Expert Systems With Applications 237 (2024) 121197

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Aquaculture defects recognition via multi-scale semantic segmentation

Waseem Akram [a], Taimur Hassan [b], Hamed Toubar [a], Muhayyuddin Ahmed [a], Nikola Miškovic [c], Lakmal Seneviratne [a], Irfan Hussain [a,*]

[a] Khalifa University Center for Autonomous Robotic Systems (KUCARS), Khalifa University, United Arab Emirates
[b] Department of Electrical, Computer and Biomedical Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates
[c] University of Zagreb, Faculty of Electrical Engineering and Computing, Laboratory for Underwater Systems and Technologies (LABUST), Unska 3, Zagreb, Croatia

ARTICLE INFO

Keywords:
Aquaculture
Marine robots
ROVs
Biofouling
Vegetation
Deep learning

ABSTRACT

Aquaculture net pen defects such as biofouling, vegetation, and holes are key challenges to efficient and sustainable fish production in aquaculture. These defects must be monitored to avoid problems with aquaculture net safety as well as fish growth. Recently, deep learning methods have been adopted to solve detection and classification problems in different applications. However, the conventional methods are challenging to meet the demands of high precision and real-time detection of aquaculture net defects in a complex marine environment. Towards this end, this paper proposes an autonomous net pen defect detection system that contains a novel multi-scale semantic segmentation topology for detecting the biofouling, vegetation, and hole problems in the aquaculture environment. In particular, we emphasize fusing the attention maps obtained across different decomposition levels of the network to generate rich feature distributions that enable the accurate extraction of biofouling, vegetation, and holes within aquaculture nets. Moreover, the proposed model is thoroughly tested on two private datasets and two public datasets which were acquired from a real-field fish farms. Across all four datasets, the proposed framework showed remarkable biofouling, vegetation, and hole detection performance, where it outperformed state-of-the-art methods by 6.58%, 3.69%, 6.44%, and 4.78% in terms of mean average precision across LABUST, KU, NDv1, and NDv2 datasets, respectively.

1. Introduction

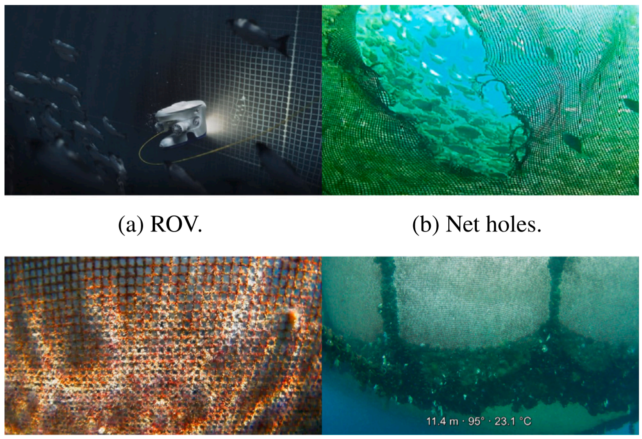
In the aquaculture environment, netting is the key component of fish farming in the open sea, but it can be easily damaged due to the dynamic marine environment. Biofouling, vegetation growth, and net holes are a few of the common net defects that cause severe economic losses. Biofouling is the colonization of various marine microorganisms such as algae, mussels, and hydroids growing on the net surface, which results in increased net weight, material damage, net deformation, and increased stress on the mooring (Fitridge, Dempster, Guenther, & De Nys, 2012). In addition, biofouling causes a reduction of the oxygen levels within the fish cages, resulting in an increased death rate for the fish. Moreover, aquatic vegetation also grows on the net surface, which causes difficulty in breathing for the fishes. In addition, net holes are one of the severe defects in the aqua-nets that allow a massive group of fish to escape into the open sea, resulting in substantial economic losses, and other marine crises. In order to enhance the efficacy and sustainability of aquaculture, it is evident that net defects detection must take place on a consistent and effective basis (Ohrem, Kelasidi, & Bloecher, 2020).

Aquaculture net pens inspection and monitoring is a hard and time-consuming task, especially when it is done manually. Moreover, it endangers the lives of the divers that carry out the manual inspection and monitoring tasks. Therefore, there is a dire need to develop autonomous systems that are capable of accurately recognizing the net pens defects using underwater imagery without any human involvement (Zion, 2012).

Currently, modern aquaculture monitoring and inspection activities are carried out through remotely operated vehicles (ROVs) (Li & Du, 2022). These vehicles are low-size and low-cost with automatic navigation and control capabilities (Akram, Casavola, Kapetanović, & Miškovic, 2022). The use of mounted-camera sensors on these vehicles

* Corresponding author.
E-mail addresses: waseem.akram@ku.ac.ae (W. Akram), taimur.hassan@adu.ac.ae (T. Hassan), hamed.toubar@ku.ac.ae (H. Toubar), muhayyuddin.ahmed@ku.ac.ae (M. Ahmed), nikola.miskovic@fer.hr (N. Miškovic), lakmal.seneviratne@ku.ac.ae (L. Seneviratne), irfan.hussain@ku.ac.ae (I. Hussain).

https://doi.org/10.1016/j.eswa.2023.121197
Received 17 May 2023; Received in revised form 21 July 2023; Accepted 11 August 2023
Available online 26 August 2023



(a) ROV. (b) Net holes.

(c) Biofouling. (d) Vegetation.

Fig. 1. Automatic aquaculture inspection using ROV. (a) shows an ROV during aquaculture net pens inspection, (b) shows an example of the hole on the net, (c) shows an example of biofouling defects on the net, and (d) shows an example of vegetation attached to the net.

allows real-time streaming of the net surface in the form of underwater marine imagery. Moreover, researchers used these imagery to develop computer vision methods that can automatically perform aquaculture inspection and monitoring tasks. The aim of these methods is to extract relevant features from the obtained visual data and subsequently perform the detection tasks, thus, providing non-invasive, objective and reproducible solutions for the aquaculture net pens inspection (Chang et al., 2021). In Fig. 1, an automatic fish farm inspection is shown in which the ROV performs the inspection tasks using the on-board camera. The use of ROVs that can provide autonomous inspection of the aquaculture environment is both time and labor saving scheme.

1.1. Motivation

Several studies have reported the use of computer vision and other visual analysis methods to detect defective regions in marine assets during the inspection and monitoring phase (Er, Chen, Zhang, & Gao, 2023). More particularly, the recent wave of methods, devoted to aquaculture net pen inspection (Kapetanović, Nađ, & Mišković, 2021; Labra et al., 2023), utilized deep learning detection models, such as YOLO (Bochkovskiy, Wang, & Liao, 2020; Jocher et al., 2022; Redmon & Farhadi, 2018), Mask R-CNN (He, Gkioxari, Dollar, & Girshick, 2017), Fast-RCNN (Girshick et al., 2015), and SSD (Liu et al., 2016). However, it should be noted that these conventional models were primarily designed to work on photographic imagery, where they employ region-based searches to detect the objects-of-interest (Liao, Zhang, Wu, An, & Wei, 2022; Zhang, Gui, Qu, & Feng, 2022). But when they are exposed to the underwater marine scans that contain very low information about the object textures (Abinaya, Susan, & Sidharthan, 2022), the performance of such models deteriorates drastically towards detecting the aquaculture net defects (Yang et al., 2021). To overcome these limitations, we propose a novel multi-scale encoder–decoder topology that uses pixel-wise recognition across various scales instead of region-based searches to effectively detect the aquaculture defects from the underwater marine scans. Moreover, the proposed model is constrained via L_t loss function which produces soft probabilities to identify the aquatic defects irrespective of their textural, color, and contextual specifications. In addition, the proposed framework is thoroughly compared with the state-of-the-art methods on four different datasets, where across each dataset, the proposed framework significantly outperformed the state-of-the-art through various metrics. More details on the contributions of this paper are presented in Section 3.

2. Related work

In this section, we present a detailed review of the recent works that are focused toward aquaculture inspection and monitoring tasks. To best present the literature, we have categorized them into two groups, where the first group of methods highlight the utilization of remotely operated vehicles (ROVs) to perform aquaculture inspection and intervention tasks. The second group of methods are focused towards employing deep learning systems in order to perform aquaculture monitoring tasks.

2.1. Utilization of ROVs for aquaculture

Recently, ROVs have become increasingly popular for use in aquaculture inspection and intervention tasks, such as biomass estimation, fish modeling, fish net pens monitoring, and other environmental monitoring (Osen, Havnegjerde, Kamsvåg, Liavaag, & Bye, 2016; Tao et al., 2018). For instance, Sato, Liu, Orita, Sakagami, and Wada (2022) developed a path-following control technique for ROVs that are used in the visual inspection of the fish net pens. Wu et al. (2022) tested an intelligent control strategy utilizing a neural network for the fish net tracking and monitoring. Another interesting work is presented by Nishitani, Kawada, and Kojima (2022) who proposed an integrated system that uses aerial and underwater vehicle to collaborate with the inspection chores. Firstly, an aerial vehicle was used to transport an ROV to the inspection site. Once the task of the ROV is finished, the aerial vehicle returns it back to its base location. To maintain the ROV's relative position and capture the net surface clearly, Lee, Jeong, Yu, and Ryu (2022) used the neural network model to design the ROV position control strategy.

The inspection of a fishnet pen relies on the ROV's pose being accurately estimated with respect to the net. Regarding pose estimation, Duda, Schwendner, Stahl, and Rundtop (2015) presented an X-junction detection (fastcross) approach, in which the knots of fishnets are recognized from camera images and their locations are estimated in the image plane. The implementation results in a lab environment showed a lower standard deviation from the reference position. Similarly, Rundtop and Frank (2016) used ultra-short baseline (USBL) and Doppler Velocity Log (DVL) sensors to estimate the vehicle position with respect to the net pens. However, this approach showed performance degradation in the field due to the noisy sensory data. Amundsen, Caharija, and Pettersen (2021) employed a DVL sensor on an Autonomous Underwater Vehicle (AUV) to calculate the position of the region-of-interest (ROI) in the net with respect to the vehicle. After calculating the position of ROI, the AUV's position and orientation are determined relative to the approximated plane and sent into a line-of-sight (LOS)-based navigation algorithm. However, noise in the DVL sensor reduces the effectiveness of this approach (Amundsen et al., 2021). Furthermore, these sensors are too bulky and expensive to be installed on the smaller AUVs (Amundsen et al., 2021).

Alternatively, many studies have installed optical cameras on the ROVs, due to their small size, cost effectiveness, and their ability to capture high quality images. Moreover, the net inspection tasks through these optical cameras are performed via vision-based algorithms due to their increased precision (Lee, Kim, Kim, Myung, & Choi, 2012; Li, Jiang, Cao, Wang, & Li, 2015). One of such works is presented by Livanos et al. (2018) where they used the visual markers to estimate the position of the ROV relative to the fish net pens. Using an omnidirectional surface vehicle equipped with sonar for location determination, Tao et al. (2018) created a revolutionary monitoring system. In this work, a vehicle-mounted, moveable camera with depth adjustment was also installed that allows the vehicle to inspect the fish net at different depths. Similarly, Duecker, Hansen, and Kreuzer (2020) worked on a distance estimation between an ROV and net pens using point clouds. To determine the position of an ROV relative to the net pens, Bjerkeng et al. (2021) proposed the utilization of camera triangulation. Apart from this, other work on the net pens inspection and monitoring, employing custom control schemes can be found in Dalhatu, Sa'ad, Cabral de Azevedo, and De Tomi (2023) and Wu et al. (2022).

2.2. Deep learning systems for aquaculture

Recently, deep learning-based image classification and object detection methods have been introduced for aquaculture monitoring tasks (Yang et al., 2021), including but not limited to, fish net hole detection (Zhang et al., 2022), biofouling detection (Fitridge et al., 2012), and vegetation detection (Chabot, Dillon, Shemrock, Weissflog, & Sager, 2018; Perrin et al., 2022).

