identified with high confidence. Two heuristics were investigated for selecting frames to generate pseudolabels:

A. Interpolated Intermediate Labels: This heuristic infers an object's presence in intermediate frames if it is detected by the model in preceding and subsequent frames, and the MOT algorithm assigns the same track ID. This method yields fewer but potentially highly accurate pseudolabels.

B. Extrapolated Labels: This more inclusive approach requires an object to be natively detected at least three times consecutively. All subsequent detections of that object by the MOT algorithm are included, forming an unbroken chain, even if the native model fails to detect it in every frame. This significantly increases the volume of pseudolabeled data.

3. Iterative Retraining: The pseudolabels generated are combined with the original labeled dataset, and the model is retrained. It is crucial to retain the original data to prevent catastrophic forgetting of classes not present in the video recordings. The learning rate of the AdamW optimizer is reset at the start of each new training phase to facilitate rapid adjustment and prevent trapping in suboptimal local minima. This iterative process (pseudolabel generation followed by retraining) is repeated for multiple cycles to progressively improve the model.

III. RESULTS

The study rigorously evaluated four configurations: ByteTrack with interpolated intermediate labels, DeepSORT with interpolated intermediate labels, ByteTrack with extrapolated labels, and DeepSORT with extrapolated labels. Each configuration was run for 100 epochs (50 supervised, followed by 50 semi-supervised with pseudolabels) which yielded the results in Table I. We see that all models perform adequately, but ByteTrack with extrapolated labels consistently outperformed the other models.

In order to explore the limits of semi-supervised training, the baseline and ByteTrack with extrapolation models were trained for 250 epochs. In Figure 4, we can see how the different components of the loss rapidly decrease both for training (top row) and validation (bottom row) data, while the four different performance measures increase correspondingly. We also observe five distinct jumps in the graphs, these are caused by introduction of new data and resetting of the learning rate for each iteration, which cause an initial worsening of scores before the model gradually converges again.

The performance statistics on the test set after 250 epochs is shown in Table II. We see that using semi-supervised training with ByteTrack and the extrapolated pseudolabeling scheme results in substantial improvements for all metrics.

Per class improvements are shown in Figure 5. As expected, classes present in the semi-supervised training data (shown in solid colors) see substantial improvements on all metrics. Classes not present (shown with faded colors) see slight degradation in precision, and mAP, but surprisingly recall improves also for these classes.

