Journal of Geography and Cartography (2023) Volume 6 Issue 1
doi:10.24294/jgc.v6i1.1980

REVIEW ARTICLE

Density estimation of the main structuring sessile species in underwater marine caves with a deep learning approach

Sergio Sierra[1,2], Elena Prado[3], Luis Rodríguez-Cobo[2,4,5], Carla Quiles-Pons[6], Pablo Roldán-Varona[2,4], David Díaz-Viñolas[6], Pedro Anuarbe-Cortés[2], Adolfo Cobo[3,4,5*], Francisco Sánchez[3]

[1] *Complutum Tecnologías de la Información Geográfica, COMPLUTIG, Alcalá de Henares 28801, Spain.*

[2] *Photonics Engineering Group, Universidad de Cantabria, Santander 39005, Spain.*

[3] *Instituto Español de Oceanografía (IEO-CSIC), Centro Oceanográfico de Santander, Santander 39004, Spain. E-mail: adolfo.cobo@unican.es*

[4] *Instituto de Investigación Sanitaria Valdecilla (IDIVAL), Santander 39011, Spain.*

[5] *CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Instituto de Salud Carlos III, Madrid 28029, Spain.*

[6] *Instituto Español de Oceanografía (IEO-CSIC), Centro Oceanográfico de Baleares, Palma de Mallorca 07015, Spain.*

ABSTRACT

Monitoring marine biodiversity is a challenge in some vulnerable and difficult-to-access habitats, such as underwater caves. Underwater caves are a great focus of biodiversity, concentrating a large number of species in their environment. However, most of the sessile species that live on the rocky walls are very vulnerable, and they are often threatened by different pressures. The use of these spaces as a destination for recreational divers can cause different impacts on the benthic habitat. In this work, we propose a methodology based on video recordings of cave walls and image analysis with deep learning algorithms to estimate the spatial density of structuring species in a study area. We propose a combination of automatic frame overlap detection, estimation of the actual extent of surface cover, and semantic segmentation of the main 10 species of corals and sponges to obtain species density maps. These maps can be the data source for monitoring biodiversity over time. In this paper, we analyzed the performance of three different semantic segmentation algorithms and backbones for this task and found that the Mask R-CNN model with the Xception101 backbone achieves the best accuracy, with an average segmentation accuracy of 82%.

Keywords: Marine Biodiversity; Underwater Caves; Underwater Images; Deep Learning; Semantic Segmentation

ARTICLE INFO

Received: 10 March 2023
Accepted: 6 May 2023
Available online: 30 May 2023

1. Introduction

In the Mediterranean Sea, underwater caves, usually formed by rocky walls with different cavities and dark or semi-dark conditions, offer an environment that attracts a multitude of species, constituting a great focus of biodiversity in our seas. The more than 3,000 known caves in the Mediterranean Sea are the subject of many studies due to their particular environmental conditions, low ecological resilience, and the presence of many species of conservation interest[1,2]. Among the sessile species, sponges are the dominant group, with a total of 311 species of all Porifera classes recorded, which represents 45.7% of Mediterranean Porifera[3]. Marine caves are considered a priority habitat for conservation included in the EU Habitats Directive (Habitat 8330). Although a large number of fragile benthic communities inhabit the interior of underwater caves, the existing knowledge about these habitats, and, in particular, multi-year studies, are very scarce[4]. This is due to access difficulties and sampling limitations that make it difficult to create detailed habitats cartography, species inventories, and studies of community dynamics or damage assessment[2,5,6].

The use of video or photographic cameras for sampling in caves is becoming a basic tool. This type of approach allows the registration of the habitats in a minimally invasive way and reduces the necessary sampling time if we compare it with the classical type of sampling[1,7]. Divers can record video transects inside the cave in a simple way, also minimizing immersion time. However, further analysis of this information still has certain limitations. The lack of automation of image analysis makes species identification a time-consuming process. In addition, image annotation has to be made by an expert taxonomist who can address this task. This means that, on many occasions, it is not possible to analyze the entire existing image data set, and subsampling strategies are imposed.

To solve the difficulties of massive image data processing, deep learning algorithms are proving to be a suitable solution[8]. Deep learning algorithms have been proposed as a powerful tool for monitoring different underwater habitats from recorded images or videos, including shallow and turbid waters[9], or deep benthic communities[10].

