Detecting salmon lice in seawater using synthetic datasets

Proceedings 2024 International Conference on Machine Learning and Applications ICMLA 2024

Zheng, Lei; Zhang, C.; Bracke, M.B.M.; Gansel, Lars Christian; da Silva Torres, R.

https://doi.org/10.1109/ICMLA61862.2024.00130

Detecting Salmon Lice in Seawater Using Synthetic Datasets

Lei Zheng[1], Chao Zhang[1], Marc Bracke[1], Lars Gansel[2], and Ricardo da Silva Torres[1]
[1]*Wageningen University and Research*, Wageningen, The Netherlands
[2]NTNU – Norwegian University of Science and Technology, Ålesund, Norway
{lei.zheng, chao.zhang, marc.bracke, ricardo.dasilvatorres}@wur.nl, lars.gansel@ntnu.no

2024 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-7488-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLA61862.2024.00130

***Abstract*—Salmon lice are one of the most serious threats to the salmon aquaculture industry. Automatic methods for monitoring the density of salmon lice larvae in seawater are paramount for taking proper response measures. Computer vision technologies are promising but require large annotated datasets to create effective detection models. This paper investigates the potential of using methods to create datasets containing synthetic images based on the combination of salmon lice masks (collected from real images) with different environment-related backgrounds. The diversity of the synthetic dataset is ensured through a spectrum of factors, including the number of synthetic images, poses, variations in brightness, contrast, and saturation of objects and background images, 'noise' injection, and scale. Orthogonal experiments are conducted to assess the impact of these factors. We also evaluate the effectiveness of a state-of-the-art object detector, YOLOv8, trained on created synthetic datasets. The results show that the use of synthetic datasets leads to significant improvements, up to 75 percent, in the recall of the salmon lice detection. Rotation is the transformation that had the most significant impact on the synthetic dataset. Moreover, this approach is applicable to datasets of different sizes and other detectors. The use of synthetic datasets seems a valuable tool for the detection of 'needles in a haystack' such as incidental salmon louse larvae free swimming in large volumes of water with variable and complex backgrounds.**

***Index Terms*—Salmon lice detection, computer vision, synthetic dataset, orthogonal experiments**

I. Introduction

The salmon farming industry is one of the most important export industries in Norway, the biggest producer of salmon today. Norwegian salmon production has increased quickly since the 1990s, but today biological challenges are slowing the further growth of this industry. In 2023, 16.7 % of all farmed salmon in Norway died during production, and especially salmon lice caused massive economic losses and increased mortality [1]. Salmon louse is an ectoparasite of salmonid fishes that feed on salmon blood, skin, and mucus, causing delayed fish growth, increased stress, and higher susceptibility to diseases [2]. Salmon lice have a relatively simple life cycle with eight stages that include nauplius (two stages), copepodite, chalimus (two stages), pre-adult (two stages), and adult. After hatching from the egg string, salmon lice larvae swim freely in the water and develop through nauplius and copepodite stages. During the infectious copepodite stage, the salmon lice larvae will seek out a host fish and live there for the rest of their lives. Obtaining data on salmon lice in the (pre-)infected stage in seawater is essential to model the spread of salmon lice, which plays a crucial role in the management of farmed salmon and the conservation of wild salmon. In addition, the detection of planktonic salmon lice larvae is crucial for early warning, and to test the efficacy of diverse and especially new methods for the prevention and treatment of salmon lice infestations.

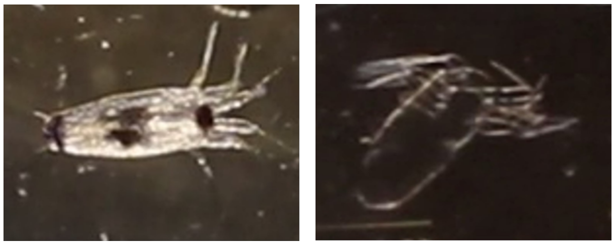
Traditional salmon lice detection in seawater relies on molecular methods or human vision methods. However, molecular methods cannot distinguish the growth stages of salmon lice and require complex pre-processing in the laboratory. Human vision methods mainly rely on manual observation microscopy to distinguish the salmon lice, which is very time-consuming and labor-intensive. It is practically impossible to monitor the abundance of different stages of salmon lice in large seawater samples (tens to hundreds of cubic meters) in different horizontal and vertical locations along the coast when employing microscopic methods.