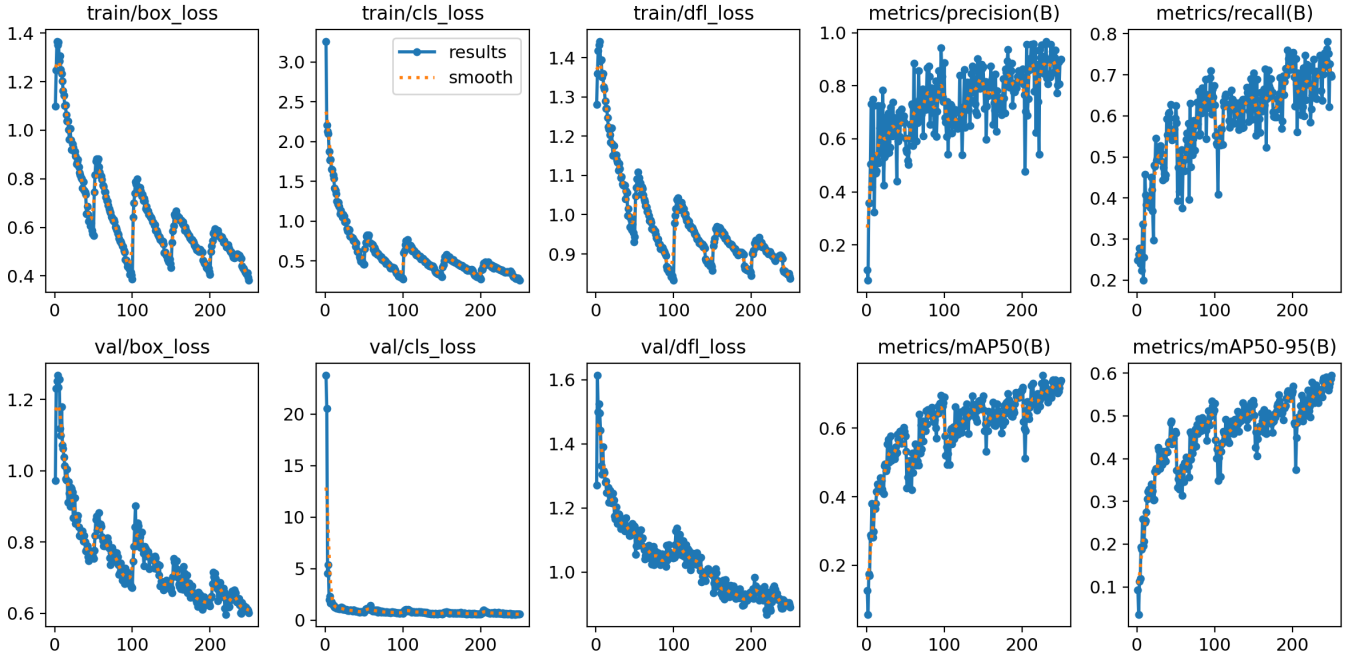

Figure 4. Training from ByteTrack tracks with extrapolated labels for 250 epochs (one iterations of supervised followed by four iterations of semi-supervised training).

TABLE II. PERFORMANCE OF BASELINE AND BYTETRACK WITH EXTRAPOLATION MODELS AFTER EXTENSIVE (250 EPOCHS) TRAINING.

Model	Prec	Recall	mAP50	mAP50-95
Base	0.75652	0.52468	0.57954	0.48811
BT, exp	0.90011	0.69644	0.73863	0.59468

IV. DISCUSSION

The most effective approach was ByteTrack combined with the extrapolated heuristic. This configuration consistently outperformed all other tested methods, as well as the baseline supervised model. The results demonstrate that leveraging temporal information through pseudolabeling significantly enhances fish detection accuracy and consistency. The substantial improvements in precision, recall, and mAP for the pseudolabeled classes, coupled with minimal negative impact on other classes, validate the effectiveness of this approach in mitigating annotation scarcity.

A crucial insight from this study is the progressive mitigation of initial model biases through iterative pseudolabeling. For instance, a systematic error where parts of the monitoring equipment were misclassified as "corkwing male" in early iterations (Figure 6) was effectively corrected and eliminated in later iterations using ByteTrack with extrapolated labels. This highlights the self-correcting nature of the temporal semi-supervised framework, guiding the model towards more accurate predictions over time.

The choice of MOT algorithm also proved critical. ByteTrack consistently outperformed DeepSORT in this semi-supervised setup. We suspect the discrepancy is caused by the use of Kalman filters in DeepSORT, which can interpolate predictions even when the object is lost by the detection model. While beneficial in predictable scenarios, this can lead to inaccurate pseudolabels for fish due to their often erratic movements, possibly creating an "off-policy" learning situation akin to the "Deadly Triad" in Reinforcement Learning, which can impede stable convergence. ByteTrack, by contrast, relies solely on the detector's predictions, ensuring a stronger alignment between pseudolabels and the model's current capabilities, thus avoiding such instability.

Semi-supervised learning has been used effectively in many different settings, but selecting pseudolabeled data to train on can be difficult. Using classifier confidence is an option [9], but tends in our experience to improve the classifier where it is already strong. Using augmentation [10] or taking class balance into account [11] may help to mitigate this, but by extracting presudolabels from temporal information removes (or at least reduces) the dependence on the classifier itself from the selection process. Although temporal pseudolabeling methods have been attempted before (*e.g.*, [12]), our approach distinguishes itself by targeting the model's weaknesses rather than reinforcing its strengths. By relying on MOT algorithms to generate labels specifically where the base model fails to detect objects, it directly addresses gaps in detection capability.