These algorithms classify an entire image at the pixel level, which allows for the accurate identification of objects and the recovery of their shape and size. There are numerous deep learning algorithms available for this type of task, including popular ones such as Mask R-CNN or U-Net, which are usually trained on standard datasets like ImageNet or COCO, and fine-tuned to specific classes for increased accuracy. Fine-tuning is a process in which a pre-trained deep learning model is further trained on a specific dataset to improve its performance for a particular task. In the context of semantic segmentation, fine-tuning involves taking a pre-trained model that has already learned features from a large and diverse dataset and adjusting it to fit a specific set of classes or a new dataset. During fine-tuning, the pre-trained model is re-trained on the new dataset with the addition of a few new output layers to classify the desired classes or objects. The weights of the original model are then updated based on the new data, while the original pre-trained weights are kept fixed. This process allows the model to learn more relevant features specific to the new dataset while still retaining the useful knowledge from the pre-trained model. Fine-tuning is a powerful technique that can significantly improve the accuracy and generalization of semantic segmentation models for specific tasks or datasets, and it has been widely used in various computer vision applications.

In the context of marine cave habitat monitoring, semantic segmentation can be used to accurately identify and track species populations, detect changes in habitat conditions over time, and identify potential threats to the ecosystem[11]. Researchers can benefit from the use of semantic segmentation as it can greatly increase the accuracy and efficiency of their data collection and analysis efforts, allowing them to make more informed decisions and take appropriate actions to protect these delicate habitats.

The objective of this study is to develop an efficient and robust underwater image segmentation algorithm for the detailed description of the main structuring species of an underwater cave habitat.

2. Methodology

2.1 Images capturing

The images were captured in the cave known as “La Catedral” (The Cathedral), Illa de L’Aire, Balearic Islands, Spain, and it is a well-known site frequented by recreational diving clubs on the island. This cave has been studied for years with different approaches[6]. For this work, a video recording campaign was carried out in April 2021, as part of the INTEMARES project, where one of its objectives is monitoring the impacts of divers on the benthic habitats in marine caves. A total of 32 minutes of video were recorded along two linear paths (transects) following the cave walls, with a total recorded path length of about 40 meters. The recording place is previously marked by placing a tape measure or rope with metric marks. Videos have a frame size of 1,920 × 1,080 pixels and 50 frames per second.

Thanks to what has been learned in these video surveys, during the VirtualMAR project, specially adapted instrumentation has been developed to cover the monitoring needs of this type of habitat. A custom underwater video camera was designed. It includes a Blackmagic Pocket Cinema 4K video camera (Blackmagic Design, San Francisco, USA) with a Micro Four Thirds sensor and a 12 mm f2.8

objective, a custom acrylic dome rated for 150-meter depth, LiPo batteries for two-hours autonomy, and LED lighting. The system can be controlled by a diver or attached to a small remotely-operated vehicle (ROV) such as Bluerobotics' BlueRov. **Figure 1** shows the underwater camera operated by a diver.

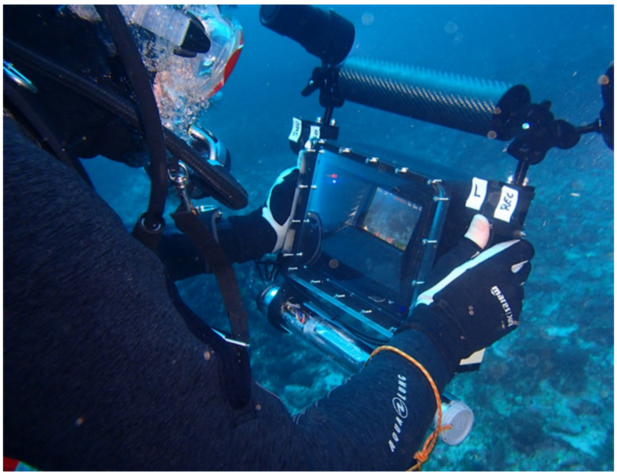


Figure 1. Custom underwater video camera developed to acquire images by a diver.

2.2 Frame overlapping

For studies of species richness and biodiversity assessment, or for the generation of species inventories, the sampling method used influences notably the results obtained[12]. For this reason, when the data source for this evaluation consists of underwater videos, care must be taken not to count the same specimens several times, and for this reason, the extraction of frames from the video files is a crucial task. To address this task, an automatic approach has been used that requires the identification of unique features that are repeated in contiguous frames. This is the problem known as Simultaneous Localization and Mapping (SLAM), which is even more challenging in underwater environments[13]. For this work, we have used the Scale Invariant Feature Transform (SIFT) algorithm[14], a powerful computer vision technology that allows us to identify and extract unique features from video. By analyzing the distribution of gradients and orientations within an image, SIFT can identify key points that are highly distinctive and can be used to match the same object or scene across multiple frames.

Using this algorithm, we have created a tool that can identify and extract unique frames from a video. Using a threshold of 10 common points, the tool can determine if a given frame contains new and unique content or is simply a duplicate or partially the same as an older frame. This allows us to automatically extract only the most significant frames from a video, saving time and reducing the amount of data that needs to be processed.

