Segmentation of Fish in Realistic Underwater Scenes using Lightweight Deep Learning Models

Gordon Böer[1], Rajesh Veeramalli[1] and Hauke Schramm[1,2]

[1] *Institute of Applied Computer Science, Kiel University of Applied Sciences, Kiel, Germany*

[2] *Department of Computer Science, Faculty of Engineering, Kiel University, Germany*

Keywords: Semantic Segmentation, Underwater Imagery, Fish Detection.

Abstract: The semantic segmentation of fish in real underwater scenes is a challenging task and an important prerequisite for various processing steps. With a good segmentation result, it becomes possible to automatically extract the fish contour and derive morphological features, both of which can be used for species identification and fish biomass assessment. In this work, two deep learning models, DeepLabV3 and PSPNet, are investigated for their applicability to fish segmentation for a fish stock monitoring application with low light cameras. By pruning these networks and employing a different encoder, they become more suitable for systems with limited hardware, such as remotely operated or autonomously operated underwater vehicles. Both segmentation models are trained and evaluated on a novel dataset of underwater images showing *Gadus morhua* in its natural behavior. On a challenging test set, which includes fish recorded at difficult visibility conditions, the PSPNet performs best, and achieves an average pixel accuracy of 96.8% and an intersection-over-union between the predicted and the target mask of 73.8%. It achieves this with a very limited parameter set of 94,393 trainable parameters.

1 INTRODUCTION

Digital imaging in marine research has become a standard tool to help marine biologists answer many scientific questions. This is largely due to rapid technological advances in recent decades that have resulted in digital cameras with higher technological capabilities, such as better image quality, larger storage capacities, and better in-situ applicability, while at the same time being available at lower prices in the consumer market. The applications of modern sensors and algorithms for underwater imaging are numerous and several reviews have been published, emphasizing either their general applicability for marine science (Durden et al., 2016; Malde et al., 2020; Fernandes et al., 2020) or for more specific research areas, like the observation of coastal marine biodiversity (Mallet and Pelletier, 2014), the monitoring of human impact on marine environments (Bicknell et al., 2016), the automatic determination of fish species (Alsmadi and Almarashdeh, 2020) or the investigation of fish connectivity (Lopez-Marcano et al., 2021).

The aim of this work is to investigate the applicability of deep learning methods to segment and outline fishes, detected in real-world underwater scenes. In addition to the information about what kind of objects are present in an image and where they are located, a successful semantic segmentation reveals what class each pixel belongs to. It thereby becomes possible, to additionally extract the outline of an object of interest and the concise area it covers in the image. The precise segmentation of a fish is an important prerequisite for the automatic determination of morphometric characteristics, like the total length, which in turn can be used to determine the fish weight (Wilhelms et al., 2013). Additionally, well-defined landmarks on the fish outline are commonly used to identify specific fish species (Rawat et al., 2017; Cavalcanti et al., 1999), while automatically localizing those landmarks becomes easier with a good segmentation result. The presented algorithm is an important software part for a currently developed underwater sensor platform, that aims to estimate fish biomass for a fish stock assessment, using non-invasive sensor technology.

Recently, there has also been a great demand to apply the undoubtedly successful deep learning algorithms using limited hardware. This is especially true for remotely controlled and autonomous vehicles, such as underwater robots, for which we intend to use

Böer, G., Veeramalli, R. and Schramm, H.
Segmentation of Fish in Realistic Underwater Scenes using Lightweight Deep Learning Models.
DOI: 10.5220/0010712700003061
In *Proceedings of the 2nd International Conference on Robotics, Computer Vision and Intelligent Systems (ROBOVIS 2021)*, pages 158-164
ISBN: 978-989-758-537-1

the studied algorithms in the future. Therefore, more light-weight segmentation models are investigated in this publication. Precisely speaking, pruned versions of DeepLabv3 (Chen et al., 2017) and PSPNet (Zhao et al., 2017) are used on image extracts containing fish to perform a binary segmentation into a fish and background class.