For instance, Hu et al. (2021) worked on the automation of feeding strategies to efficiently and effectively process feeding and reduce feed waste and water pollution. This work proposed a real-time detection and monitoring system for feed pellet consumption in fish cages using improved YOLOv4 (Bochkovskiy et al., 2020). Moreover, the results obtained by the improved YOLOv4 model demonstrated better performance as compared with the original YOLOv4 model (Bochkovskiy et al., 2020) by achieving an increase of 27.21% average precision. Abinaya et al. (2022) worked on biomass estimation in the fish cages, where they proposed the use of the YOLOv4 (Bochkovskiy et al., 2020) model to detect the head, body and tail of the fish and then the nearest neighbor association technique to develop a sequence for fish length estimation. The proposed method demonstrated a higher mAP of 91.52%. Voskakis, Makris, and Papandroulakis (2021) also performed biomass estimation using a stereo camera. The images obtained from the stereo camera were fed as an input to the system from which the fish length is estimated via CNN model. The system was tested on a real dataset and resulted in a small error of 0.3% compared with the actual measurement. Rosales et al. (2021) used Faster R-CNN (Ren, He, Girshick, & Sun, 2015) to detect fishes, where they achieved a mAP score of 0.99 on a private dataset taken in fishpond. Wang et al. (2022) worked on the fish behavior detection task in the fish cage which is an effective way to ensure the welfare and safety of the fish. The study proposed an improved YOLOv5 (Jocher et al., 2022) model. In the backbone network, the multi-head attention mechanism is used instead of the Cross-Stage Partial (CSP) (Wang et al., 2020) structure to obtain global information. The proposed improved model was trained and tested on a real dataset for fish behavior detection tasks considering blur and small-sized targets. The results of this method showed better accuracy of 97% compared to the original model. Similarly, fish behavior detection was also studied in Hu, Zhao, Zhang, Zhou and Chen (2021) by considering an improved YOLOv3 (Redmon & Farhadi, 2018) by modifying the network backbone with the MobileNetV2. The work showed better performance by achieving 10%–20% improvements in terms of precision and recall compared with the original YOLOv3 (Redmon & Farhadi, 2018). Ahmed, Aurpa, and Azad (2021) worked on fish disease detection by using a support vector machine (SVM) model. The salmon fish dataset was acquired and various image pre-processing was performed to extract useful features that demonstrate fish health, e.g., fresh or infected. Moreover, the study was also validated on the real dataset that showed better performance of 94.12% accuracy for the identification of fresh and infected fishes. In aquaculture, water quality also plays a vital role in fish growth (Haq & Harigovindan, 2022). The unsuitable water can cause fish deaths which eventually can result in an economic loss. Tran-Quang and Ha-Ngoc (2021) proposed long short-term memory networks (LSTM) model (Matsubara & Torikai, 2015) to predict the environmental parameters such as temperature, salinity, and pH to provide a real-time water monitoring system.

Although recent works are focused towards using deep learning models for the detection of fish and biofouling within the aquaculture environment. But, since the methods proposed in these works were built using conventional detection models that were primarily designed for photographic imagery, the performance of these methods deteriorates when they are applied to underwater imagery, which contains complex textural and contextual features of the aquaculture environment (Liao et al., 2022; Zhang et al., 2022). To overcome these issues, we propose a novel multi-scale semantic segmentation model that accurately recognizes different regions of aquaculture defects depicted within the complex underwater scans. A detailed description of the proposed model and the contributions of this work are presented in the subsequent sections.

3. Contributions

In this paper, we solve the aquaculture net pen defect detection problem, including biofouling, vegetation, and hole detection, through a multi-scale encoder–decoder topology. Unlike the competitive methods, we leverage detection by segmentation with the proposed network to effectively recognize the presence of net defects across the pixel level and detect them autonomously. Furthermore, the proposed network fuses attentional features of the input scan across different decomposition levels in order to boost the extraction of biofouling, vegetation, and holes in the aquaculture nets. The proposed model is constrained via L_t loss function, which produces soft probabilities to give more exposure to the model for identifying aquatic defects irrespective of their textural, color, and contextual specifications.

Apart from this, the proposed framework is thoroughly tested on four datasets to detect aquaculture net pen defects. The first dataset that we used to validate the proposed framework was adopted from the real field of fish farming based in Zadar, Croatia, by Cromaris Fishery. Apart from this, we also collected data at the marine pool based at Khalifa University, UAE. To further evaluate the performance of the proposed scheme, we also tested the proposed framework on two publicly available datasets of aquaculture net defects, namely NDv1 (Roboflow, 2022a), and NDv2 (Roboflow, 2022b). The first two datasets are also one of the contributions of this work, which we will release publicly to facilitate the research community upon acceptance of the paper. Moreover, the performance of the proposed method towards biofouling, vegetation, and hole detection is thoroughly evaluated through different metrics, which not only confirm the potential of the proposed model towards robust aqua-net defect detection but also provide ease to the end user by making the detection and cleaning processes faster and more efficient during fish farming. The proposed framework is also compared with the state-of-the-art methods across all four datasets, where it fairly outperformed them in terms of mAP, μIoU, and μDC scores. To summarize, the main contributions of this paper are:

- A novel attempt to detect aquaculture defects such as biofouling, vegetation, and net holes, by segmentation via proposed multi-scale encoder–decoder topology.
- A new loss function, dubbed L_t, that can constrain the segmentation networks to extract aquatic defects from the underwater marine scans irrespective of their textural and contextual characteristics.
- A release of the two highly annotated aquaculture datasets that can facilitate the research community in developing and validating the deep learning models for aquaculture defects detection. The datasets will be released publicly upon the acceptance of the paper.

The rest of the paper is organized as follows: Section 4 presents the detailed description of the proposed approach. Section 5 presents the experimental setup. The experimental results and discussions are discussed in Section 6. Lastly, Section 7 concludes the paper and highlight the prospects of the proposed framework.

4. Proposed approach

In this section, we discuss the detailed description of the proposed model, and the L_t loss function through which we constrained the proposed network in order to perform robust extraction of aquaculture defects across four different datasets.

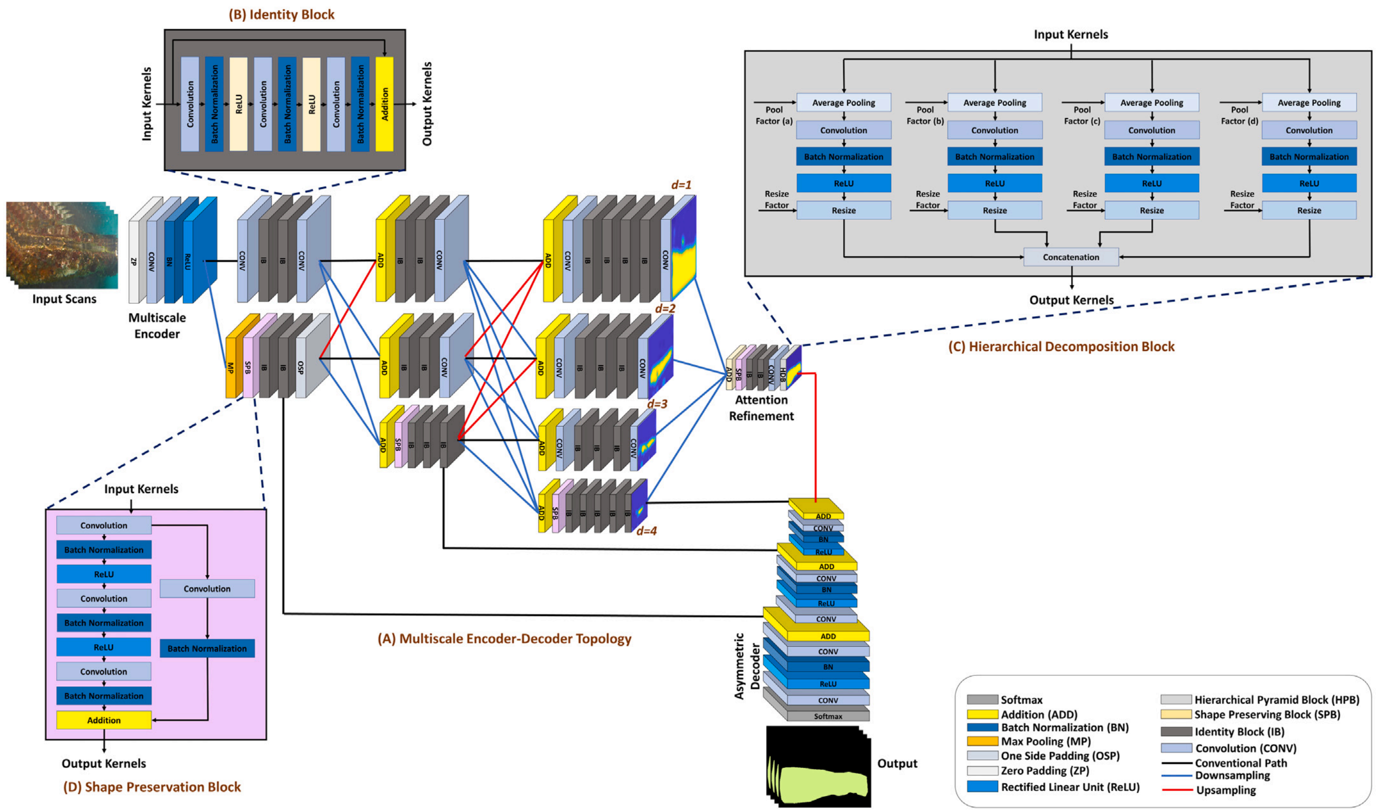


Fig. 2. The architectural diagram of the proposed multi-scale encoder–decoder topology. When the input scan is fed to the proposed model, the encoder backbone first extracts distinct feature representations of the aquaculture defects across various depths of the network via custom identity, shape preservation, and hierarchical decomposition blocks. These blocks refine the attention of the model to focus on the defect regions irrespective of the scan characteristics. Moreover, these latent features are then passed to the asymmetric decoder that accurately segments aquaculture defects from the candidate scan.

4.1. Proposed multi-scale encoder-decoder network

We designed a novel multi-scale encoder–decoder pipeline to perform aquaculture defects detection using underwater marine imagery. The architectural description of the proposed model is shown in Fig. 2. Here, we can observe that the encoder backbone is used to generate the latent space representations by distributing feature maps across different network depth levels. In contrast to traditional encoders, the proposed encoder can maintain the high-resolution features of the candidate input by summing features maps across each depth in a butterfly manner via upsampling and downsampling the kernel sizes as needed. Moreover, each block within the proposed network consists of custom identity blocks (IB), hierarchical decomposition blocks (HDB), and shape-preservation blocks (SPB). These blocks refine the attention of the model so that it only focuses on the defected regions, irrespective of the scan's textural and contextual attributes. Moreover, in contrast to other state-of-the-art networks (Wang et al., 2021; Yu et al., 2021), in which the latent vectors are generated by the high-resolution feature maps, the proposed encoder design permits the generation of latent features across any depth of the network. Such design schematics allow the decoder end to not only rely on the features generated from the top layers of the encoder, which can be pruned to noise and vendor artifacts. But also to analyze the fine-grained features generated across various depths of the encoder end that allow the decoder to preserve the context and shape of the defected regions for their accurate extraction. Apart from this, the proposed model does not use bilinear decimation or interpolation but rather convolutional layers with a varying stride size to carry out the upsampling and downsampling process, as the strided convolutions produce more accurate segmentation results than the bilinear decimation or interpolation schemes (Yu et al., 2021). Apart from this, in the subsequent sections, we elaborate further on the custom blocks that are used within the proposed model.