2.3 True size estimation

Estimating the true size of the recorded surface and marine species on it is an essential task for researchers studying these environments. The area covered by an image or swept area is the value that allows the transformation of species presence data into densities. The density parameter of a species together with its geographic distribution is key to addressing any study and being able to draw relevant conclusions about its conservation status. In this work, we have developed a custom algorithm that provides the scale information from reference cords placed over the surface, with blue-over-white marks spaced at 10 cm or 20 cm intervals. From the distance in pixels between consecutive color marks, the true size of the image is estimated (**Figure 2**).

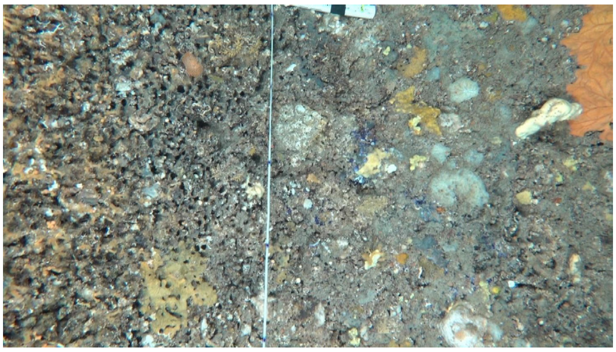


Figure 2. A reference cord with spaced marks and an automatic identification algorithm has been used to estimate the true size of each frame.

The code takes an input image and a binary mask as input. It first applies the mask to the image, then applies a color threshold to the resulting image to isolate the blue marks. It then finds the contours of the blue objects and calculates the centers of each contour. The code then calculates the distance between each pair of centers, finds the maximum distance, and calculates the real-world area of the image by converting the maximum distance into a distance in meters and using that to calculate the area.

The area is calculated based on the maximum distance between two white circles (balls) detected in the image. This maximum distance is obtained by finding the Euclidean distance between the centers of all possible pairs of circles. Once the maximum distance is obtained, it is used to calculate the approximate size of the area in square meters, assuming a fixed distance of 10 cm or 20 cm (depending on the reference cord) between the points in the real world. The area is calculated as the product of the height and width of the image in meters divided by the square of the maximum distance between circles.

2.4 Semantic segmentation of species

2.4.1 Dataset

The dataset of images has been automatically created from the recorded videos with the above-described algorithms. It consists of a total of 153 images, which were annotated by experts from the Spanish Oceanography Institute (IEO-CSIC) with over 30 different species (see **Table 1**). However, more than half of these species have a very low density in the explored area, so they were discarded, and the semantic segmentation algorithm was trained with 10 species, which are listed below. The selection criteria for these species were that the entire dataset should contain at least 50 instances of the species, that is, they should appear at least fifty times in the images. The dataset was split into a 60% training set, a 20% validation set, and a 20% testing set. This methodology ensures that the model is trained on a balanced dataset with sufficient samples of the selected species, which should improve its accuracy and robustness.

The training set was used to train the model, the validation set was used to evaluate the model's performance and hyperparameter optimization, and the test set was used to assess the final performance of the model on unseen data. The test set was selected beforehand to be representative of the overall dataset and was kept separate from the training and validation sets to avoid bias and overfitting.

2.4.2 Selected species

The selected classes, i.e., species for the semantic segmentation model are composed of sponges (*Agelas oroides*, *Spirastrella cunctatrix*, *Acanthella acuta*), corals (*Parazoanthus axinellae*, *Axinella*), tunicata (*Didemmnum*) and bryozoa (*Schizoretepora serratimargo*, *Reteporella*, *Frondipora verrucosa*, *Myriaphora truncata*). By focusing on these classes, the model can be trained to accurately identify and segment these species in new images. This should enable researchers to study and monitor the populations of these species in the exploration area more effectively.

2.4.3 Preprocessing techniques

One common technique used for pre-processing images for semantic segmentation models is normalization[15]. Normalization involves scaling the pixel values of an image so that they fall within a certain range (typically [0,1] or [–1,1]). This can be important because different images may have different brightness and contrast levels, which can affect the performance of the model. By normalizing the images, we can ensure that the model is not biased toward certain brightness or contrast levels.

Other pre-processing techniques that may be used for semantic segmentation models include resizing or cropping the images to a standard size; we have tried this pre-processing technique over several resolutions. The best results have been achieved by pre-processing images into crops of 640 × 640, 512 × 512, and 256 × 256 pixels. This could be because most models have these resolutions as their native resolution. When an image is processed, it is often resized to fit the input requirements of the model being used. By using these common resolutions as crops, the input to the model can be processed more efficiently and effectively.