All the used underwater videos were recorded with specific low-light cameras adapted for an underwater usage, that allowed the recording of underwater scenes without active lighting.

The main contributions of this work are:

- Alternative configurations of PSPNet and DeepLabv3 which are better suited for hardware-limited devices.
- The comparison of those segmentation models regarding their inference times and performance for fish segmentation.

The rest of this article is organized as follows. Section 2 provides an overview of related publications. In section 3, the used recording setup and the employed algorithms, together with the metrics used to evaluate them, are described. Section 4 details the experimental evaluation, and section 5 concludes this paper.

2 RELATED WORK

Besides the automatic localization and classification of arbitrary objects in digital images, the semantic segmentation of those images is another active field of research. It is of great relevance in many application areas like autonomous driving (Grigorescu et al., 2020), remote sensing (Marmanis et al., 2016), 3D-sensing (Tchapmi et al., 2017) or cancer prediction (Kourou et al., 2015). As in many areas of computer vision, the most powerful algorithms to date are based on deep learning architectures, well established ones being Fully Convolutional Networks (Long et al., 2015), Mask R-CNN (He et al., 2017), U-Net (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017) or SegNet (Badrinarayanan et al., 2017).

Regarding the segmentation of fish in digital images, several efforts have been published so far. We will focus on methods that have been tested on free-swimming fish, recorded in underwater scenarios. Due to the different settings, the results obtained for dead fish photographed in air, e.g. on a photo-table (Yu et al., 2020; Konovalov et al., 2019; Baloch et al., 2017), conveyor belt (Storbeck and Daan, 2001) or fish-trawler (Yang et al., 2018; French et al., 2015), are not directly comparable.

So far, various standard algorithms have been applied to extract segmentation masks for fish, like Otsu thresholding, edge detection, Grabcut, mean-shift, matrix decomposition or the curvature scale space transform (Abdeldaim et al., 2018; Qin et al., 2014; Spampinato et al., 2010; Abbasi and Mokhtarian, 1999). However, all those algorithms do not profit from recent advances possible with novel deep learning methods, e.g. the possibility to automatically learn important image features, as opposed to hand-crafted ones.

To our knowledge, published efforts to use deep learning methods for semantic segmentation of free-swimming marine animals are still rather limited. Recently, the Enhanced Cascade Decoder Network (ECD-Net) has been proposed (Li et al., 2021), which is utilized for the segmentation of marine animals, including several fish species. It builds upon a pre-trained ResNet-50 (He et al., 2016) backbone, followed by several feature enhancement and cascade decoder modules which are trained using a mixture loss. The authors report superior results, on a self-published dataset, as compared to several state-of-the-art models. Although the ECD-Net, in terms of trainable parameters, has been reported to be smaller than most of the other considered networks, with the remaining 207 million parameters it is much too complex for the usage in an embedded system. In another work, images from the stereoscopic camera system Deep Vision (Rosen and Holst, 2013), which records the water volume at the opening of a net trawl, were automatically processed to determine the length of visible fish. For this purpose, a Mask R-CNN was used to detect and segment fish, followed by a refinement step to distinguish between individual overlapping fish. The authors report an average IoU of 84.5% for an independent test set of 200 images (Garcia et al., 2020). Given the used camera setup, the fish is recorded showing an unnatural behavior in front of a rather uniform background, therefore, the obtained results may not be directly transferable to real world underwater scenarios.

3 MATERIALS AND METHODS

3.1 Data Acquisition

All experiments were carried out on videos of realistic underwater scenes, which are part of a larger collection of underwater sensor data, recorded by a prototype underwater sensor platform. The device was developed in a joint project of German universities and marine engineering companies. Over the period of several months, the sensor platform was deployed at seafloor level, at a depth of 22m in the North Sea, about 45 nautical miles west of the island of Sylt, and in the Kiel Fjord, an inlet of the Baltic Sea in northern Germany. During the measurement campaigns, the primary goal was to conduct a continuous recording of stereo video and sonar data as well as various oceanic parameters at a high temporal sampling rate. In total, the raw data set spans approximately 240 days, resulting in nearly 3000 hours of video footage, as the cameras were not operated during the night hours.