4.2. Identity Block (IB)

The identity block, used within the proposed model, is inspired from residual blocks in ResNets (He, Zhang, Ren, & Sun, 2016). Architecturally, each IB block combines three convolutional layers, three batch normalization layers and two ReLUs in a cascaded fashion, and fuses with them the input kernels in a residual manner (see Fig. 2). Such design allows the generation of discriminative features for locating the defective areas within the input scans. In addition, it aids in maintaining the object's geometrical properties even while dealing with noisy data.

4.3. Hierarchical Decomposition Blocks (HDB)

Fig. 2(C) shows the proposed hierarchical decomposition block (HDB) that is added at the end of the encoder backbone. HDB consists of a multi-level decomposition architecture containing four average pooling, four convolution, four batch normalization, four ReLU, and four resizing factors. Such design layout aids the proposed model in fusing coarser to finer representations of the aqua-net defects across several scales, making it possible for the encoder end to generate rich feature representations that can significantly aid the proposed model in producing a detailed and accurate segmentation of aqua-net defects as an output.

4.4. Shape-Preservation Blocks (SPB)

The SPB block within the proposed model consists of a total of four convolutional layers, four batch normalization layers, and two ReLU activations (see Fig. 2), and it aids the network in preserving the shape of the aquatic defect regions within the input scan. Furthermore, SPB

Table 1
Datasets description.

Dataset	No. of Images	Biofouling	Vegetation	Holes
LABUST	156,600	Yes	Yes	No
KU	37,926	No	Yes	Yes
NDv1	1864	Yes	Yes	No
NDv2	1411	Yes	Yes	No

block employs a residual connection that fuses the output of a single convolutional layer and a single batch normalization layer with the cascaded combination of convolutional, batch normalization and ReLU layers. This kind of design topology helps in preserving the shape of the aquatic defects by generating the rich contextual and geometrical features. Moreover, it also ensures that the model learns to discriminate between similar-textured regions belonging to different categories of the aquatic defects during the training process.

4.5. Training strategy

During training, the model is constrained via proposed L_t loss function to extract aquaculture defect regions from the underwater imagery. The L_t loss function is the sum of two objective functions, dubbed L_{s1} and L_{s2}. Including these sub-objectives into the loss function allows the model to be trained while being exposed to a wider range of underlying net defects. For instance, in the event of an unbalanced distribution of background and foreground pixels within the input scan, yielding the size of the defect regions to be in a much smaller size compared to the background region, L_{s1} effectively minimizes the pixel-level errors in such cases, enabling the underlying model to carry out the segmentation tasks despite the presence of imbalance distribution of foreground and background pixels. However, achieving the convergence through L_{s1} is a difficult task because of the fact that the gradient of L_{s1} can overshoot in case we get smaller values of predicted logits and ground truths. This problem is resolved by introducing L_{s2} in the L_t loss function that enables the model to converge even in the presence of smaller values of predicted logits and the ground truths. In addition to this, the trade-off between L_{s1} and L_{s2} within L_t is administered through the β_1 and β_2 hyperparameters. Mathematically, the objective functions are expressed as follows:

$$L_t = \beta_1 L_{s1} + \beta_2 L_{s2}, \tag{1}$$

where

$$L_{s1} = \frac{1}{b_s}\sum_{i=0}^{b_s-1}\left(1 - \frac{2\sum_{j=0}^{c_{se}-1} T_{i,j}^{se}\, p(\mathcal{L}_{i,j}^{se,\tau})}{\sum_{j=0}^{c_{se}-1}\left((T_{i,j}^{se})^2 + p(\mathcal{L}_{i,j}^{se,\tau})^2\right)}\right), \tag{2}$$

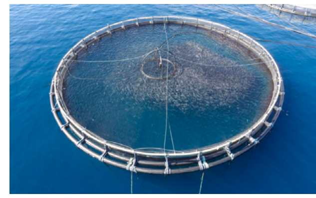
and

$$L_{s2} = -\frac{1}{b_s}\sum_{i=0}^{b_s-1}\sum_{j=0}^{c_{se}-1} T_{i,j}^{se}\log(p(\mathcal{L}_{i,j}^{se,\tau})). \tag{3}$$

$T_{i,j}^{se}$ shows the ground truth label for the ith sample belonging to jth category of aqua-net defects, i.e., biofouling, vegetation, and net holes, $p(\mathcal{L}_{i,j}^{se,\tau})$ denote the predicted probability distribution obtained from the output logit $\mathcal{L}_{i,j}^{se,\tau}$ belonging to ith sample and jth net defects category, τ denotes the temperature constant that softens the probabilities generated through the softmax function to ensure robust learning of recognizing aqua-net defects, b_s shows the batch size, and c_{se} shows the total classes.

5. Experimental setup

This section presents the details about the datasets, the implementation protocols, as well as the evaluation metrics.



(a) Fish farm.

(b) Blueye Pro ROV.

Fig. 3. Data acquisition at Crosmaris fish farms, Croatia.

5.1. Datasets

We used four different datasets to evaluate the performance of the proposed framework, and compare it with the state-of-the-art methods. All of these four datasets includes the aquaculture defects like biofouling, vegetation, and holes defects depicted within the underwater imagery. Table 1 summarizes all of the four datasets, and their detailed description is presented below:

5.1.1. LABUST dataset

The first dataset on which we evaluated the proposed framework is the LABUST dataset. It was taken from the Laboratory for Underwater Systems and Technologies (LABUST) at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. The data was collected as a part of the research activity of the Heterogenous Autonomous Robotic System in Viticulture and Mariculture (HECTOR) project (Kapetanović et al., 2022). This project is aimed at deploying aerial vehicles, surface vehicles and ROVs in an integrated and coordinated manner to accomplish surveying and inspection missions both in viticulture as well as in the mariculture domain.

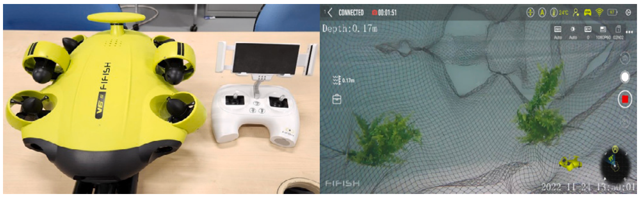
The data acquisition was conducted at the commercial fish farming sites of Cromaris Fishery, a leading positioned company in the production and processing of high-quality Mediterranean fish located in Zadar, Croatia. Blueye Pro ROV, developed by Blueye robotics company, was deployed during the experiments for the data acquisition. The ROV is equipped with a battery allowing two hours capacity and can go down 300 m in depth with 400 m tethering cable in saltwater, brackish or freshwater. A full HD forward-looking camera with 25–30 frames per second (fps) is also installed allowing the production of high-quality video streaming of the environment (see Fig. 3).

During the inspection campaign, the vehicle is allowed to move from top to bottom and around the fish cage to capture the current status of the net surface. The captured video has a size of 1920 × 1080 with 30fps. The image sequences were extracted from the video by taking one frame every five seconds from the video. Apart from this, the dataset contains around 156,600 scans which were acquired from the camera mounted on the ROV. From these scans, we used 31,320 scans for training purposes, and the rest for testing purposes.[1]

5.1.2. KU dataset

The second dataset that is used in this work was collected from the net pool at Khalifa University, United Arab Emirates, using the FiFish ROV, as shown in Fig. 4. This ROV is equipped with an HD camera, allowing seamless recording of underwater imagery. Moreover, during the data acquisition phase, the ROV is controlled and moved around the net via a remote controller. The acquired scans were then annotated in a pixel-wise manner, using the MATLAB Image Labeler application, to identify vegetation, holes, and clean net regions. In addition, the total

[1] The annotated LABUST, KU datasets along with the complete pixel-wise annotations for the NDv1 and NDv2 datasets will be released publicly upon the acceptance of the paper.



(a) FiFish ROV. (b) Aqua-net input scan during experiment.

Fig. 4. KU dataset collection. (a) shows a FiFish ROV used in this paper. (b) shows an ROV during aquaculture net pens inspection.

of scans within the KU dataset are 37,926, from which 7421 scans are to be used for training and the rest of 30,505 scans are to be used for testing purposes.[1] It is worth mentioning that the KU dataset includes underwater scans that contain extreme water disturbances, such as light reflections, jerky water waves, and water ripples. These disturbances cause distortions and artifacts, making it difficult for the deep learning models to accurately detect and identify aquaculture defect regions.

5.1.3. NDv1 dataset

The third dataset that we used in this work is NDv1 (Roboflow, 2022a). The NDv1 dataset is publicly available and it consists of a total of 1864 underwater images containing aqua-net defects, such as, biofouling and vegetation. Moreover, from 1864 scans, 1307 scans are to be used for training, and 194 scans are to be used for testing purposes, and 363 scans are to be used for validation as per the dataset standards. A summary of the dataset is reported in Table 1. The images within NDv1 contain a diverse range of textures, brightness, and variations within net material and type (Roboflow, 2022a). In the images, biofouling and vegetations classes can be observed either separately, grouped together, or intersecting with each other. Initially, the dataset did not have pixel-wise annotations for the biofouling, vegetation, and hole extraction. However, we managed to annotate the dataset ourselves using the MATLAB Image Labeler application.[1]

5.1.4. NDv2 dataset

The fourth dataset which we used in this work is the NDv2 dataset (Roboflow, 2022b). NDv2 is also publicly available and it includes biofouling and vegetation classes as described in Table 1. Moreover, the dataset contains a total of 1411 underwater scans from which 854 scans are to be used for training, 194 scans are to be used for testing, and 363 scans are to be used for validation purposes as per the dataset standards. Moreover, the scans within NDv2 contain uneven illumination, noise, and object size differences. Apart from this, like NDv1 dataset, NDv2 also did not have pixel-wise annotations for training the segmentation model for extracting the biofouling and vegetation defects. Therefore, we managed to mark the pixel-wise annotations for the NDv2 dataset for extracting aquaculture defects using the MATLAB Image Labeler application.[1]

5.2. Implementation protocols

The proposed framework is implemented on Core i9-10940@3.30 GHz processor, 128 GB RAM and a single NVIDIA Quadro RTX 6000 GPU having a CUDA toolkit v11.0.221 and cuDNN v7.5. Moreover, the proposed model is developed using Python 3.7.9 and TensorFlow 2.1.0. Also, it is trained for 40 epochs having 512 iterations in each epoch, where for LABUST, and KU datasets, we used 20% of the training data for validation, and for NDv1, and NDv2 datasets, we used 363 scans for validation purposes, as per the dataset standards. The optimizer used during the training was ADADELTA (Zeiler, 2012) having a default learning rate and decay rate of 1.00 and 0.95, respectively.

5.3. Evaluation metrics

In order to assess the performance of the proposed framework towards identifying biofouling, vegetation, and net holes regions, we used mean intersection-over-union (μIoU), mean dice coefficient (μDC), and mean average precision (mAP) scores.

6. Results and discussions

In this section, we present a detailed experimental results of the proposed framework across all four datasets. Three of these datasets (LABUST, NDv1 (Roboflow, 2022a), and NDv2 (Roboflow, 2022b)) were acquired from the real-field marine fish farms, whereas, the fourth one (KU dataset) was acquired from the pool environment within Khalifa University. Moreover, the proposed framework was primarily tested on these datasets to detect biofouling, vegetation (both horizontal and vertical vegetation) and holes within the aquaculture nets, where its detection performance is also compared with the state-of-the-art methods under fair experimental settings.