Another technique that is often used is data

Table 1. Sessile species of interest in habitat 8830 considered in this study

Demospongiae sp.	*Acanthella acuta*	*Caryophyllia inornate*
Axinella sp.	*Bryozoa* sp.	*Didemmnum* sp.
Serpulidae sp.	*Filograna implexa*	*Reptadeonella violacea*
Parazoanthus axinellae	*Palmophyllum crassum*	*Terpios* sp.
Leptopsammia pruvoti	*Mesophyllum* sp.	*Schizoretepora serratimargo*
Reteporella sp.	*Smittina cervicornis*	*Miniacina miniacea*
Petrosia ficiformis	*Spirastrella cunctatrix*	*Madracis pharensis*
Agelas oroides	*Haliclona mucosa*	*Frondipora verrucosa*
Myriapora truncata	*Ascidiacea* sp.	*Porifera* sp.
Peyssonnelia sp.	*Chondosia reniformis*	*Cliona* sp.

augmentation[15,16]. This involves generating new training images by applying various transformations (such as rotation, scaling, and flipping) to the original images. This can help to increase the diversity of the training data and improve the model's ability to generalize to new, unseen images.

2.4.4 Semantic segmentation algorithms

These algorithms can classify an image into classes at the pixel level, thus estimating not only the number of species detected by their shapes. We have experimented with several state-of-the-art algorithms for semantic segmentation, including Mask R-CNN and its various backbone networks (ResNet101, ResNet50, and Xception101), as well as U-Net with VGG16, and some newer models like ResNeSt and ConvNeXt.

Mask R-CNN is a popular algorithm that builds on the success of Faster R-CNN by adding a mask prediction branch to enable pixel-level segmentation. It has been shown to be effective in a range of tasks, including object detection and instance segmentation. The choice of backbone network can affect the performance of Mask R-CNN, with larger networks like ResNet101 generally performing better at the cost of increased computational requirements. Adapting the Mask R-CNN algorithm to assign the same color to all instances of a particular class, you have essentially transformed it from an instance segmentation model into a semantic segmentation model.

Instance segmentation involves identifying and segmenting each individual instance of an object class within an image, whereas semantic segmentation involves labeling each pixel in an image with a corresponding class label. By assigning a consistent color to all instances of a class, you are effectively treating them as a single entity for the purpose of semantic segmentation. This modification may be useful in scenarios where the specific instances of an object class are not as important as the overall distribution of that class within an image. By simplifying the output of the segmentation model in this way, it may be easier for human analysts to quickly understand and interpret the results.

U-Net is another popular algorithm that has been widely used in medical image segmentation[17]. It uses a symmetric encoder-decoder architecture that allows for the precise localization of objects and has been shown to be effective in a variety of applications. The choice of backbone network can also affect the performance of U-Net, with VGG16 being a popular choice due to its good performance on image classification tasks.

ResNeSt[18] and ConvNeXt[19] are newer models that have shown promising results on image classification tasks and have also been adapted for use in semantic segmentation. ResNeSt is designed to improve the accuracy and efficiency of ResNet by introducing a nested scale-reduction strategy, while ConvNeXt introduces a new convolution operation that combines the benefits of both depthwise and pointwise convolutions.

2.4.5 Model selection

Precision, recall, and F1 scores are common evaluation metrics used in machine learning to assess the performance of classification models[20]. Precision measures the proportion of true positives among all positive predictions made by the model. It indicates the model's ability to avoid false positives. Recall, on the other hand, measures the proportion of true positives identified by the model among all actual positives in the dataset. It indicates the model's ability to avoid false negatives. F1 score is the harmonic mean of precision and recall, which provides a balanced view of both metrics. It is often used when precision and recall have to be balanced, such as in cases where false positives and false negatives have similar costs. It's important to evaluate and compare different models on the specific task and data to determine which one is the best fit, try out several different models and hyperparameters, and fine-tune them on the data to achieve the best performance.

3. Results and conclusions

3.1 Frame overlapping

The results of the tool that identifies and extracts unique frames from a video were impressive, as it was validated on a 3-minute video whose frames were manually extracted by an expert, and the output was the same number of frames minus one. This proves that the tool is reliable and efficient in identifying unique frames in a video. This is especially valuable when studying underwater