All stereo videos were recorded using two Photonis ”Nocturn XL” monochrome CMOS image sensors with an optical resolution of 1280x1024 pixels and a sampling rate of up to 20 frames/second, housed in a specially designed underwater case with a flat view port. The selected camera system is particularly suitable for low-light scenarios, which made it possible to record without active lighting from dawn to dusk, which was verified to a maximum depth of 22m. The exclusive use of passive lighting is justified by the main concept of the platform, which is to be as less invasive as possible, e.g. by avoiding attraction effects from light sources (Marchesan et al., 2005), thus ensuring that the fish can be recorded in their natural behavior.

The same stereo camera setup will be duplicated on a mobile platform that is currently being built, making it possible to reuse many of the insights gained from the stationary system as has been used in the present work.

3.2 Methodology

In general, the semantic segmentation of an image can be defined as a pixel-wise classification, where each pixel is assigned to a specific object class. Depending on the complexity, the task can further be divided into:

- Binary segmentation, if only 2 classes are separated, e.g. foreground and background.
- Multi-class segmentation, if multiple classes are considered, e.g. car, pedestrian, street, building.
- Instance segmentation, if each pixel additionally is assigned to a unique instance of an object class, thereby making it possible to distinguish between several, possibly overlapping, occurrences of the same object class.

The problem addressed in this work is a binary segmentation problem since only the two classes *fish* and *background* are considered.

A simple approach to define a segmentation network is to stack multiple convolutional layers with the same padding to preserve the resolution of the input image. This type of architecture is computationally intensive, as it preserves the input resolution through all layers of the network. Therefore, most recent segmentation networks follow an encoder-decoder architecture, which is comparatively more efficient. First, a sequence of convolutional and either downsampling or pooling layers creates a low-resolution image representation which encodes the high-resolution information of the input image. Second, to reconstruct the original resolution in the output image, again a sequence of convolutional layers, followed by upsampling layers or transposed convolutions is added, to gradually increase the size of the spatial features.

In the present work, the two segmentation models DeepLabv3 and PSPNet are used, both of which follow the described type of encoder-decoder architecture. As opposed to the original architectures, which utilize a ResNet (He et al., 2016) as decoder, we employ MobileNetV2 (Sandler et al., 2018) for each of the two segmentation networks. MobileNetV2 makes use of depthwise convolutions and inverted residuals, which are grouped together into several subsequent blocks using ReLU6 as non-linearity function. The features from the first 3 stages of the MobileNetV2, which has been pre-trained on ImageNet, are passed to the decoder part of the respective network. These changes reduce the model complexity and size, which makes them better suited for systems with limited hardware.

3.2.1 PSPNet

In the original implementation of PSPNet, a pretrained ResNet was employed in combination with a dilated network strategy to extract the feature maps. As was mentioned above, the ResNet was interchanged with MobileNetV2 in the proposed system. The spatial pyramid pooling (SPP) module in PSPNet first downsamples the feature maps from the encoder at 4 different scales (1 x 1, 2 x 2, 3 x 3 and 6 x 6), all of which are afterwards upsampled and fused together. The integrated feature maps from the SPP module are concatenated along with the feature maps from the encoder. The decoder, followed by a bilinear upsampling layer with a scale of 8, then converts the concatenated feature maps to the segmentation output. The SPP module eliminates the requirement for a fixed input size.

3.2.2 DeepLabv3

In the DeepLabv3 architecture, a series of 3 x 3 atrous convolutions are built in cascade, which are able to capture long-range information from the inputs.