At first, we discuss the ablation experiments through which we determined the hyperparameters of the proposed framework. Then, we present quantitative and qualitative evaluations through which we measured the performance of the proposed framework and compared it with the state-of-the-art models. Lastly, we present the detailed discussion on the prospect of the proposed framework and its limitations.

6.1. Ablation studies

We performed an extensive list of ablative experiments to determine (1) optimal β values in L_t, (2) best encoder backbone, (3) optimal encoder depth, (4) the best value of temperature constant (τ) in L_t, and (5) the optimal loss function, that yield better performance for the proposed framework to detect aqua-net defects.

6.1.1. Optimal β values in L_t

The first series of ablation experiments are related to determining the optimal hyper-parameters $\beta_{1,2}$ in L_t that produce the best segmentation performance across each dataset. For β_1, we varied the values ranging from 0.1 to 0.9 with an increment of 0.2, and for each value of β_1, we computed β_2 through $\beta_2 = 1 - \beta_1$. Apart from this, we trained the proposed model with each combination of β_1 and β_2 parameter, and then the segmentation performance of the model, for each combination, is observed at the inference stage, across each dataset, in terms of mAP scores, as reported in Table 2. Here, it can be noticed that the better performance of the proposed framework is achieved with larger weight of β_1, i.e., 0.9 within the L_t loss function. For example, with $\beta_1 = 0.9$, and $\beta_2 = 0.1$, the proposed model achieved the mAP scores of 0.7858, 0.8169, 0.7144, and 0.7278, across LABUST, KU, NDv1, and NDv2 datasets, respectively. Thus, in the rest of the experiments, we used $\beta_1 = 0.9$ and $\beta_2 = 0.2$ to train the proposed model.

6.1.2. Optimal encoder backbone

The second series of ablation experiments are related to identifying the optimal network backbone for aquaculture defects detection tasks. For this purpose, we plugged different pre-trained models, such as HRNet (Wang et al., 2021), Lite-HRNet (Yu et al., 2021), EfficientNet-B4 (Tan & Le, 2019), DenseNet-201 (Huang, Liu, Van Der Maaten, & Weinberger, 2017), and ResNet-101 (He et al., 2016), within the proposed model, and compared their performance with the proposed backbone which we specifically designed to detect aquaculture defects using the underwater imagery. The results for these experiments are reported in Table 3. From Table 3, we can observe that the proposed encoder demonstrated better results as compared with the state-of-the-art models where it outperformed them by 3.22%, 1.96%, 3.34%, and 2.54% in terms of μIoU, μDC, mAP, and AUC scores, respectively,

Table 2
Determining the optimal $\beta_{1,2}$ values that give the best segmentation performance. Moreover, bold indicates the best performance in terms of mAP score, while '-' indicates that the respective combination of β_i and β_j is invalid. The abbreviation 'LB' represents LABUST dataset.

LB	β_1	0.1	0.3	0.5	0.7	0.9
β_2	0.1	–	–	–	–	**0.7858**
	0.3	–	–	–	0.7531	–
	0.5	–	–	0.7126	–	–
	0.7	–	0.6964	–	–	–
	0.9	0.6483	–	–	–	–
KU	β_1	0.1	0.3	0.5	0.7	0.9
β_2	0.1	–	–	–	–	**0.8169**
	0.3	–	–	–	0.7745	–
	0.5	–	–	0.7508	–	–
	0.7	–	0.7283	–	–	–
	0.9	0.6824	–	–	–	–
NDv1	β_1	0.1	0.3	0.5	0.7	0.9
β_2	0.1	–	–	–	–	**0.7144**
	0.3	–	–	–	0.6821	–
	0.5	–	–	0.6573	–	–
	0.7	–	0.6249	–	–	–
	0.9	0.5968	–	–	–	–
NDv2	β_1	0.1	0.3	0.5	0.7	0.9
β_2	0.1	–	–	–	–	**0.7278**
	0.3	–	–	–	0.6924	–
	0.5	–	–	0.6702	–	–
	0.7	–	0.6383	–	–	–
	0.9	0.6087	–	–	–	–

Table 3
Determining the best backbone network to perform aquaculture defect detection tasks across all datasets. The input size for all the model was 576 × 768 × 3. Moreover, the abbreviations are: DS: Dataset, LB: LABUST, ND1: NDv1, ND2: NDv2, PEB: Proposed Encoder Backbone, LHR: Lite-HRNet-30 (Yu et al., 2021), HRN: HRNetv2 (Wang et al., 2021), EN4: EfficientNetB4 (Tan & Le, 2019), D201: DenseNet-201 (Huang et al., 2017), R101: ResNet-101 (He et al., 2016).

DS	BB	μIoU	μDC	mAP	AUC	Params
LB	PEB	**0.8265**	**0.9050**	**0.7858**	**0.9492**	56.2M
	LHR	0.7721	0.8714	0.7337	0.9013	6.84M
	HRN	0.7943	0.8854	0.7524	0.9238	50.9M
	EN4	0.7409	0.8512	0.7139	0.8756	21.7M
	D201	0.7263	0.8415	0.6846	0.8419	96.5M
	R101	0.7075	0.8287	0.6607	0.8204	85.1M
KU	PEB	**0.9602**	**0.9795**	**0.8169**	**0.9668**	58.1M
	LHR	0.9042	0.9497	0.7528	0.9316	9.25M
	HRN	0.9126	0.9543	0.7671	0.9443	53.4M
	EN4	0.8847	0.9388	0.7354	0.9058	22.8M
	D201	0.8625	0.9262	0.7218	0.8874	98.3M
	R101	0.8503	0.9191	0.7045	0.8715	89.6M
ND1	PEB	**0.7730**	**0.8720**	**0.7144**	**0.9285**	57.4M
	LHR	0.7216	0.8383	0.6583	0.8806	8.94M
	HRN	0.7382	0.8494	0.6726	0.9073	52.6M
	EN4	0.7069	0.8283	0.6241	0.8635	21.5M
	D201	0.6864	0.8140	0.5865	0.8407	97.3M
	R101	0.6623	0.7968	0.5607	0.8226	89.2M
ND2	PEB	**0.8087**	**0.8942**	**0.7278**	0.9314	57.4M
	LHR	0.7483	0.8560	0.6614	0.9024	8.93M
	HRN	0.7605	0.8640	0.6853	**0.9315**	52.6M
	EN4	0.7231	0.8393	0.6402	0.8974	21.5M
	D201	0.7069	0.8283	0.5996	0.8693	97.3M
	R101	0.6843	0.8126	0.5785	0.8521	89.2M

across LABUST dataset. Similarly, across KU dataset, the proposed framework achieved the performance improvement of 5.60%, 2.52%, 4.98%, and 2.25%, in terms of μIoU, μDC, mAP, and AUC scores, respectively. The performance gains achieved on the NDv1 dataset is 3.48%, 2.26%, 4.18%, and 2.12% in terms of μIoU, μDC, mAP, and AUC scores, respectively. Moreover, we also observed the performance gains across NDv2 dataset: 4.82% of μIoU, 3.02% of μDC, and 4.25% of mAP, while in terms of AUC, the model showed 2.90% performance improvement over Lite-HRNet (Yu et al., 2021). These significant performance improvements are achieved because of the fact that the proposed encoder backbone employed a novel SPB, IB, and HDB blocks in a butterfly structure allowing the model to compute distinct latent features from the input images. The integration of these novel blocks in the network model yields better performance for aquaculture defects detection tasks compared to the state-of-the-art models (He et al., 2016; Huang et al., 2017; Wang et al., 2021; Yu et al., 2021). Although, the proposed scheme is computationally expensive compared to Lite-HRNet (Yu et al., 2021). However, at the same time, it significantly outperformed Lite-HRNet (Yu et al., 2021) in terms of detection performance. Since our prime objective here was to achieve maximum detection performance. Therefore, we chose the proposed encoder backbone to generate distinct feature representations in the rest of the experiments.

6.1.3. Optimal encoder depth

The proposed model is capable of generating latent features from any depth of the encoder backbone. This unique attribute makes the proposed model flexible and robust towards detecting the aquaculture defects, as compared to its competitors. Furthermore, most of the models that are proposed in the literature only focus on the high-resolution channels for features generation, and these features may also be susceptible to unwanted high-frequency contents, such as noise and vendor specifications (Wang et al., 2021). Thus, using such models as feature extractor within the proposed model may cause noticeable performance degradation as evident from Table 3. To overcome these issues, we propose generating the multi-scale latent features (from different depths of the encoder backbone), and fuses them together via addition (see Fig. 2) to effectively detect the aquaculture defects from the underwater imagery. Additionally, we performed another series of experiments in which we analyzed the performance improvements which the proposed network can achieve by using these fused features to extract aquaculture defects as compared to the variants in which the latent features are generated from each depth of the encoder individually. The results for these experiments are reported in Table 4, in which we can observe that, across each dataset, the proposed model produced a better performance as compared to the other variants in terms of μIoU, μDC, mAP, and AUC scores. For example, across the LABUST dataset, the fused features led to the performance improvement of 1.89% in terms of μIoU, 1.14% in terms of μDC, 3.09% in terms of mAP, and 2.35% in terms of AUC scores, as compared to other variants. Similarly, across KU dataset, we observed the performance improvement of 1.74% in terms of μIoU, 0.89% in terms of μDC, 2.16% in terms of mAP, and 2.42% in terms of AUC scores, as compared to other variants. The proposed model showed performance improvement of 2.48%, 1.60%, 3.15%, and 2.42% in terms of μIoU, μDC, mAP, and AUC scores, respectively. Moreover, we also noticed the performance improvement of 2.83%, 1.75%, 3.16%, and 1.58% in terms of μIoU, μDC, mAP, and AUC scores, respectively. The results in Table 4 illustrate that increasing the network depth from 1 to 4 leads to improved performance. However, when generating latent features solely from a single depth of the network, we observed a decline in performance, as evident in Table 4. These performance improvements emanate due to the fact that the fused features are inherently resistant to the noisy artifacts, because of the inclusion of low-scale coarser representations of the candidate scans, while they can simultaneously capture the finer representations of aquaculture defects, that lead to their accurate extraction at the decoder end. Moreover, generating the latent features from the single depth of the network leads to the performance degradation as it is evident from Table 4. Therefore, in the rest of the experimentation, we used fused features to generate the latent representation of the aquaculture defect regions.

Table 4
Determining the optimal depth of the network to generate latent feature representations across different datasets.

Dataset	d	μIoU	μDC	mAP	AUC
LABUST	1	0.7564	0.8613	0.7034	0.8195
	2	0.7753	0.8734	0.7152	0.8432
	3	0.7821	0.8777	0.7306	0.8746
	4	0.8076	0.8936	0.7549	0.9059
	Fused	**0.8265**	**0.9050**	**0.7858**	**0.9294**
KU	1	0.8875	0.9404	0.7086	0.8703
	2	0.9039	0.9495	0.7394	0.8954
	3	0.9174	0.9569	0.7629	0.9194
	4	0.9428	0.9706	0.7953	0.9426
	Fused	**0.9602**	**0.9795**	**0.8169**	**0.9668**
NDv1	1	0.6643	0.7983	0.6116	0.8239
	2	0.6975	0.8218	0.6357	0.8451
	3	0.7219	0.8385	0.6604	0.8726
	4	0.7482	0.8560	0.6829	0.9043
	Fused	**0.7730**	**0.8720**	**0.7144**	**0.9285**
NDv2	1	0.6976	0.8219	0.6347	0.8448
	2	0.7281	0.8427	0.6516	0.8629
	3	0.7592	0.8631	0.6794	0.8972
	4	0.7804	0.8767	0.6962	0.9156
	Fused	**0.8087**	**0.8942**	**0.7278**	**0.9314**

Table 5
Determining the temperature constant (τ) that leads to the best aquaculture defect detection performance across each dataset. Here, we used the proposed backbone encoder to generate the latent features with different values of τ.