Machine learning on images/video offers the possibility of rapid detection of salmon lice at different life stages, as it can make predictions about new samples by learning relevant visual patterns from training data. Applying machine learning often requires a lot of labeled data. However, the concentration of salmon lice is low in general and other similar-sized plankton are abundant and variable, making it very time-consuming and labor-intensive to obtain sufficient labeled data required for training and updating the model. While samples can be spiked with laboratory-hatched salmon lice to obtain large data sets of annotated salmon lice, the background cannot easily be replaced by similar methods to create physical samples with a variety of realistic plankton communities and densities. In addition, changes, both physical and in the settings of the imaging system delivering the data to be analysed, may impact the performance of models trained on data from an imaging system with different properties.

To cope with the lack of sufficient data, many methods have been explored. A promising direction relies on the use of synthetic datasets [3]. The goal is to enhance the model's adaptability to diverse scenarios and backgrounds encountered in real-world applications. Compared to on-site data collection and labeling, generating synthetic data is more cost-effective and efficient, with the added flexibility of creating data tailored for specific tasks or scenarios. However, the quality of the syn-

This work was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek [grant number Nwa.1332.20.002].

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00130



(a) (b)

Fig. 1: Examples of “salmon louse nauplius” (a) and “salmon louse nauplius shell” (b).

thesized data may directly impact model performance. Also, synthesized images may not fully encapsulate the complexity and diversity of the real world. Insufficient diversity in the synthesis strategy or the presence of fixed patterns can predispose models to overfitting. Likewise, if the data generation method is biased, models might inadvertently incorporate and perpetuate these biases.

This paper investigates the use of synthetic datasets targeting the detection of free-swimming salmon lice. Our methodology is based on the combination of salmon lice masks (collected from real images) with different environment-related backgrounds. The diversity of the synthetic dataset is ensured through the use of different factors, such as the number of synthetic images, poses, variations in brightness, contrast, and saturation of objects and background images, ‘noise’ injection, and scale. We perform an orthogonal experiment to evaluate the importance of each factor, as well as assess the impact of using synthetic datasets for the training of a state-of-the-art detector, YOLOv8. Experimental results show that employing synthetic datasets leads to gains of up to 75.37 % in terms of recall. Rotation is identified as the most relevant factor.

II. Material & Methods

A. Dataset description

In total, 2100 images with salmon lice or salmon lice shells were obtained with a resolution of 1920 × 1080 pixels. Two classes, “salmon louse nauplius” and “salmon louse nauplius shell” (Fig. 1), were annotated with bounding boxes in the images. As this study focuses on the case of lacking sufficient data for training, we used a very small set with only 100 annotated images (referred to as raw data – RD) as input for generating the synthetic training set. We used 1000 images (with 799 nauplii and 543 shells) as validation set (VD) to select the model during the training progress, and 1000 images (with 514 nauplii and 840 shells) as test set (TD) to evaluate the final performance of the models. In addition, we also had 100 background images with plankton and without salmon lice, which were used to generate synthetic images.