Since in the original architecture, the last block of the ResNet encoder is duplicated 3 times, we are duplicating the last blocks of the MobileNetV2 encoder as well. The atrous spatial pyramid pooling module in DeepLabv3 consists of multiple parallel atrous convolutional layers with different dilation rates. This module resamples the feature maps at multiple rates, which have a complementary field of view compared to normal convolution filters. Following this idea, it becomes possible to classify objects or regions of arbitrary size.

3.3 Evaluation Metrics

The metrics used for the validation of the models are the pixel-level accuracy and Intersection-over-Union (IoU):

- The overall per-pixel accuracy measures the ratio of correctly classified pixels to total pixels.
- The IoU is measured as the area of overlap, divided by the area of union, between the predicted segmentation and the ground truth segmentation. This metric is often referred to as Jaccard Index.

The investigated task of fish segmentation can be defined as an unbalanced, binary segmentation problem, since the background class dominates the fish class in the dataset. In this case, the pixel-level accuracy is sensitive to the class-imbalance problem, which is why we rely on the IoU as an additional metric.

4 EXPERIMENTAL RESULTS

In this section, we provide a brief description of the used data, the implementation and training details, and the results achieved, as compared for the two networks.

4.1 Data

The used dataset consists of 600 labeled grayscale images with a resolution of 1280x1024 pixels, that depict fish of the species *Gadus morhua*, which is of significant importance to the fishery industry. The full image set was randomly split into 80% for training, 10% for testing and 10% for validating the models. Since the images in these data splits have been extracted from video sequences, very similar samples may appear in the test and training set due to the simple random selection, e.g. if a fish is swimming slowly in front of the camera. Because of this, a separate test set of 1148 images, that where recorded on a different day have been annotated as well. Using this test set, we aim to fairly assess the generalization ability of the segmentation models used. All images have been annotated with binary segmentation masks, where pixels, belonging to a fish, are marked as foreground, everything else as background, respectively. In most of the samples, background pixels predominate the number of foreground pixels. Each thereby annotated fish is included in a bounding-box of a fixed size of 512x512 pixels. In several cases, a bounding box may contain more than one fish.

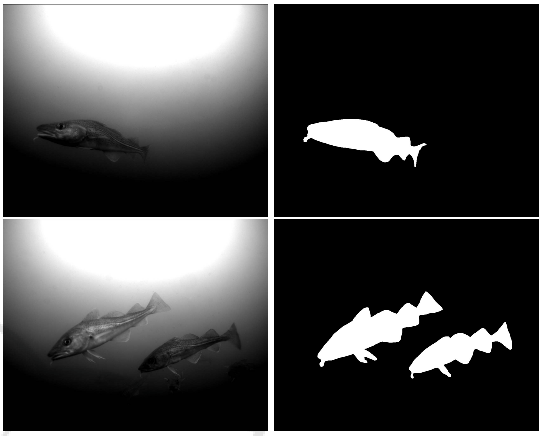

Figure 1: Examples of an annotated segmentation mask.

4.2 Training Setup

The models have been implemented with PyTorch and were trained and evaluated on a NVIDIA TITAN XP GeForce RTX 2080 TI GPU. For the model optimization, the Dice loss was used to determine the error between the prediction and the ground truth, which is calculated at pixel-level. The Dice loss for a complete image is defined by the formula:

$$DL = \frac{2\sum_i^N y_i \hat{p}_i}{\sum_i^N \hat{p}_i^2 + \sum_i^N y_i^2} \tag{1}$$

with y_i and $\hat{p}_i$ being the values of corresponding pixels representing the true object class and the predicted class for this pixel, respectively. The network parameters were optimized using the Adam backpropagation algorithm, with a learning rate of 0.001. Both models were trained with a batch size of 4, each sample being flipped with a probability of 0.5 in the horizontal or vertical direction. Other augmentation techniques were not used, since the possible data interpolation may produce an unnatural fish appearance. Both models were trained for 30 epochs on the same data splits. All segmentation results were thresholded at a confidence value of 0.5 to obtain binary masks.