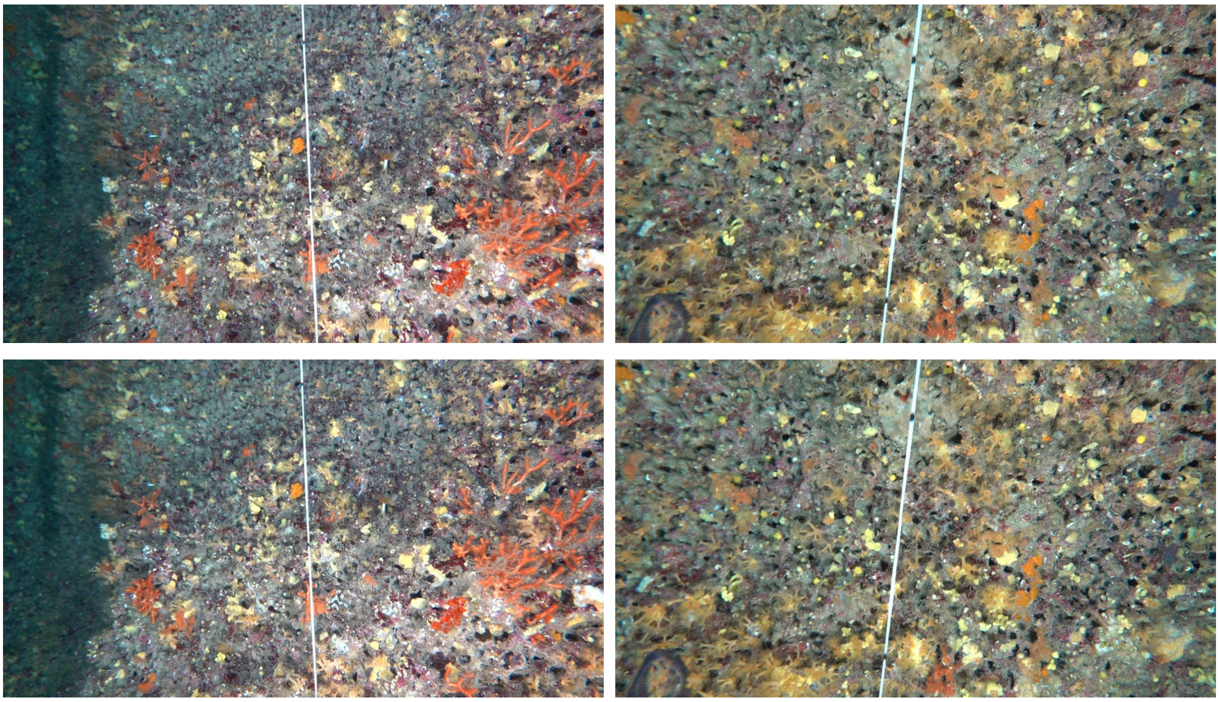


Figure 3. Comparison of the manual extraction of a unique frame (frames #2,400 and #5,350) by an expert (top) and by SIFT algorithm (bottom).

caves, where divers often record long, slow videos that contain many repetitive frames. Currently, the task of manually extracting frames with unique content is time-consuming and labor-intensive, requiring human experts to carefully review each frame to identify those that contain new information. A comparison between the manually extracted single frame (frames #2,400 and #5,350 in the video sequence) at the top of the figure and the frames automatically extracted by the SIFT algorithm (below) is shown in **Figure 3**.

3.2 Best performance model

Precision, recall, and F1 score obtained after the training of the models are shown in **Table 2**. The architecture with the best results was Mask R-CNN[21]. In any of the three backbone variants used, better results were obtained than in the other three models. We have found that it is not always the case that the newest or most advanced models will perform better than older models in all scenarios.

Table 2. Performance metrics of the tested models. Mask R-CNN has the best performance

Model	**Mean precision**	**Mean recall**	**Mean F1 score**
U-Net	0.579	0.633	0.587
Mask R-CNN	0.822	0.736	0.777
ConvNeXt	0.632	0.654	0.643
ResNeSt	0.595	0.61	0.603

Here is the general architecture of the Mask R-CNN model (**Figure 4**) with ResNet101 backbone:

a. Backbone Network: The input image is passed through the ResNet101 network to extract features. The output of the last convolutional layer is a feature map of size H/32 × W/32 × 2,048, where H and W are the height and width of the input image.
b. Region Proposal Network (RPN): The RPN takes the feature map generated by the backbone network as input and generates a set of region proposals. These proposals are regions in the image that are likely to contain objects. The RPN predicts the objectness score and bounding box coordinates for each proposal.
c. Region of Interest (RoI) Align: The RoI Align layer extracts features from each region proposal and produces a fixed-size feature map for each proposal.
d. Mask Head: The Mask Head takes the feature maps produced by the RoI Align layer and produces a binary mask for each object proposal. The Mask Head is a fully convolutional network that takes as input the RoI feature map and produces a mask of size m × m, where m is a fixed size.
e. Classification Head: The Classification Head takes the same RoI feature maps as the Mask Head and produces class probabilities for

each object proposal.

f. Loss Function: The Mask R-CNN model uses a multi-task loss function that combines the losses for object detection, object classification, and mask prediction.