Dataset	τ	μIoU	μDC	mAP	AUC
LABUST	1	0.7824	0.8779	0.7632	0.9063
	1.5	**0.8265**	**0.9050**	**0.7858**	**0.9492**
	2	0.7759	0.8738	0.7515	0.9128
	2.5	0.7618	0.8648	0.7362	0.8957
KU	1	0.8718	0.9315	0.7652	0.9126
	1.5	**0.9602**	**0.9795**	**0.8169**	**0.9668**
	2	0.8549	0.9218	0.7408	0.9049
	2.5	0.8401	0.9131	0.7296	0.8774
NDv1	1	0.7528	0.8590	0.6859	0.8718
	1.5	**0.7730**	**0.8720**	**0.7144**	**0.9285**
	2	0.7594	0.8632	0.6925	0.8863
	2.5	0.7413	0.8514	0.6739	0.8504
NDv2	1	0.7761	0.8739	0.6942	0.8875
	1.5	**0.8087**	**0.8942**	**0.7278**	**0.9314**
	2	0.7824	0.8779	0.7013	0.8921
	2.5	0.7629	0.8655	0.6784	0.8756

6.1.4. Determining the optimal temperature constant

The temperature constant (τ) is a hyper-parameter within the proposed L_t loss function that softens the target probabilities, hence giving more exposure to the model towards recognizing the aquaculture defect regions irrespective of the scanner characteristics. In the fourth series of ablation experiments, we determined the optimal value for τ that leads to the accurate extraction of aquaculture defects. For this purpose, we first varied the value of τ from 1 to 2.5 with an incremental of 0.5 increment within the L_t loss function, and then continued with training the proposed model across each dataset. Once the training is completed, we measured the aquaculture defects detection performance of the proposed framework across each dataset at the inference stage. The results for these experiments are reported in Table 5. We can observe from Table 5 that when the value of τ is increased from 1 to 1.5, the proposed framework achieved significant performance boost across all four datasets. For example, it achieved the performance improvement of 4.41% in terms of μIoU, 2.71% in terms of μDC, 2.26% in terms of mAP, and 4.29% in terms of AUC scores across the LABUST dataset. Similarly, across KU dataset, it achieved the performance boost of 8.84% in terms of μIoU, 4.80% in terms of μDC, 5.17% in terms of mAP, and 5.42% in terms of AUC scores. Experiments across NDv1 dataset demonstrated the performance improvement of 1.72% in terms of μIoU, 1.30% in terms of μDC, 2.85% in terms of mAP, and 5.67% in terms of AUC scores. Moreover, experiments across NDv2 dataset has shown the performance improvement of 3.26% in terms of μIoU, 2.03% in terms of μDC, 3.36% in terms of mAP, and 4.39% in terms of AUC scores. It should also be noticed that increasing τ does not always result in the performance improvements. For example, when we increased the value of τ from 1.5 to 2, and from 2 to 2.5, the performance of the proposed framework deteriorated. This performance degradation is explained by the fact that when we increase the value of τ beyond a certain threshold, it loses its ability to accurately differentiate between logits that represent different categories (e.g., the aquaculture defect regions and the background) within the underwater imagery. Moreover, since the optimal detection results were achieved with $\tau = 1.5$ for the proposed framework across each dataset. Therefore, we chose to train the proposed framework with $\tau = 1.5$ in order to perform the rest of the experimentation.

6.1.5. Optimal loss function

The fifth series of ablation experiments are related to analyzing the performance of the proposed model when it is trained using L_t as compared to other state-of-the-art loss functions, such as the soft nearest neighbor L_{sn} (Frosst, Papernot, & Hinton, 2019), the focal Tversky loss function L_{ft} (Abraham & Khan, 2019), the dice-entropy loss function L_{de} (Raja, Hassan, Akram, & Werghi, 2020), and the conventional cross-entropy loss function L_{ce}. The results for these experiments are shown in Table 6. From Table 6, we can observe that the proposed model, trained using L_t loss function, outperformed its variants which are trained using state-of-the-art loss functions across each dataset. For example, on LABUST dataset, L_t led to the performance improvement of 2.53% in terms of μIoU, 1.54% in terms of μDC, 2.65% in terms of mAP, and 2.46% in terms of AUC scores. Similarly, across the KU dataset, it led to the performance improvement of 2.84% in terms of μIoU, 1.48% in terms of μDC, 4.16% in terms of mAP, and 1.93% in terms of AUC scores. On NDv1 dataset, the proposed model led to the performance improvement of 2.11% in terms of μIoU, 1.36% in terms of μDC, 3.15% in terms of mAP, and 3.57% in terms of AUC scores. Moreover, on NDv2 dataset, it led to the performance improvement of 3.61% in terms of μIoU, 2.25% in terms of μDC, 4.35% in terms of mAP, and 2.58% in terms of AUC scores. These performance improvements are due to the fact that the proposed L_t loss function makes use of both contextual and semantic differences within the underwater scans in order to constrain the proposed model towards recognizing aquaculture defects regions irrespective of the scan characteristics. Therefore, in the rest of the experimentation, we used L_t loss function in order to train the proposed model towards extracting the aquaculture defects across all the four datasets.

6.2. Evaluation on LABUST dataset

In this section, we discuss the quantitative and qualitative evaluations of the proposed framework on the LABUST dataset.

First of all, we compared the performance of the proposed framework with state-of-the-art models that have been proposed for detecting the aquaculture defects such as biofouling, horizontal and vertical vegetation using the underwater imagery. To the best of our knowledge, all of these methods employ the pre-trained detection and segmentation models, such as, Fast R-CNN (Girshick et al., 2015), Faster R-CNN (Ren et al., 2015), Mask R-CNN (He et al., 2017), YOLOv4 (Bochkovskiy et al., 2020), YOLOv5 (Jocher et al., 2022), and SSD (Liu et al., 2016), where the performance of these models are measured using mAP, μIoU, μDC, and/or AUC scores. Therefore, to achieve a fair comparison with these methods across LABUST dataset, we also trained their underlying models using the same experimental protocols which we followed for training the proposed model. Moreover, the performance evaluation of the proposed framework is performed using all of the above-mentioned metrics (i.e., μIoU, μDC, and mAP). The performance comparison of the

Table 6
Determining the optimal loss function that gives the best aquaculture defects detection performance across all four datasets. Here, we used the proposed backbone encoder to generate the latent features when the model is being constrained using different loss functions. Moreover, the abbreviations are DS: Dataset, LB: LABUST Dataset, and (P) indicates the proposed L_t loss function. In addition, L_{sn} from Frosst et al. (2019), L_{ft} from Abraham and Khan (2019), and L_{de} from Raja et al. (2020) are taken.

DS	Loss Function	μIoU	μDC	mAP	AUC
LB	L_t (P)	**0.8265**	**0.9050**	**0.7858**	**0.9492**
	L_{ce}	0.7556	0.8608	0.7059	0.8963
	L_{sn}	0.8012	0.8896	0.7593	0.9246
	L_{ft}	0.7784	0.8754	0.7147	0.9081
	L_{de}	0.7926	0.8843	0.7376	0.9175
KU	L_t (P)	**0.9602**	**0.9795**	**0.8169**	**0.9668**
	L_{ce}	0.8834	0.9381	0.7239	0.8594
	L_{sn}	0.9318	0.9647	0.7753	0.9475
	L_{ft}	0.9162	0.9563	0.7448	0.8843
	L_{de}	0.9273	0.9623	0.7582	0.8951
NDv1	L_t (P)	**0.7730**	**0.8720**	**0.7144**	**0.9285**
	L_{ce}	0.7193	0.8367	0.6271	0.8517
	L_{sn}	0.7519	0.8584	0.6829	0.8928
	L_{ft}	0.7402	0.8507	0.6695	0.8746
	L_{de}	0.7257	0.8411	0.6438	0.8623
NDv2	L_t (P)	**0.8087**	**0.8942**	**0.7278**	**0.9314**
	L_{ce}	0.7274	0.8422	0.6254	0.8471
	L_{sn}	0.7726	0.8717	0.6843	0.9056
	L_{ft}	0.7641	0.8663	0.6675	0.8829
	L_{de}	0.7533	0.8593	0.6492	0.8603

proposed framework with these state-of-the-art methods are reported in Tables 7, and 8. We can observe from these tables that, under the fair experimental settings, the proposed framework outperforms the second-best method (based on Mask R-CNN Zhang et al., 2022) by 6.58% in terms of mAP score. In addition to this, in terms of μIoU, and μDC scores, the proposed framework outperformed the second-best method (Zhang et al., 2022) by 4.65% and 3.48%, respectively. Apart from Mask R-CNN (He et al., 2017), which is although a pioneer model for segmentation (especially instance segmentation) (Yu et al., 2022), we also considered some recently proposed segmentation models based on convolutional networks and transformers, such as DSRL (Liu et al., 2019), SegFormer (Xie et al., 2021), and SETR (Zheng et al., 2021), and compared their performances with the proposed model towards extracting the aquaculture defect regions. The purpose of comparing the proposed model with these models is to further verify its applicability towards robustly detecting the aquaculture defects as compared to the state-of-the-art deep learning models, and from Table 8, we can observe that the proposed model indeed outperform these models in terms of μIoU, and μDC scores. For example, it leads the DSRL model by 10.65% and 7.48% in terms of μIoU, and μDC scores, respectively. Similarly, it leads the SETR by 6.65% and 4.48%, in terms of μIoU, and μDC scores, respectively, and SegFormer by 5.65% and 3.48%, in terms of μIoU, and μDC scores, respectively. These performance improvements are evident from the fact that the proposed model, unlike its competitors, generates latent feature representations across various decomposition scales which are further expanded across multiple levels by the HDB block. Such design schematics allow the proposed model to pay attention to both coarser and finer details of the aquaculture defect regions for their accurate extraction. Although, the recently introduced SegFormer (Xie et al., 2021), Multi-scale attention feature extraction module (MAFEM) (Li, Yang, Zhao, & Yang, 2023), Mask R-CNN (Zhou et al., 2023), and SETR (Zheng et al., 2021) architectures employ multi-head self-attention mechanisms to segment the regions of interest. But since the latent features generated by these models are from the single depth of the network, their performance towards detecting the aquaculture defects is constrained. SegFormer (Xie et al., 2021), to some extent, overcome this limitation by expanding the latent features (generated from the single scale) to multiple levels by analyzing the scan content across various fields-of-view. But, still, its performance

Table 7
Performance comparison of the proposed framework with state-of-the-art methods in terms of mAP scores across different dataset. Moreover, the abbreviation 'LB' denotes the LABUST dataset, FRN means Faster R-CNN (Ren et al., 2015), MRN represents Mask R-CNN (He et al., 2017), and 'MEN' denotes the Proposed Multi-scale Encoder-Decoder Network. In addition, Fast R-CNN from Paraskevas and Ergina (2022), FRN from Rosales et al. (2021), SSD from Liao et al. (2022), YOLOv5 from Hu, Zhao et al. (2021), and YOLOv4 from Abinaya et al. (2022) are taken.

Metric	Method	LB	KU	NDv1	NDv2
mAP	MEN	**0.79**	**0.82**	**0.71**	**0.72**
	Fast R-CNN	0.64	0.69	0.56	0.59
	FRN	0.71	0.74	0.63	0.66
	MRN	0.73	0.78	0.65	0.68
	SSD	0.67	0.70	0.60	0.63
	YOLOv5	0.68	0.72	0.59	0.62
	YOLOv4	0.65	0.69	0.57	0.58

Table 8
Performance comparison of the proposed framework with state-of-the-art methods in terms of μIoU and μDC scores across different dataset. Moreover, the abbreviation LB denotes the LABUST dataset. Moreover, the abbreviation are: 'MEN': Proposed Multi-scale Encoder-Decoder Network, 'DSRL' Deep Super-Resolution Learning Network (Liu et al., 2019), 'SETR': Segmentation Transformer (Zheng et al., 2021), and Mask R-CNN (Zhang et al., 2022).