4.3 Results and Discussion

The performance of each model on the independent test set of 1148 images, using the aforementioned evaluation metrics, is listed in Table 1. Regarding the pixel accuracy, both models perform almost equally well with an average accuracy of 96.8% for PSPNet and 96.2% for DeepLabv3, respectively. Considering the IoU, PSPNet performs better with an average value of 73.8% compared to the 69.9% as achieved by DeepLabv3.

Table 1: Average pixel level accuracy and IoU as obtained by the two utilized models.

Model	Pixel Accuracy	IoU
PSPNet	96.8%	73.8%
DeepLabv3	96.2%	69.9%

To assess which model is better suited for constrained hardware devices, we investigate their respective inference times and model sizes, as listed in Table 2. While both models perform equally well in terms of segmentation accuracy, it is obvious, that the PSPNet achieves this by using much less memory at a slightly higher framerate. The difference in model size can be explained by the number of trainable parameters, which adds up to 94,393 for the PSPNet and 121,382,5 for DeepLabv3, respectively. Deeplabv3 has much more trainable parameters because of the used trainable atrous convolutions in the decoder. Given this insight, we suggest using the PSPNet for a hardware limited device.

Table 2: Inference speed, in frames per second (FPS), and model size of the two adapted models.

Model	Inference speed (FPS)		Model size
	CPU	GPU	
PSPNet	5.5	155	463 KB
DeepLabv3	3.45	148	5 MB

We have investigated those examples, where both models fail to generate a perfect segmentation mask, which revealed, that both models made similar errors on the same samples. As is illustrated in Figure 2, smaller errors can occur in the border regions of a fish. Although the larger part of the fish body and the background have been segmented correctly, a small discrepancy can be observed at the very edge of the fish. In our opinion, this happens largely due to an inaccurate ground-truth mask, which was annotated rather coarsely, while the automatically generated mask is characterized by much smoother and more detailed edges. Another source of error are examples with a low signal-to-noise ratio as depicted in Figure 3. Those are typically images which were recorded at times of twilight, i.e. when the sun was below the horizon. A comparable case is shown in Figure 4, in which the fish becomes hardly visible while swimming out of the imaged area, towards the ocean floor. Since the recording setup does not use active lighting, to avoid an unnatural attraction effect, the lower part of the image is less well illuminated by the passive lighting from the surface. This explains the poor visibility in the lower part of the image. For those cases, the models were not able to provide a good segmentation, although the human annotator was able to clearly mark the fish. To understand this effect, it is important to consider that the human annotator could switch between frames during the annotation process, revealing the movement of the fish and making it easier for the human eye to see or guess even barely visible outlines. The segmentation models, on the other hand, only work on single frames, without any knowledge of the temporal context. Therefore, a possible exten-

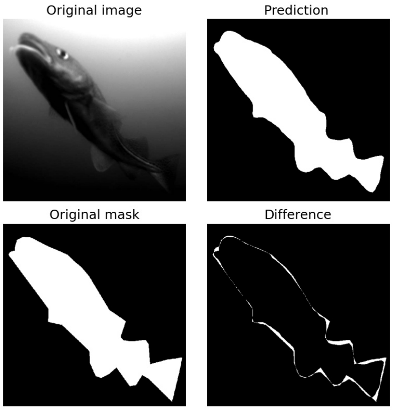

Figure 2: Illustration of a good segmentation result obtained with PSPNet, showing the input image, the predicted segmentation, the ground-truth and the difference between them. The difference between the predicted and the target mask is highlighted by white pixels.

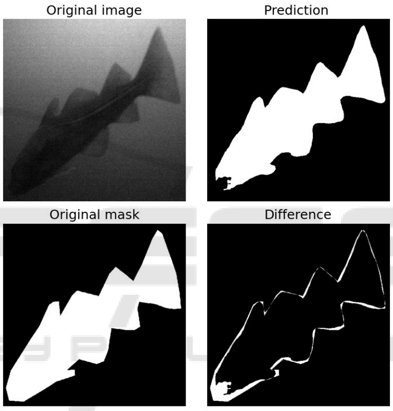

Figure 3: Example of a partially false segmentation, due to difficult viewing conditions.

sion of the algorithm would be to operate on the full video sequence, utilizing the connected frames, e.g. by an object tracking or averaging of generated segmentation masks.