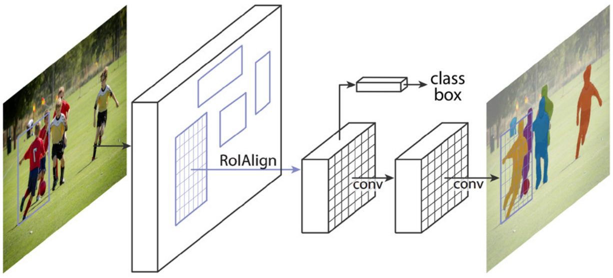
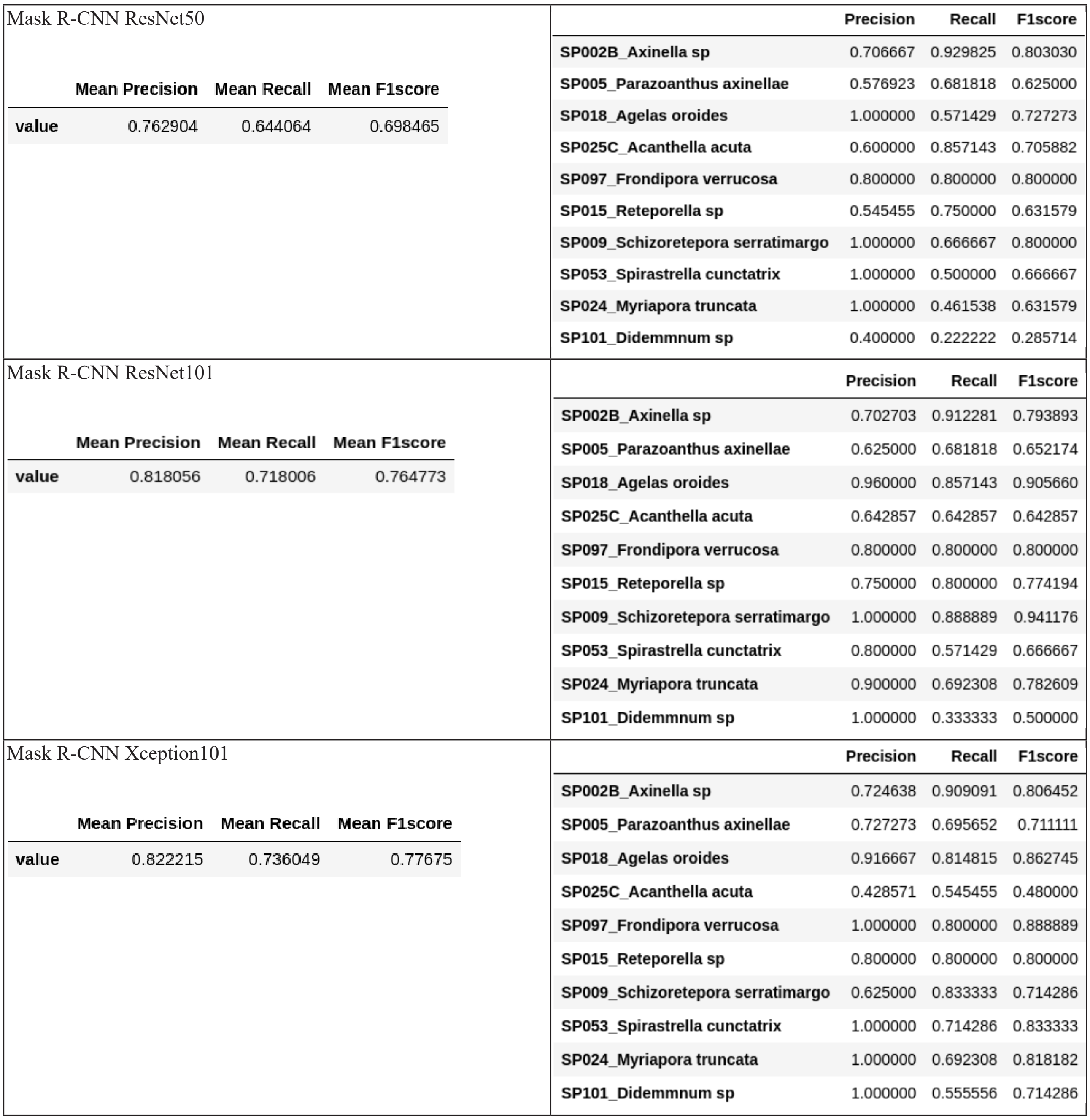


Figure 4. Mask R-CNN architecture, reproduced from [21].

Overall, the Mask R-CNN with ResNet101 backbone architecture is a deep neural network that is trained end-to-end for object detection and instance segmentation tasks. It is a powerful and widely used model in computer vision applications.