B. Data synthesis method

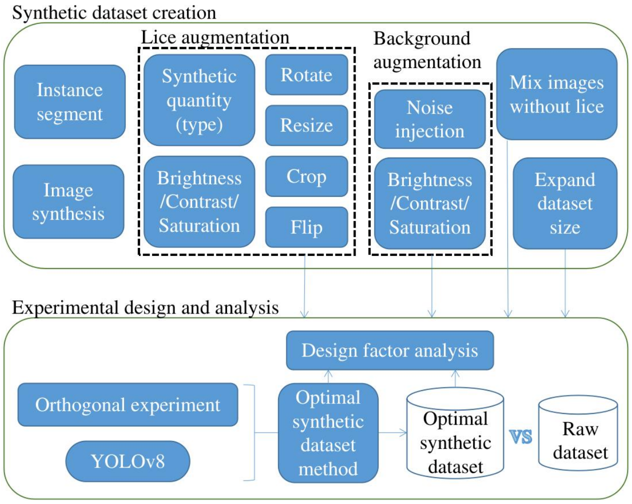


Fig. 2: Figure of data synthesis method. The creation of the synthetic dataset comprised two basic synthetic methods, six object-related (lice) data augmentation methods, two background-related enhancement methods, and two methods to increase the diversity of the dataset. The orthogonal experimental design and analysis focused on finding the optimal synthetic dataset method and quantifying the impact of the factors on improved model performance by comparing the performance metrics of the YOLOv8 model trained on the synthetic datasets versus RD.

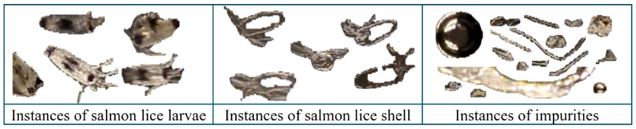

Instances of salmon lice larvae
Instances of salmon lice shell
Instances of impurities

Fig. 3: Examples of the instances extracted from images.

The data synthesis methodology proposed in this study is shown in Fig. 2. The synthetic images were mainly created by copying and pasting object instances with transformation. Firstly, the targets of interest (Fig. 3) in the original images were cropped out, and then pasted into the new image. Some transformations, such as cropping and rotation, were also applied to the target instances or the whole image to increase the diversity of newly created images. In this study, the cropped instances were salmon louse nauplii and salmon louse nauplius shells in addition to other impurities in the image.

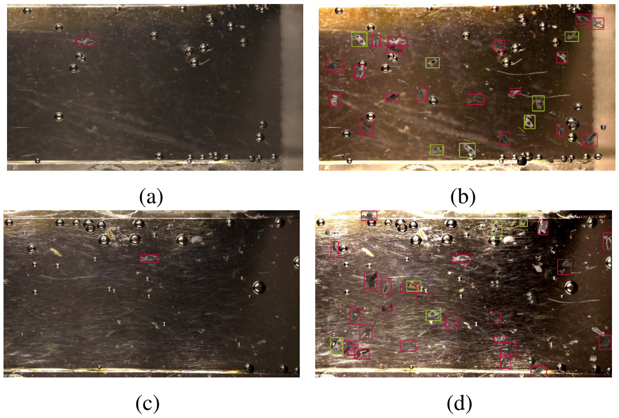
Segment Anything [4] was utilized for image segmentation. As this project focuses on the processing and analysis of 2D images, the varied poses of salmon lice in different scenes play a crucial role in model learning. Hence, all salmon lice nauplii (47 instances), shells (87 instances), and some impurities (140 instances) in the RD were extracted. When creating a synthetic image, instances were randomly pasted over an empty background to form a new image, or pasted on an RD image. Also, different instances were pasted over the same background image multiple times, leading to a diverse set of new images. When pasting an instance, a random coordinate was selected within the pixel dimensions of the background image, and the instance was allocated to this coordinate. Thus, with the dimensions of the instance image and background image, the annotation information for the synthetic instance was obtained. To increase the diversity of the synthetic images, we simulated various environmental conditions, including various transformations or factors:

1) **Number of synthetic objects** indicates the number of target objects (salmon lice nauplius and their shell) pasted onto the background. Pasting different quantities of objects onto background images can enhance the diversity of the dataset. However, an excessive number of objects may lead to significant overlap, which could degrade the quality of the dataset. The number of objects synthesized was designed to range from 2 to 30.
2) **Rotation** indicates the objects are rotated randomly when pasting to the new images.
3) **Resizing** indicates the objects are resized randomly within a range when pasting to the new images. The scale of resizing was set between 0-40 %.
4) **Cropping** indicates the objects are cropped at random scales when pasting to the new images. The scale of cropping was set between 0-40 %.
5) **Flipping** indicates the objects are flipped randomly when pasting to the new images.
6) **B/C/S to instances** indicates the objects undergo random variations in brightness, contrast, and saturation when pasting to the new images. The range of variation factors for adjusting brightness, contrast, and saturation was set between 0 and 2.
7) **Noise** indicates a certain number of impurities (e.g., plankton objects) were introduced into the new images to create "noise" in the data, further enhancing the diversity of the dataset. The number of impurities introduced ranged from 0 to 30.
8) **B/C/S to background** indicates the background images were subjected to random adjustments in brightness, contrast, and saturation. The range of variation factors for adjusting brightness, contrast, and saturation was set between 0 and 2.
9) **Background without objects** indicates the number of background images (with plankton but without salmon lice and shells) from outside of RD. The number of background images introduced in this study was set between 0 and 100.
10) **Expand** indicates the ratio of the synthetic dataset to the original dataset to represent the magnification factor. The magnification factor was set to a range of 1-50.

The final synthetic dataset was a result of the combined effects of these factors, as illustrated in the examples shown in Fig. 4.

C. Experimental protocol

YOLOv8 [5] is a machine-learning architecture released by Ultralytics in 2023. It enables tasks such as object detection and object segmentation. Compared to other mainstream models, such as Faster R-CNN [6], YOLOv5 [7], etc., it has a smaller model size, faster inference speed, and better performance. Since the detection of salmon lice in seawater requires the analysis of a large volume of seawater, the inference speed and performance are prioritised. Therefore, the YOLOv8 model was chosen as the detector for the experiments in this paper.



(a) (b) (c) (d)

Fig. 4: Examples of synthetic images. (a) and (c) were original images and (b) and (d) were synthetic images. Red boxes refer to salmon louse nauplius shells while green boxes refer to live salmon louse nauplii. The images to the left are the samples from the raw dataset (RD), while the images to the right depict corresponding samples from the synthetic dataset composed under the influence of different factors.

TABLE I: Levels considered for the different factors in the orthogonal experiment protocol.

Scope	Factor	Level 1	Level 2	Level 3	Level 4
Instance	Number of synthetic objects	2	2-10	2-20	2-30
	Rotation	False	True	-	-
	Resizing	0	0-20	0-30	0-40
	Cropping	0	0-20	0-30	0-40
	Flipping	False	True	-	-
	B/C/S to instances	1	0.7-1.3	0.3-1.7	0-2
Background	Noise	0	0-10	0-20	0-30
	B/C/S to background	1	0.7-1.3	0.3-1.7	0-2
	Background without objects	0	30	60	100
Both	Expand	1	10	25	50

To efficiently evaluate the impact of each factor on the results in a multi-factorial experiment while maximally reducing the number of experiments, an orthogonal experiment approach was employed [8]. The levels of each factor are shown in Table I. A $L32(4^8 2^2)$ orthogonal array was chosen corresponding to 32 factor level combinations. This resulted in 32 training datasets with sizes ranging from 100 to 10,000.

In this study, precision, recall, F1-score, mAP50, and mAP95 were chosen as the evaluation metrics. For a definition of those metrics, the reader is referred to [9].

Range analysis [8] was used to determine the importance of factors. The maximum difference in the effects of a factor at different levels reflects the extent to which variations in the factor lead to changes in the outcomes. Therefore, the results of the 32 experiments can be used to determine the optimal combination of levels (the combination of different levels that maximizes the effect difference). Additionally, The impact of the factors may vary depending on the evaluation metrics. For example, a change in the same factor may have a large effect on the model's precision and a small effect on recall. Therefore, there may be different optimal combinations of factors for different evaluation metrics. Taking recall as an example, range analysis can be used to determine the level of every factor that most improves the trained model. Subsequently, based on this combination of factors, a new synthetic dataset can be created and a new model (Model_Recall) can be trained. Comparing the performance of this refined model with the model trained on the original dataset allows for an assessment of the effectiveness of the synthetic dataset. The gain in model performance regarding salmon lice detection is a measure of the value of the use of the synthetic dataset. Similarly, models based on other evaluation metrics were also trained (e.g., Model_Precision, Model_F1, Model_mAP50, and Model_mAP95) and tested.