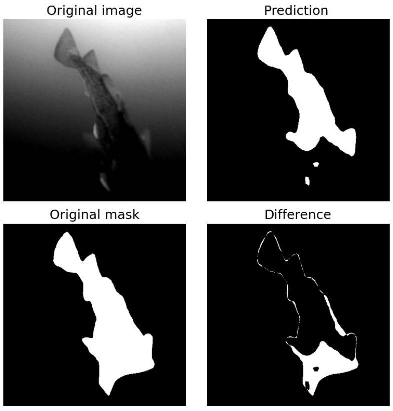

Figure 4: Example of a partially false segmentation, due to the fish swimming out of the imaged area.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have investigated PSPNet and DeepLabv3 for their applicability to fish segmentation for a fish stock monitoring application with low light cameras. The original architectures of both models were adapted for a usage on limited hardware, by interchanging their respective decoder with MobileNetV2 and by reducing the number of used layers in the encoder. We have trained and evaluated the segmentation and processing performance of each model on a custom dataset, which depicts freely swimming fishes in an unconstrained underwater environment. A larger test set of 1148 images was used to assess the models generalization capability on completely unseen data. This set includes difficult, realistic samples like fish recorded at difficult visibility conditions and swimming in and out of the imaged area. While both models perform comparably well in terms of segmentation accuracy, the PSPNet outperforms DeepLabv3 in regard to inference speed, while achieving an average pixel accuracy of 96.8% and an intersection-over-union between the predicted and the target mask of 73.8%.

In the future, we plan to extend this framework for multi-class segmentation by considering additional marine species and investigating whether segmentation results can be improved by considering the entire video sequence instead of individual frames, especially in cases of difficult visibility. Additionally, we are going to investigate several refinement steps that can enhance the segmentation results, such as Conditional Random Fields (Krähenbühl and Koltun, 2011) or affinity matrices (Liu et al., 2017). In a subsequent work, the generated segmentation masks will be used to extract morphometric features, such as fish length, and to locate specific key points on the fish outline. This information will be utilized to determine the species and biomass of the detected fish. We plan to deploy the developed algorithms on an embedded device to be used in remotely operated and autonomous underwater vehicles that complement a network of stationary underwater sensors, with the goal of performing a continuous fish stock assessment

REFERENCES

Abbasi, S. and Mokhtarian, F. (1999). Robustness of shape similarity retrieval under affine transformation. In *Challenge of Image Retrieval*, pages 1–10.

Abdeldaim, A. M., Houssein, E. H., and Hassanien, A. E. (2018). Color image segmentation of fishes with complex background in water. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 634–643. Springer.

Alsmadi, M. K. and Almarashdeh, I. (2020). A survey on fish classification techniques. *Journal of King Saud University-Computer and Information Sciences*.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

Baloch, A., Ali, M., Gul, F., Basir, S., and Afzal, I. (2017). Fish image segmentation algorithm (fisa) for improving the performance of image retrieval system. *International Journal of Advanced Computer Science and Applications*, 8(12):396–403.

Bicknell, A. W., Godley, B. J., Sheehan, E. V., Votier, S. C., and Witt, M. J. (2016). Camera technology for monitoring marine biodiversity and human impact. *Frontiers in Ecology and the Environment*, 14(8):424–432.