3.3 Metrics

The model Mask R-CNN has been trained with three different backbones (ResNet 50, ResNet 101, and Xception101) searching for the best segmentation performance. **Table 3** shows the results, which are better for the later backbone network. The detailed segmentation metrics are shown in **Figure 5**. It can be seen that the Xception101 backbone gives



Mask R-CNN ResNet50

	Mean Precision	Mean Recall	Mean F1score
value	0.762904	0.644064	0.698465

	Precision	Recall	F1score
SP002B_Axinella sp	0.706667	0.929825	0.803030
SP005_Parazoanthus axinellae	0.576923	0.681818	0.625000
SP018_Agelas oroides	1.000000	0.571429	0.727273
SP025C_Acanthella acuta	0.600000	0.857143	0.705882
SP097_Frondipora verrucosa	0.800000	0.800000	0.800000
SP015_Reteporella sp	0.545455	0.750000	0.631579
SP009_Schizoretepora serratimargo	1.000000	0.666667	0.800000
SP053_Spirastrella cunctatrix	1.000000	0.500000	0.666667
SP024_Myriapora truncata	1.000000	0.461538	0.631579
SP101_Didemmnum sp	0.400000	0.222222	0.285714

Mask R-CNN ResNet101

	Mean Precision	Mean Recall	Mean F1score
value	0.818056	0.718006	0.764773

	Precision	Recall	F1score
SP002B_Axinella sp	0.702703	0.912281	0.793893
SP005_Parazoanthus axinellae	0.625000	0.681818	0.652174
SP018_Agelas oroides	0.960000	0.857143	0.905660
SP025C_Acanthella acuta	0.642857	0.642857	0.642857
SP097_Frondipora verrucosa	0.800000	0.800000	0.800000
SP015_Reteporella sp	0.750000	0.800000	0.774194
SP009_Schizoretepora serratimargo	1.000000	0.888889	0.941176
SP053_Spirastrella cunctatrix	0.800000	0.571429	0.666667
SP024_Myriapora truncata	0.900000	0.692308	0.782609
SP101_Didemmnum sp	1.000000	0.333333	0.500000

Mask R-CNN Xception101

	Mean Precision	Mean Recall	Mean F1score
value	0.822215	0.736049	0.77675

	Precision	Recall	F1score
SP002B_Axinella sp	0.724638	0.909091	0.806452
SP005_Parazoanthus axinellae	0.727273	0.695652	0.711111
SP018_Agelas oroides	0.916667	0.814815	0.862745
SP025C_Acanthella acuta	0.428571	0.545455	0.480000
SP097_Frondipora verrucosa	1.000000	0.800000	0.888889
SP015_Reteporella sp	0.800000	0.800000	0.800000
SP009_Schizoretepora serratimargo	0.625000	0.833333	0.714286
SP053_Spirastrella cunctatrix	1.000000	0.714286	0.833333
SP024_Myriapora truncata	1.000000	0.692308	0.818182
SP101_Didemmnum sp	1.000000	0.555556	0.714286

Figure 5. Detailed segmentation metrics for each species.

Table 3. Performance metrics for Mask R-CNN model with three different backbones

Model + backbone	Mean precision	Mean recall	Mean F1 score
Mask R-CNN + ResNet50	0.763	0.644	0.698
Mask R-CNN + ResNet101	0.81	0.718	0.764
Mask R-CNN + Xception101	0.822	0.736	0.777