To more comprehensively assess whether the synthetic method is applicable under broader conditions, this study also used a different detector, RT-DETR [10], to conduct experiments following the methodology above. The test results of the five trained models (referred to as Model_precision_detr, Model_recall_detr, Model_F1-score_detr, Model_mAP50_detr, Model_mAP95_detr) on TD were compared with the results of Model_RD_detr, which was trained on RD using RT-DETR and tested on TD.

III. Results & Discussion

The YOLOv8 model was trained on each of the 32 synthetic datasets and tested on TD. The results (Fig. 5) revealed that models trained on the synthetic datasets showed significant improvements over the model trained on RD in terms of precision, recall, F1-score, mAP50, and mAP95. For instance, the highest recall achieved by the models trained on the synthetic dataset was 0.725, compared to 0.467 for the model trained on RD (second bar).

Range analysis was conducted on models trained on 32 datasets using precision, recall, F1-score, mAP50, and mAP95. Five optimal levels of factor combinations were calculated and five models were trained on the synthetic datasets using optimal combinations. These models were then tested using TD to compare their performance with the model trained on the RD. Models trained on the five synthetic datasets performed significantly better than those of Model_RD. Moreover, the test results of Model_mAP50 were better in all other metrics compared to the models trained on the other four synthetic datasets. At this point, a preliminary calculation can be made on the improvement levels of the salmon lice detection model using the optimal synthetic dataset generated based on the currently designed factors on RD. The gains of Model_mAP50 on TD compared to Model_RD on TD were 65.28% in the precision performance metric, 75.37% for the recall metric, 70.6% for mAP50, and 84.25% for mAP95, as shown in Fig. 6. The visualisation of the model predictions is shown in Fig. 7.

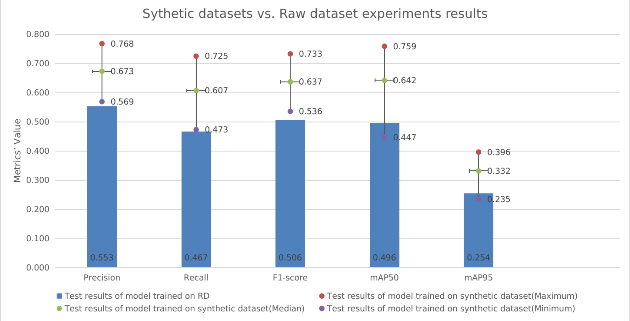


Fig. 5: Based on the results of 32 experiments, the maximum, median, and minimum values of Precision, Recall, F1-score, mAP50, and mAP95 were statistically determined. The blue bars represent the results of experiments conducted using the raw dataset, while the red, green, and purple dots respectively indicate the maximum, median, and minimum values obtained across five metrics using the synthetic dataset.

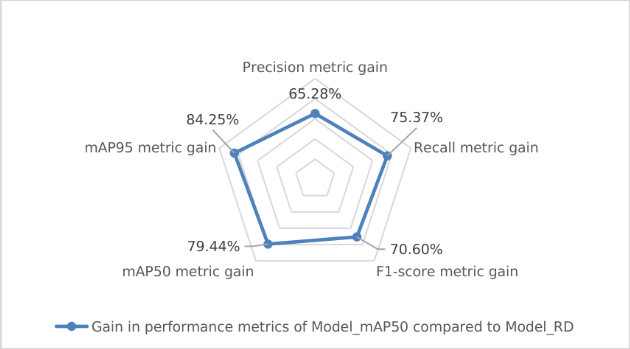


Fig. 6: Gains of Model_mAP50 on TD compared to Model_RD on TD, shown across the 5 performance metrics.