V. CONCLUSION AND FUTURE WORK

We have successfully demonstrated the significant potential of semi-supervised learning leveraging temporal information for enhancing object detection in marine life monitoring. By integrating YOLOv8 with ByteTrack, a robust methodology was developed to generate high-quality pseudolabels from unlabeled video data, substantially improving model performance

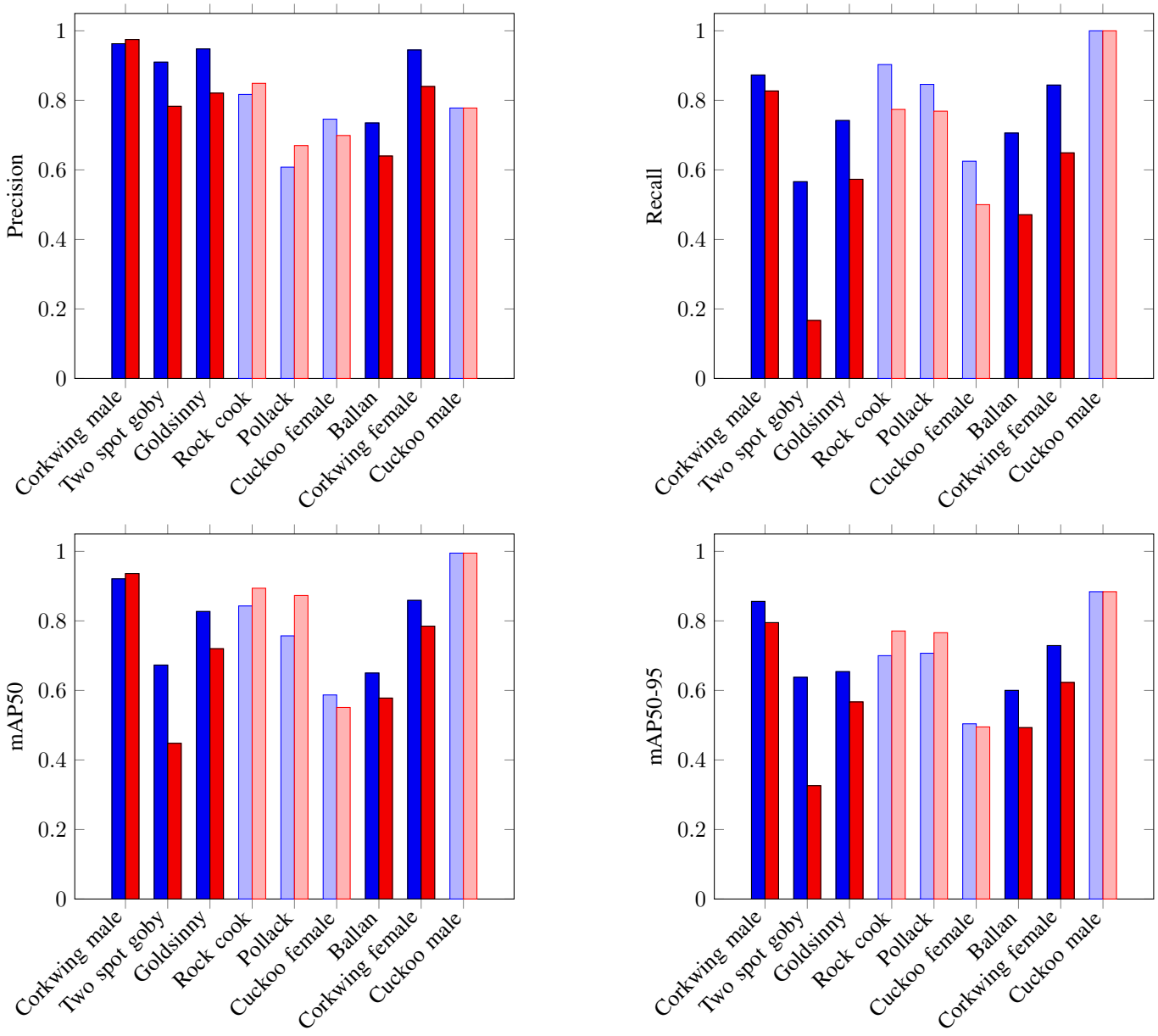

Figure 5. Precision (top left), recall (top right), mAP50 (bottom left) and mAP50-95 (bottom right) for baseline (red) and semi-supervised (blue) models. Classes not present in the pseudolabeled data shown with faded color.

for fish species. This approach reduces the reliance on costly and labor-intensive manual annotations, paving the way for more sustainable and scalable marine life assessment practices. The insights gained regarding iterative bias mitigation and the critical role of MOT algorithm selection provide valuable directions for future research and practical deployment in real-world marine conservation efforts.