Metric	Method	LB	KU	NDv1	NDv2
μIoU	MEN	**0.82**	**0.96**	**0.76**	**0.79**
	Mask R-CNN	0.78	0.92	0.73	0.76
	DSRL	0.72	0.86	0.68	0.70
	SETR	0.77	0.90	0.72	0.75
	SegFormer	0.76	0.88	0.71	0.73
μDC	MEN	**0.90**	**0.97**	**0.86**	**0.88**
	Mask R-CNN	0.87	0.95	0.84	0.86
	DSRL	0.83	0.92	0.80	0.82
	SETR	0.87	0.94	0.83	0.85
	SegFormer	0.86	0.93	0.83	0.84

on the multi-textured underwater scans characterized by color cast, unbalanced contrast and blur, to detect aquaculture defect regions is limited.

Apart from this, the qualitative evaluation of the proposed framework for extracting the aquaculture defect regions across the LABUST dataset is shown in Fig. 5(A)–(L). We can observe here that how accurately the proposed framework is able to recognize the biofouling, horizontal and vertical regions of the vegetation irrespective of the diversity within the scan textures, and the background artifacts.[2] These results further encourage the deployment of the proposed model around the aquaculture farms to timely alert the farmers about the defective regions so that they can take appropriate actions.

6.3. Evaluation on KU dataset

The second dataset on which we evaluated the proposed framework is the KU dataset. The scans within the KU dataset contain vegetation and net hole defects. Moreover, the KU dataset scans contain a large amount of water disturbance, like, light reflections, water wave jerks, water ripples, etc. These disturbances can cause the segmentation and detection models to produce pixel-level or box-level false positives and false negatives, which can result in performance reduction.

First of all, we performed an exhaustive set of experiments through which we compared the robustness of the proposed framework with state-of-the-art methods towards detecting the aquaculture defects

[2] We also submitted a demo video, as the supplementary material of this article, that showcases the seamless recognition of aquaculture defects by the proposed framework on the LABUST dataset scans.

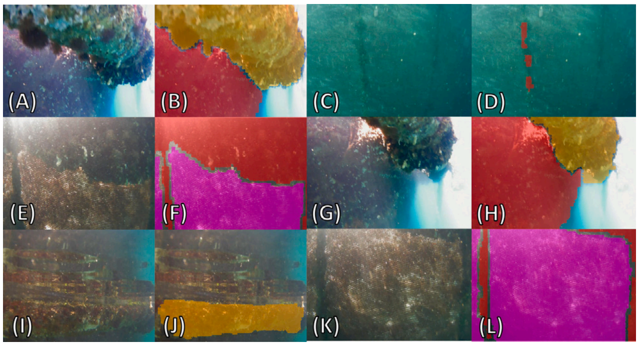


Fig. 5. Qualitative evaluations of the proposed model on the LABUST dataset. The first and third column shows the original scans, while the second and fourth column shows the extracted results. Moreover, orange color indicate vertical vegetation, red color indicates horizontal vegetation, and pink color represents biofouling region.

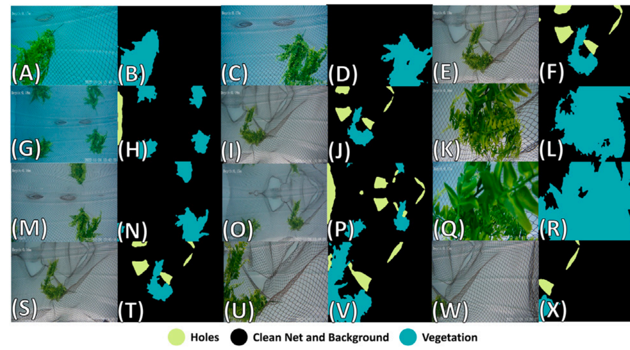


Fig. 6. Qualitative evaluations of the proposed model on the KU dataset. The first, third, and fifth column shows the original scans, while the second, fourth, and sixth column shows the extracted results. Moreover, cyan color indicates vegetation and light-green color represents holes region.

across the KU dataset. The comparison, in terms of mAP scores, is reported in Table 7. From Table 7, we can observe that the proposed framework outperforms the second-best method (Zhang et al., 2022) that is based on Mask R-CNN (He et al., 2017) by 3.69%. Apart from this, it outperformed the third-best method (Rosales et al., 2021) based on Faster R-CNN (Ren et al., 2015) by 7.69%. In addition to this, we also evaluated the performance of the proposed framework with Mask R-CNN (He et al., 2017), DSRL (Liu et al., 2019), and recently introduced transformers-based segmentation models, such as, SETR (Zheng et al., 2021), and SegFormer (Xie et al., 2021). The comparison is reported in Table 8 in which we can observe that the proposed framework leads the state-of-the-art method (Zhang et al., 2022) based on Mask R-CNN by 4.02% and 2.95% in terms of μIoU and μDC, respectively. Similarly, it is leading the DSRL, SETR and SegFormer models by 10.02%, 6.02% and 8.02% in terms of μIoU, and 5.95%, 3.95% and 4.95% in terms of μDC score, respectively. The performance improvements are a result of the fused features being naturally resilient to noisy artifacts, thanks to the incorporation of low-scale, coarser representations of the candidate scans. This also allows for finer representations of aquaculture defects to be detected, enabling their accurate extraction at the decoder end. Moreover, it is worth noting that the KU dataset contains input scans with extreme water disturbances like light reflections, water waves jerks, and water ripples, etc. But even in the presence of these disturbances, the proposed framework has able to accurately recognize the defected regions, with minimal pixel-level false positives, and false negatives.[3]

To further elaborate on the detection capability of the proposed framework, we showcase its qualitative evaluations across KU dataset in Fig. 6(A)–(X). Here, the first, third, and fifth columns show the input images, and the second, fourth, and sixth columns show the results of the proposed framework. Moreover, the light-green color shows the hole regions, the black color shows the clean and background portions, whereas the cyan color shows the vegetation region on the net surface. From the results reported in Fig. 6, we can observe how accurately the proposed framework can recognize the aquaculture defects irrespective of the water reflections and disturbances depicted within the underwater scans. This robust extraction of the defected regions is because of the inclusion of the custom IB, SPB, and HDB blocks within a proposed encoder topology that allows the generation of discriminative latent features for the background, cleaner net, and aquaculture defect categories, enabling the decoder end to accurately recognize them when the proposed model is trained using the proposed L_t loss function.

[3] The water disturbances on KU dataset can be observed in the demo video submitted as a supplementary material of this article.

6.4. Evaluation on NDv1 dataset

The third dataset on which we evaluated the proposed framework is the NDv1 dataset. NDv1 is a public dataset that contains complex underwater marine scans having biofouling and vegetation regions within aquaculture farms.

First of all, we report the quantitative evaluations of the proposed framework on NDv1 dataset through which we measured the segmentation performance of the proposed model, and compared it with state-of-the-art methods under fair experimental settings. Table 7 reports the performance of the proposed framework in terms of mAP scores. Here, we can observe that the proposed framework is leading the state-of-the-art (Zhang et al., 2022) via (He et al., 2017), by 6.44%. Similarly, in terms of μIoU and μDC scores, it is leading the state-of-the-art method (Zhang et al., 2022) (driven via Mask R-CNN He et al., 2017) by 3.30%, and 2.20%, respectively, as evident from Table 8. Additionally, we also compared the performance of the proposed framework with recently introduced segmentation models based on convolutional and transformer networks, such as, DSRL (Liu et al., 2019), SETR (Zheng et al., 2021), and SegFormer (Xie et al., 2021). Under fair comparison, the proposed framework is leading these models by 8.30%, 4.30%, and 5.30%, respectively, in terms of μIoU, and 6.20%, 3.20%, and 3.20%, respectively, in terms of μDC (see Table 8). These performance improvements are explained by the fact that the proposed model generates latent feature representations at different decomposition scales, which are then expanded across multiple levels by the HDB block. As a result, the model is able to focus on both the broader and finer details of the aquaculture defect regions, which results in their accurate extraction.

Apart from this, Fig. 7 shows the qualitative evaluation of the proposed framework on the NDv1 dataset for the biofouling and vegetation detection. The input scans are shown on the left, while the results of the proposed framework are shown on the right. Vegetation is represented by cyan, biofouling by light green, and the background region is displayed with the black color. Here, we can observe how accurately the proposed model has recognized the defects within the candidate scans at the inference stage. It is important to note that in some examples, the object size is blurry and small, yet thanks to the proposed L_t loss function, the proposed framework is optimally trained to these aquaculture defect regions accurately.

6.5. Evaluation on NDv2 dataset

The fourth dataset on which we evaluated the proposed framework is the publicly available NDv2 dataset. NDv2 is the second version

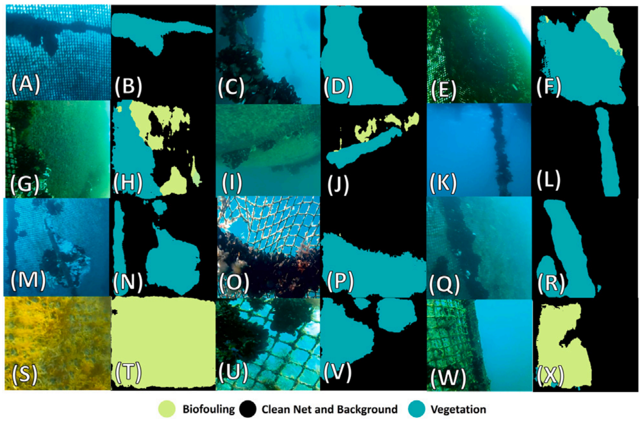


Fig. 7. Qualitative evaluations of the proposed model on the NDv1 dataset. The first, third, and fifth column shows the original scans, while the second, fourth, and sixth column shows the extracted results. Moreover, cyan color indicates vegetation and light-green color represents biofouling.

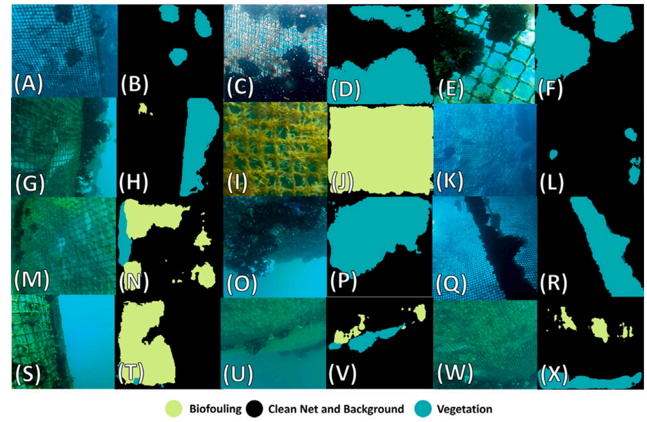


Fig. 8. Qualitative evaluations of the proposed model on the NDv2 dataset. The first, third, and fifth column shows the original scans, while the second, fourth, and sixth column shows the extracted results. Moreover, cyan color indicates vegetation and light-green color represents biofouling.

of NDv1 dataset that contains more complex scans. But similar to NDv1, it also allows the extraction of biofouling and vegetation defect categories.