The impact of 10 factors was also determined through range analysis. The results of Model_mAP50, which had the best overall performance, are arranged in descending order of range, as shown in Table II. It was found that the 'Rotate' factor had the greatest impact on the quality of the synthetic dataset, with an impact measured by a range of 0.107, accounting for 21.73 % of the influence relative to other factors. 'Crop' and 'Number' follow, with their impact percentages at 15.17 % and 13.17 %, respectively. Lice's and background's 'B/C/S', as well as 'Expand' have similar impacts with percentages around 10 %. 'Resize' and 'Background without objects' are also close, with impact percentages around 7%. 'Flipping' has the smallest percentage at 0.38 %, significantly lower than the ninth-ranked 'Noise' (5.87 %). This implies that 'Flip' does not seem to play a substantial role in affecting the quality of the model trained on the synthetic dataset. It is worth mentioning that 'Rotate', identified in Table II as having the highest impact on synthetic dataset quality, was not identified as relevant in other studies [11], where the method of randomly pasting objects was considered sufficient. This may relate to

TABLE II: Detailed range analysis results for the optimal level combination selected based on the mAP_50 criterion. The first row shows the absolute impact of each factor on the quality of the synthetic dataset, measured by the value of the range. The second row displays the relative impact of each factor in relation to all the designed factors. B/C/S stands for brightness, contrast, and saturation.

Item	Rotate	Crop	Number	B/C/S(Background)	B/C/S(Instance)	Expand	Resize	Background w/ objects	Noise	Flip
Impact magnitude	0.107	0.075	0.065	0.052	0.046	0.046	0.039	0.032	0.029	0.002
Impact proportion	21.73%	15.17%	13.17%	10.48%	9.39%	9.37%	7.93%	6.51%	5.87%	0.38%

the task category targeted in the research. Our study focused on object detection tasks, whereas the research by Ghiasi et al. [11] mainly concentrates on instance segmentation tasks.

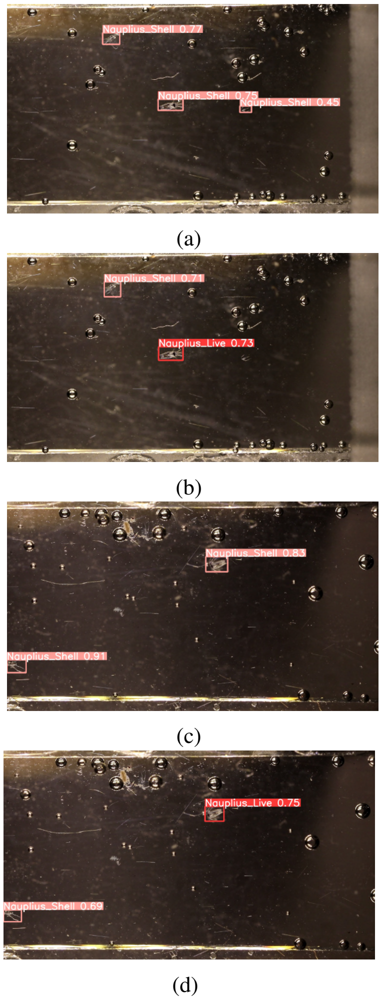
Similarly, RT-DETR trained on synthetic datasets created using the best combination of levels determined from orthogonal experiments with YOLOv8 also showed significant improvements over RD. Model_mAP95_detr outperformed the models trained on the other four synthetic datasets across all evaluation metrics. When tested on the same TD, Model_mAP95_detr exhibited improvements in precision by 11.44%, recall by 29.75%, mAP50 by 20.27%, and mAP95 by 26.43% compared to Model_RD_detr, as shown in Fig. 8. The results showed that the proposed method can be integrated into the training process of other detectors as well.