Cavalcanti, M. J., Monteiro, L. R., and Lopes, P. (1999). Landmark-based morphometric analysis in selected species of serranid fishes (perciformes: Teleostei). *ZOOLOGICAL STUDIES-TAIPEI-*, 38(3):287–294.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Durden, J. M., Schoening, T., Althaus, F., Friedman, A., Garcia, R., Glover, A. G., Greinert, J., Stout, N. J., Jones, D. O., Jordt, A., et al. (2016). Perspectives in visual imaging for marine biology and ecology: from acquisition to understanding. *Oceanography and marine biology: an annual review*, 54:1–72.

Fernandes, A. F., Dorea, J. R., and Rosa, G. J. (2020). Image analysis and computer vision applications in animal sciences: an overview. *Frontiers in Veterinary Science*, 7:800.

French, G., Fisher, M., Mackiewicz, M., and Needle, C. (2015). Convolutional neural networks for counting fish in fisheries surveillance video.

Garcia, R., Prados, R., Quintana, J., Tempelaar, A., Gracias, N., Rosen, S., Vågstøl, H., and Løvall, K. (2020). Automatic segmentation of fish using deep learning with application to fish size measurement. *ICES Journal of Marine Science*, 77(4):1354–1366.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. (2017). Mask R-CNN. *CoRR*, abs/1703.06870.

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

Konovalov, D. A., Saleh, A., Efremova, D. B., Domingos, J. A., and Jerry, D. R. (2019). Automatic weight estimation of harvested fish from images. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17.

Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24:109–117.

Li, L., Dong, B., Rigall, E., Zhou, T., Donga, J., and Chen, G. (2021). Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., and Kautz, J. (2017). Learning affinity via spatial propagation networks. *arXiv preprint arXiv:1710.01020*.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Lopez-Marcano, S., Brown, C. J., Sievers, M., and Connolly, R. M. (2021). The slow rise of technology: Computer vision techniques in fish population connectivity. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31(1):210–217.

Malde, K., Handegard, N. O., Eikvil, L., and Salberg, A.-B. (2020). Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77(4):1274–1285.

Mallet, D. and Pelletier, D. (2014). Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). *Fisheries Research*, 154:44–62.

Marchesan, M., Spoto, M., Verginella, L., and Ferrero, E. A. (2005). Behavioural effects of artificial light on fish species of commercial interest. *Fisheries research*, 73(1-2):171–185.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016). Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, 3:473–480.

Qin, H., Peng, Y., and Li, X. (2014). Foreground extraction of underwater videos via sparse and low-rank matrix decomposition. In *2014 ICPR Workshop on Computer Vision for Analysis of Underwater Imagery*, pages 65–72. IEEE.

Rawat, S., Benakappa, S., Kumar, J., Naik, K., Pandey, G., and Pema, C. (2017). Identification of fish stocks based on truss morphometric: A review. *Journal of Fisheries and Life Sciences*, 2(1):9–14.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rosen, S. and Holst, J. C. (2013). Deepvision in-trawl imaging: Sampling the water column in four dimensions. *Fisheries Research*, 148:64–73.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y.-H. J., Fisher, R. B., and Nadarajan, G. (2010). Automatic fish classification for underwater species behavior understanding. In *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, pages 45–50.

Storbeck, F. and Daan, B. (2001). Fish species recognition using computer vision and a neural network. *Fisheries Research*, 51(1):11–15.

Tchapmi, L., Choy, C., Armeni, I., Gwak, J., and Savarese, S. (2017). Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE.

Wilhelms, I. et al. (2013). Atlas of length-weight relationships of 93 fish and crustacean species from the north sea and the north-east atlantic. Technical report, Johann Heinrich von Thünen Institute, Federal Research Institute for Rural ....

Yang, X., Zeng, Z., Teo, S. G., Wang, L., Chandrasekhar, V., and Hoi, S. (2018). Deep learning for practical image recognition: Case study on kaggle competitions. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 923–931.

Yu, C., Fan, X., Hu, Z., Xia, X., Zhao, Y., Li, R., and Bai, Y. (2020). Segmentation and measurement scheme for fish morphological features based on mask r-cnn. *Information Processing in Agriculture*, 7(4):523–534.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.