The quantitative comparison of the proposed framework with state-of-the-art on NDv2 dataset, in terms of mAP, μIoU, and μDC scores, are reported in Tables 7, and 8. From Table 7, we can observe that the proposed framework, under the fair experimental settings, has outperformed the state-of-the-art works (Abinaya et al., 2022; Hu, Zhao et al., 2021; Liao et al., 2022; Paraskevas & Ergina, 2022; Rosales et al., 2021; Zhang et al., 2022) based on Fast R-CNN (Girshick et al., 2015), Faster R-CNN (Ren et al., 2015), Mask R-CNN (He et al., 2017), SSD (Liu et al., 2016), YOLOv4 (Bochkovskiy et al., 2020), and YOLOv5 (Jocher et al., 2022) by 13.78%, 6.78%, 4.78%, 9.78%, 10.78%, and 14.78% in terms of mAP scores, respectively. These performance improvements are due to the fact that the proposed encoder backbone utilizes unique SPB, IB, and HDB blocks arranged in a butterfly structure to extract distinct latent features from input scans, as compared to the state-of-the-art. Furthermore, constraining the proposed model through proposed L_t loss function led to the generation of distinct feature representations from the encoder end, which ultimately results in the robust extraction of aquaculture defect regions, such as biofouling, vegetation, and net holes, at the decoder end.

In another series of experiments, we compared the performance of the proposed framework with Mask R-CNN (He et al., 2017) (employed by Zhang et al. (2022)) and the recently introduced segmentation models based on convolutional and transformer networks, such as DSRL (Liu et al., 2019), SETR (Zheng et al., 2021), and SegFormer (Xie et al., 2021) in terms of μIoU, and μDC scores. The comparison is reported in Table 8 from which we can observe that the proposed framework, under fair experimental settings, outperformed DSRL (Liu et al., 2019), SETR (Zheng et al., 2021), and SegFormer (Xie et al., 2021) by 10.87%, 5.87%, and 7.87%, respectively, in terms of μIoU scores. Similarly, in terms of μDC scores, the proposed framework is leading DSRL (Liu et al., 2019), SETR (Zheng et al., 2021), and SegFormer (Xie et al., 2021) by 7.42%, 4.42%, and 5.40%, respectively, under fair experimental settings. The reason for these improvements is because of the discriminative latent features that are generated by the encoder end. These features are inherently resistant to the noisy artifacts because they are composed of coarser to finer representations of the aquaculture defect regions obtained across multiple depths of the encoder end from the input scans. Afterwards, these features allow the decoder-end to effectively recognize and localize the biofouling, and vegetation regions from the candidate scans irrespective of the scan attributes. On the other hand, generating a single-level or multi-level features from the single depth of the network, as seen in models like SegFormer (Xie et al., 2021) and SETR (Zheng et al., 2021), leads to the sub-optimal results when these models are applied to the complex underwater imagery.

Apart from this, Fig. 8(A)–(X) shows the qualitative evaluation of the proposed framework on the NDv2 dataset. The first, third, and fifth columns show the input scan, whereas the second, fourth, and sixth column shows the results of the proposed framework. Moreover, the cyan color represents the detect vegetation defects, the black color represents the background region, and the light-green color shows the biofouling regions. From Fig. 8, we can observe how well the proposed model recognized the aquaculture defect regions, even from the complex scans. For example in (G) and (U), the vegetation were largely grown on the aquaculture net but the proposed framework was able to effectively detect them as evident from (H) and (V), respectively. Similarly, from (S), and (W), we can observe how effectively the proposed framework has detected the biofouling regions from the complex underwater scans, as shown in (T), and (X). These improvements are again due to the inclusion of the IB, SPB, and HDB blocks within the encoder end that enabled the generation of distinct feature representations (when trained via L_t loss function). These features then aided in the accurate identification of the defected regions while maintaining their semantic and spatial appearances. Overall, this robust extraction of aquaculture defects demonstrates the applicability of the proposed framework in the real world to assist farmers by timely generating the alarm of any biofouling, and vegetation outgrowths on the aquaculture net.

6.6. Limitations

Although the proposed method has demonstrated remarkable performance for the biofouling, vegetation, and hole detection tasks in aquaculture environment. But, there is still some work that is needed to be done in order to improve the aquaculture defects detection in a dynamic marine environment. For instance, currently, the proposed framework only considers the biofouling occurrence on the net surface. However, biofouling can also grow among shellfish, finfish, seaweed, and other regions. We did not apply the proposed model to such cases due to the lack of datasets. So, dataset collection and formulation are the remedies that should be performed as one of the future tasks. Moreover, the tracking and monitoring the growth of biofouling regions over time will also aid the farmer in effectively performing the intervention and prevention tasks. Apart from this, the proposed framework is currently tested to only extract the vegetation regions from the net surface. However, the categorization of vegetation regions into different

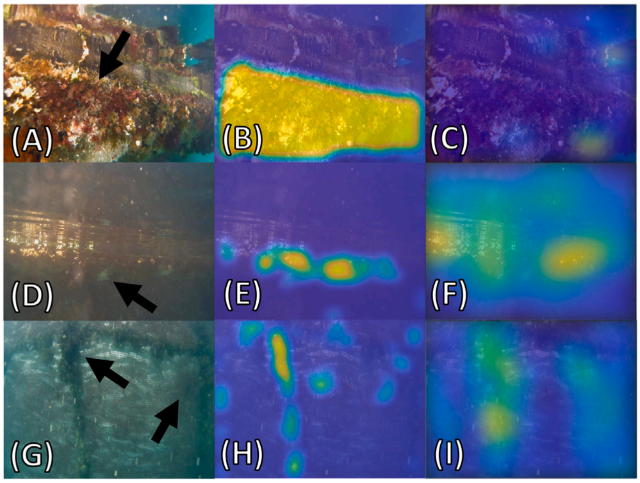


Fig. 9. Attention of the proposed network. The first column here shows the original images where the region of interests are hidden with black arrows. Moreover, the second column shows the attention maps of the proposed model, and the third column shows the attention map obtained by the Mask R-CNN (He et al., 2017) used in state-of-the-art works (Zhang et al., 2022).

categories, such as macroalgae, phytoplankton, and seagrass, was not performed in this work. These cases were not taken into consideration due to the absence of such classes in the datasets. Nevertheless, it is important to monitor and manage different types of vegetation growth on the net pens over time to ensure the health and safety of the fish and the surrounding environment. Moreover, the net holes considered in this study are captured parallel to the camera from a constant distance, and the size of the hole does not vary too much within each image frame due to the perspective projection. It is also important to consider and approximate the physical size of the damaged area on the net in order to take further maintenance actions. We plan to conduct these experiments in the future by collaborating with the farmers after deploying the proposed framework in their aquaculture setups.

6.7. Discussion

This paper presents a novel multi-scale semantic segmentation model for the aquaculture defect detection applications. The proposed framework has undergone rigorous testing across a variety of datasets, namely, LABUST, KU, NDv1, and NDv2, where it was effectively trained using the proposed L_t loss function. Moreover, the experimental results across these datasets showcased the resilience and efficiency of the proposed framework for the aquaculture defect detection tasks, particularly for the biofouling, vegetation, and net holes detection. The qualitative and quantitative evaluations across four datasets demonstrate a significant performance gain of the proposed framework when compared to the state-of-the-art methods as evident from Tables 7–8, and Figs. 5–9. The performance improvements, produced by the proposed framework, are because of its unique design which fuses the multi-scale features, obtained from various decomposition levels, to accurately extract the aquaculture defect regions irrespective of the scan attributes. Apart from this, each level of the encoder block is built upon custom IB, and SPB blocks that allow to model to learn the high-resolution contextual and semantic information about the aquaculture defect regions, whereas the model attention is further focused through the inclusion of the HDB block. Afterwards, the latent features, generated by the encoder end, are utilized by the decoder end towards accurately extracting the aquaculture defects from the underwater imagery.

The robustness and significance of the model is further demonstrated by the attention maps produced by the proposed multi-scale encoder as shown in Fig. 9. Here, we also compared the attention maps generated by the proposed model against the Mask R-CNN (He et al., 2017) which is used in state-of-the-art works (Zhang et al., 2022). We can observe from the attention maps that the proposed model only focuses on the biofouling and vegetation regions while ignoring the irrelevant regions of the scan, as compared to its competitor. For example, see the accurate extraction of biofouling, horizontal, and vertical vegetation in Fig. 9(B, E, H), as compared to (C, F, I). Furthermore, since Mask R-CNN (He et al., 2017) is primarily designed to work on the conventional photographic imagery, it produces sub-optimal attention maps when exposed to the underwater imagery (as evident from F and I), which ultimately leads toward the inaccurate detection of biofouling and vegetation regions. On the contrary, the proposed model, thanks to its shape preservation and hierarchical decomposition blocks fuses the attention of the network across various depths and refines them in order to generate accurate localization of the net defects areas.

The results indicate that the proposed framework is the most suitable approach for detecting defects in aquaculture nets across various datasets obtained from video sequences during net inspection campaigns. However, different types and levels of biofouling, vegetation, and holes over time are not considered in this study. As a part of our future research, we aim to enhance the management of net defects by accounting for scenarios involving different degrees of perspective distortion and by integrating the characterization of biofouling, vegetation, and holes over time. Additionally, the proposed approach can be employed for inspecting open-sea fish net pens, which can aid in the timely identification of net defects, repairing damaged regions, and ensuring the overall stability of the net.

7. Conclusion

In this paper, we presented a novel multi-scale semantic segmentation network for the robust extraction of aquaculture defects, such as biofouling, vegetation, and net holes, from the complex underwater imagery. The key benefits of the proposed approach lie in its ability to extract the fused multi-scale latent features that allows it to focus its attention only on the aquaculture defect regions in order to robustly recognize them irrespective of the scan attributes. The autonomous detection of aquaculture defects also aids the workers in accurately managing their schedule and looking after the aquaculture net more effectively. The proposed framework has been rigorously tested on four different datasets, dubbed, LABUST, KU, NDv1, and NDv2, where across each dataset, it achieved significant performance improvements over state-of-the-art. For example, on the LABUST dataset, it achieves the performance boost of 6.58%, 4.65%, and 3.48% in terms of mAP, μIoU, and μDC scores, respectively. Similarly, on the KU dataset, it achieves the performance boost of 3.69%, 4.02%, and 2.95% in terms of mAP, μIoU, and μDC scores, respectively. The results obtained on the NDv1 dataset showed performance boost of 6.44%, 3.30%, and 2.20% in terms of mAP, μIoU, and μDC scores, respectively. Moreover, on the NDv2 dataset, the proposed model also showed performance boost of 4.78%, 7.87%, and 4.42% in terms of mAP, μIoU, and μDC scores, respectively.

In the future, we aim to address all the limitations as mentioned in Section 6.6, especially the ones related to formulating a new dataset to identify different categories of biofouling, vegetation, and holes within aquaculture net scans. Additionally, other challenges such as fish detection will also be investigated in the future to determine the extent of the proposed framework.

CRediT authorship contribution statement

Waseem Akram: Conceptualization, Methodology, Software, Writing – original draft, Formal analysis, Writing – review & editing. **Taimur Hassan:** Conceptualization, Methodology, Software, Writing – original draft, Formal analysis, Writing – review & editing. **Hamed**

Toubar: Dataset collection, Software, Contributions in writing. **Muhayyuddin Ahmed:** Methodology, Writing – original draft, Investigation. **Nikola Miškovic:** Methodology, Writing – original draft, Investigation. **Lakmal Seneviratne:** Methodology, Writing – original draft, Investigation. **Irfan Hussain:** Methodology, Writing – original draft, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This publication is based upon work supported by the Khalifa University of Science and Technology under Award No. CIRA-2021-085, FSU-2021-019, RC1-2018-KUCARS.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2023.121197.