A. *Conclusions*

The synthetic dataset approach developed in this study provides a promising solution for the detection of salmon lice using computer vision. Especially in the case of limited training data. This method is not limited to applications for detecting salmon lice in seawater, but can also be extended to other environments and objects of interest that face similar data scarcity challenges. In the marine environment, obvious similar challenges exist for other parasites and toxic algae, amongst others. Of the 10 synthetic factors, six were found to significantly improve the quantity and diversity of data. Range analyses helped to identify the best combination of these factors to improve the quality of the synthetic dataset. Our study shows that salmon lice detection models can be improved using synthetic datasets. Moreover, the synthetic dataset approach can be applied to raw datasets of varying sizes and different detectors. By using synthetic datasets, the model can be adapted to the nuances of image quality in the output of different video systems and is more applicable to samples from different geographical locations and planktonic communities where the target organisms are to be found in relatively large volumes of water.

Acknowledgment

This work was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek [grant number Nwa.1332.20.002]. C. Zhang gratefully acknowledges financial support from the China Scholarship Council. The collection of videos and images of salmon lice and plankton was supported by the Norwegian University of Science and Technology via research license M-VS-00016.



(a)

(b)

(c)

(d)

Fig. 7: Visualization of model prediction results. Pink boundaries indicate predictions for salmon louse nauplius shells, while red bounding boxes indicate predictions for salmon louse nauplii, with numbers representing confidence levels. The images on the left (a and c) show the prediction results of the model trained using RD, whereas the images on the right (b and d) represent the prediction results of Model_mAP50. In (a) an impurity and a "Nauplius Live" were mistakenly detected as "Nauplius Shell". In (c) a "Nauplius Live" was mistakenly detected as "Nauplius Shell". In (b) and (d) all classes were detected correctly.

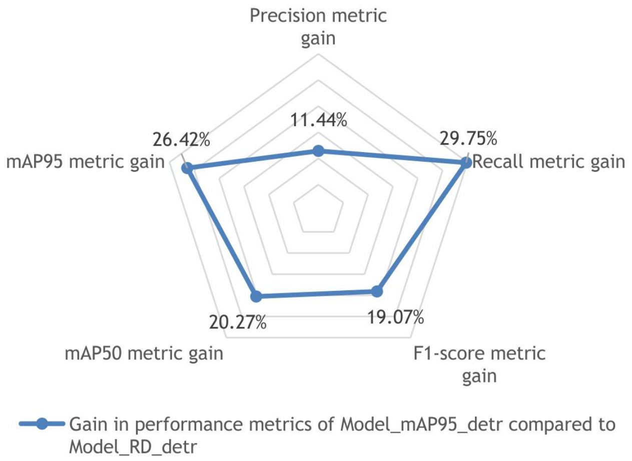


Fig. 8: Gains of Model_mAP95_detr on TD compared to Model_RD_detr on TD, shown across the 5 performance metrics.

REFERENCES

[1] Av Ingunn Sommerset, Av Victor H S Oliveira, Torfinn Moldal, Eve Marie Louise Zeyl Fiskebeck, and Hege Løkslett et al. Fiskehelserapporten 2023, 2024.

[2] Per Gunnar Fjelldal, Thomas WK Fraser, Tom J Hansen, Ørjan Karlsen, and Samantha Bui. Effects of laboratory salmon louse infection on mortality, growth, and sexual maturation in atlantic salmon. *ICES Journal of Marine Science*, 79(5):1530–1538, 2022.

[3] Stefan Hinterstoisser, Olivier Pauly, Hauke Heibel, Marek Martina, and Martin Bokeloh. An annotation saved is an annotation earned: Using fully synthetic training for object detection. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2787–2796. IEEE Computer Society, 2019.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[5] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[7] Glenn Jocher Ultralytics. Yolov5 by ultralytics, Oct 2020.

[8] Yueying Zhu, Weiyan Wei, Chuantian Yang, and Yan Zhang. Multi-objective optimisation design of two-phase excitation switched reluctance motor for electric vehicles. *IET Electric Power Applications*, 12(7):929–937, 2018.

[9] Irem Catal and Yavuz Selim Taşpinar. Automatic classification and detection of faulty packaging using deep learning algorithms: A study for industrial applications. *Intelligent Methods In Engineering Sciences*, 3(1):13–21, 2024.

[10] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[11] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021.