Acknowledgement

This work is based on results from the Master's degree thesis of VHE [13], which includes a more detailed exploration of the methods and data. The image data was collected and annotated as part of the CoastVision project, RCN grant number 325862.

References

[1] A. W. Bicknell, B. J. Godley, E. V. Sheehan, S. C. Votier, and M. J. Witt, "Camera technology for monitoring marine biodiversity and human impact," *Frontiers in Ecology and the Environment*, vol. 14, no. 8, pp. 424–432, 2016. DOI: https://doi.org/10.1002/fee.1322. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/fee.1322. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/fee.1322.

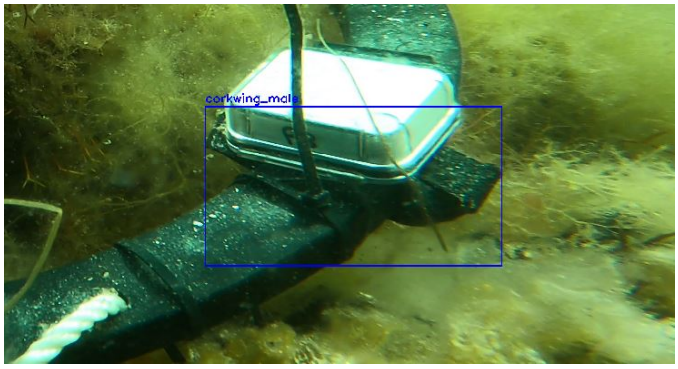

Figure 6. Equipment incorrectly predicted as “corkwing male” by the initial classifier.

[2] S. K. Whitmarsh, P. G. Fairweather, and C. Huveneers, “What is Big BRUVver up to? methods and uses of baited underwater video,” *Reviews in Fish Biology and Fisheries*, vol. 27, no. 1, pp. 53–73, 2017, ISSN: 1573-5184. DOI: 10.1007/s11160-016-9450-1. [Online]. Available: https://doi.org/10.1007/s11160-016-9450-1.

[3] H. Liu, X. Ma, Y. Yu, L. Wang, and L. Hao, “Application of deep learning-based object detection techniques in fish aquaculture: A review,” *Journal of Marine Science and Engineering*, vol. 11, no. 4, p. 867, 2023.

[4] P. Rubbens et al., “Machine learning in marine ecology: An overview of techniques and applications,” *ICES Journal of Marine Science*, vol. 80, no. 7, pp. 1829–1853, 2023.

[5] J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Machine learning*, vol. 109, no. 2, pp. 373–440, 2020.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection*, 2016. arXiv: 1506.02640 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1506.02640.

[7] Y. Zhang et al., *Bytetrack: Multi-object tracking by associating every detection box*, 2022. arXiv: 2110.06864 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2110.06864.

[8] N. Wojke, A. Bewley, and D. Paulus, *Simple online and realtime tracking with a deep association metric*, 2017. arXiv: 1703.07402 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1703.07402.

[9] K. Sohn et al., “A simple semi-supervised learning framework for object detection,” *arXiv preprint arXiv:2005.04757*, 2020.

[10] J. Jeong, S. Lee, J. Kim, and N. Kwak, “Consistency-based semi-supervised learning for object detection,” *Advances in neural information processing systems*, vol. 32, 2019.

[11] M. Xu et al., “End-to-end semi-supervised object detection with soft teacher,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3060–3069.

[12] R. J. Veiga et al., “Autonomous temporal pseudo-labeling for fish detection,” *Applied Sciences*, vol. 12, no. 12, p. 5910, 2022.

[13] V. H. Elvevoll, “Semi-supervised object detection using temporal information,” M.S. thesis, Department of Informatics, The University of Bergen, 2025.