References

Abinaya, N., Susan, D., & Sidharthan, R. K. (2022). Deep learning-based segmental analysis of fish for biomass estimation in an occulted environment. *Computers and Electronics in Agriculture*, *197*, Article 106985.

Abraham, N., & Khan, N. M. (2019). A Novel Focal Tversky loss function with improved Attention U-Net for lesion segmentation. In *IEEE 16th international symposium on biomedical imaging*. ISBI.

Ahmed, M. S., Aurpa, T. T., & Azad, M. A. K. (2021). Fish disease detection using image based machine learning technique in aquaculture. *Journal of King Saud University-Computer and Information Sciences*.

Akram, W., Casavola, A., Kapetanović, N., & Miškovic, N. (2022). A visual servoing scheme for autonomous aquaculture net pens inspection using ROV. *Sensors*, *22*(9), 3525.

Amundsen, H. B., Caharija, W., & Pettersen, K. Y. (2021). Autonomous ROV inspections of aquaculture net pens using DVL. *IEEE Journal of Oceanic Engineering*, *47*(1), 1–19.

Bjerkeng, M., Kirkhus, T., Caharija, W., T. Thielemann, J., B. Amundsen, H., Johan Ohrem, S., et al. (2021). ROV navigation in a fish cage with laser-camera triangulation. *Journal of Marine Science and Engineering*, *9*(1), 79.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv, 2020.

Chabot, D., Dillon, C., Shemrock, A., Weissflog, N., & Sager, E. P. (2018). An object-based image analysis workflow for monitoring shallow-water aquatic vegetation in multispectral drone imagery. *ISPRS International Journal of Geo-Information*, *7*(8), 294.

Chang, C.-C., Wang, J.-H., Wu, J.-L., Hsieh, Y.-Z., Wu, T.-D., Cheng, S.-C., et al. (2021). Applying artificial intelligence (AI) techniques to implement a practical smart cage aquaculture management system. *Journal of Medical and Biological Engineering*, *41*(5), 652–658.

Dalhatu, A. A., Sa'ad, A. M., Cabral de Azevedo, R., & De Tomi, G. (2023). Remotely operated vehicle taxonomy and emerging methods of inspection, maintenance, and repair operations: An overview and outlook. *Journal of Offshore Mechanics and Arctic Engineering*, *145*(2), Article 020801.

Duda, A., Schwendner, J., Stahl, A., & Rundtop, P. (2015). Visual pose estimation for autonomous inspection of fish pens. In *OCEANS 2015-Genova* (pp. 1–6). IEEE.

Duecker, D. A., Hansen, T., & Kreuzer, E. (2020). RGB-D camera-based navigation for autonomous underwater inspection using low-cost micro AUVs. In *2020 IEEE/OES autonomous underwater vehicles symposium* AUV, (pp. 1–7). IEEE.

Er, M. J., Chen, J., Zhang, Y., & Gao, W. (2023). Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: A review. *Sensors*, *23*(4), 1990.

Fitridge, I., Dempster, T., Guenther, J., & De Nys, R. (2012). The impact and control of biofouling in marine aquaculture: a review. *Biofouling*, *28*(7), 649–669.

Frosst, N., Papernot, N., & Hinton, G. (2019). Analyzing and improving representations with the soft nearest neighbor loss. In *International conference on machine learning* (pp. 2012–2020). PMLR.

Girshick, R., et al. (2015). Fast R-CNN. In *ICCV, 2015*.

Haq, K. R. A., & Harigovindan, V. (2022). Water quality prediction for smart aquaculture using hybrid deep learning models. *IEEE Access*.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. In *IEEE international conference on computer vision (ICCV), 2017*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, X., Liu, Y., Zhao, Z., Liu, J., Yang, X., Sun, C., et al. (2021). Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network. *Computers and Electronics in Agriculture*, *185*, Article 106135.

Hu, J., Zhao, D., Zhang, Y., Zhou, C., & Chen, W. (2021). Real-time nondestructive fish behavior detecting in mixed polyculture system using deep-learning and low-cost devices. *Expert Systems with Applications*, *178*, Article 115051.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Jocher, G., et al. (2022). *YOLOv5*. Ultralytics, Retrieved: November 2nd, 2022.

Kapetanović, N., Goričanec, J., Vatavuk, I., Hrabar, I., Stuhne, D., Vasiljević, G., et al. (2022). Heterogeneous autonomous robotic system in viticulture and mariculture: Vehicles development and systems integration. *Sensors*, *22*(8), 2961.

Kapetanović, N., Nađ, Đ., & Mišković, N. (2021). Towards a heterogeneous robotic system for autonomous inspection in mariculture. In *OCEANS 2021: San Diego–Porto* (pp. 1–6). IEEE.

Labra, J., Zuniga, M. D., Rebolledo, J., Ahmed, M. A., Carvajal, R., Jara, N., et al. (2023). Robust automatic net damage detection and tracking on real aquaculture environment using computer vision. *Aquacultural Engineering*, *101*, Article 102323.

Lee, H., Jeong, D., Yu, H., & Ryu, J. (2022). Autonomous underwater vehicle control for fishnet inspection in turbid water environments. *International Journal of Control, Automation and Systems*, *20*(10), 3383–3392.

Lee, D., Kim, G., Kim, D., Myung, H., & Choi, H.-T. (2012). Vision-based object detection and tracking for autonomous navigation of underwater robots. *Ocean Engineering*, *48*, 59–68.

Li, D., & Du, L. (2022). Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish. *Artificial Intelligence Review*, *55*(5), 4077–4116.

Li, Y., Jiang, Y., Cao, J., Wang, B., & Li, Y. (2015). AUV docking experiments based on vision positioning using two cameras. *Ocean Engineering*, *110*, 163–173.

Li, D., Yang, Y., Zhao, S., & Yang, H. (2023). A fish image segmentation methodology in aquaculture environment based on multi-feature fusion model. *Marine Environmental Research*, Article 106085.

Liao, W., Zhang, S., Wu, Y., An, D., & Wei, Y. (2022). Research on intelligent damage detection of far-sea cage based on machine vision and deep learning. *Aquacultural Engineering*, *96*, Article 102219.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). SSD: Single Shot MultiBox Detector. In *ECCV, 2016*.

Liu, T., De Haan, K., Rivenson, Y., Wei, Z., Zeng, X., Zhang, Y., et al. (2019). Deep learning-based super-resolution in coherent imaging systems. *Scientific Reports*, *9*(1), 1–13.

Livanos, G., Zervakis, M., Chalkiadakis, V., Moirogiorgou, K., Giakos, G., & Papandroulakis, N. (2018). Intelligent navigation and control of a prototype autonomous underwater vehicle for automated inspection of aquaculture net pen cages. In *2018 IEEE international conference on imaging systems and techniques* IST, (pp. 1–6). IEEE.

Matsubara, T., & Torikai, H. (2015). An asynchronous recurrent network of cellular automaton-based neurons and its reproduction of spiking neural network activities. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(4), 836–852.

Nishitani, A., Kawada, R., & Kojima, J. (2022). Development of combined aerial and underwater drone system. In *OCEANS 2022, Hampton Roads* (pp. 1–5). IEEE.

Ohrem, S. J., Kelasidi, E., & Bloecher, N. (2020). Analysis of a novel autonomous underwater robot for biofouling prevention and inspection in fish farms. In *2020 28th mediterranean conference on control and automation* MED, (pp. 1002–1008). IEEE.

Osen, O. L., Havnegjerde, A., Kamsvåg, V., Liavaag, S., & Bye, R. T. (2016). A low cost USV for aqua farm inspection. In *2016 Techno-Ocean (Techno-Ocean)* (pp. 291–298). IEEE.

Paraskevas, K., & Ergina, K. (2022). Biofouling detection in fish farming nets using neural networks. Available at SSRN 4144161.

Perrin, J. E., Jernigan, S. R., Thayer, J. D., Howell, A. W., Leary, J. K., & Buckner, G. D. (2022). Sensor Fusion with deep learning for autonomous classification and management of aquatic invasive plant species. *Robotics*, *11*(4), 68.

Raja, H., Hassan, T., Akram, M. U., & Werghi, N. (2020). Clinically verified hybrid deep learning system for retinal ganglion cells aware grading of glaucomatous progression. *IEEE Transactions on Biomedical Engineering*, *68*(7), 2140–2151.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv, 2018.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*.

Roboflow (2022a). Net defect detection dataset-v1. Retrieved: December 14th, 2022.

Roboflow (2022b). Net defect detection dataset-v2. Retrieved: November 20th, 2022.

Rosales, M. A., Palconit, M. G. B., Almero, V. J. D., Concepcion, R. S., Magsumbol, J.-A. V., Sybingco, E., et al. (2021). Faster R-CNN based fish detector for smart aquaculture system. In *2021 IEEE 13th international conference on humanoid, nanotechnology, information technology, communication and control, environment, and management* HNICEM, (pp. 1–6). IEEE.

Rundtop, P., & Frank, K. (2016). Experimental evaluation of hydroacoustic instruments for ROV navigation along aquaculture net pens. *Aquacultural Engineering*, *74*, 143–156.

Sato, E., Liu, H., Orita, Y., Sakagami, N., & Wada, T. (2022). Cooperative path-following control of a remotely operated underwater vehicle for human visual inspection task. *Frontiers in Control Engineering*, 31.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Tao, Q., Huang, K., Qin, C., Guo, B., Lam, R., & Zhang, F. (2018). Omnidirectional surface vehicle for fish cage inspection. In *OCEANS 2018 MTS/IEEE Charleston* (pp. 1–6). IEEE.

Tran-Quang, V., & Ha-Ngoc, A. (2021). Aquaculture environment prediction based on improved LSTM deep learning model. In *2021 8th NAFOSTED conference on information and computer science* NICS, (pp. 486–491). IEEE.

Voskakis, D., Makris, A., & Papandroulakis, N. (2021). Deep learning based fish length estimation. An application for the Mediterranean aquaculture. In *OCEANS 2021: San Diego–Porto* (pp. 1–5). IEEE.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., et al. (2021). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Wang, H., Zhang, S., Zhao, S., Lu, J., Wang, Y., Li, D., et al. (2022). Fast detection of cannibalism behavior of juvenile fish based on deep learning. *Computers and Electronics in Agriculture*, *198*, Article 107033.

Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., & Chen, L.-C. (2020). Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV* (pp. 108–126). Springer.

Wu, Y., Liu, J., Wei, Y., An, D., Duan, Y., Li, W., et al. (2022). Intelligent control method of underwater inspection robot in netcage. *Aquaculture Research*, *53*(5), 1928–1938.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, *34*, 12077–12090.

Yang, X., Zhang, S., Liu, J., Gao, Q., Dong, S., & Zhou, C. (2021). Deep learning for smart fish farming: applications, opportunities and challenges. *Reviews in Aquaculture*, *13*(1), 66–90.

Yu, C., Hu, Z., Li, R., Xia, X., Zhao, Y., Fan, X., et al. (2022). Segmentation and density statistics of mariculture cages from remote sensing images using mask R-CNN. *Information Processing in Agriculture*, *9*(3), 417–430.

Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., et al. (2021). Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10440–10450).

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. arXiv:1212.5701.

Zhang, Z., Gui, F., Qu, X., & Feng, D. (2022). Netting damage detection for marine aquaculture facilities based on improved mask R-CNN. *Journal of Marine Science and Engineering*, *10*(7), 996.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. CVPR.

Zhou, H., Kim, S. H., Kim, S. C., Kim, C. W., Kang, S. W., & Kim, H. (2023). Instance segmentation of shrimp based on contrastive learning. *Applied Sciences*, *13*(12), 6979.

Zion, B. (2012). The use of computer vision technologies in aquaculture–a review. *Computers and Electronics in Agriculture*, *88*, 125–132.