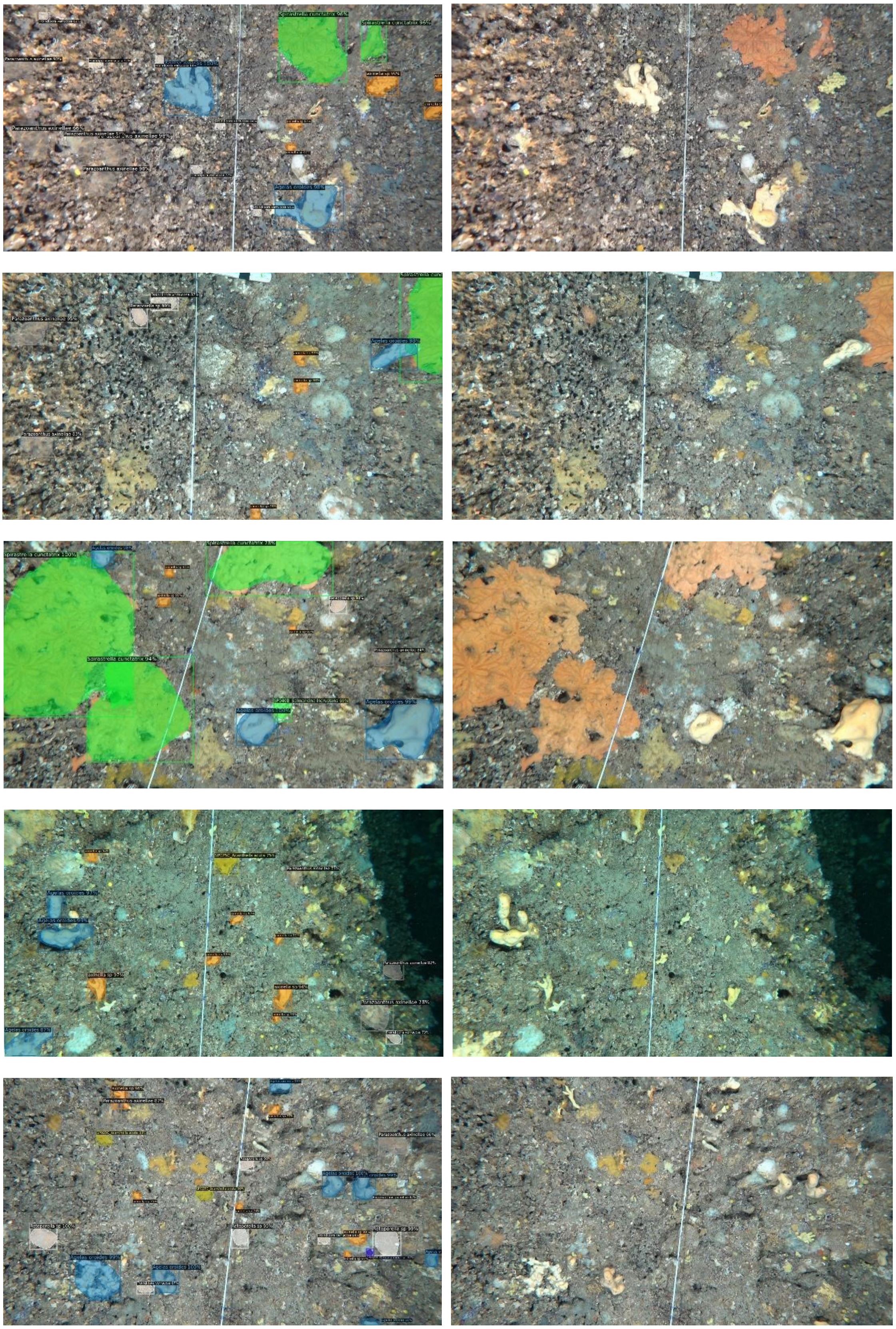

Spirastrella cunctatrix 100%
Agelas oroides 98%
Spirastrella cunctatrix 78%
Spirastrella cunctatrix 94%
SP040B_Scalarispongia incrustans 99%
Agelas oroides 100%
Agelas oroides 99%
SP025C_Acanthella acuta 75%

Figure 6. Examples of the semantic segmentation output. The algorithm automatically extracts the species and shape of each specimen.

the best averaged results, and the sponge *Aanthella acuta* is the most difficult to identify.

3.4 Sessile species density

In this study, we utilized the SIFT model to select frames from a video that covers new areas, a custom algorithm to obtain the true size of the images from reference points, and semantic segmentation models to obtain density data of sessile species. Ten of the main species that conform to the habitat of a submerged marine cave have been selected, including sponges, corals, and bryozoans. A benchmarking of different models resulted in the Mask R-CNN model with the Xception101 network as a backbone as the best-performing model with our dataset. By doing so, we were able to obtain several outputs, including units of each species across the transect, density per square meter of each species, and an estimation of the mean size of each species.

In **Figure 6**, some examples of the semantic segmentation algorithm output are shown. The shape and inferred species can be seen as colored areas. From this pixel-level identification, the total number of species in the transect and the density of specimens per square meter are extracted (**Table 4**), according to the total area covered in the transect of 139.87 m^2. **Figure 7** summarizes the relative abundance of species in this transect. The randomly chosen test set did not include any specimens of *Schizoretepora serratimargo*, that is the reason that the number of specimens is zero and their density could be calculated.

Table 4. Number of specimens and their density

Species	Number of specimens across the complete transect	Density (specimens/m^2) of species
Axinella sp.	630	4.5
Parazoanthus axinellae	196	1.4
Agelas oroides	155	1.11
Acanthella acuta	83	0.59
Myriapora truncata	77	0.55
Reteporella sp.	74	0.53
Frondipora verrucosa	66	0.47
Spirastrella cunctatrix	33	0.24
Didenmnum sp.	15	0.11
Schizoretepora serratimargo	0	0

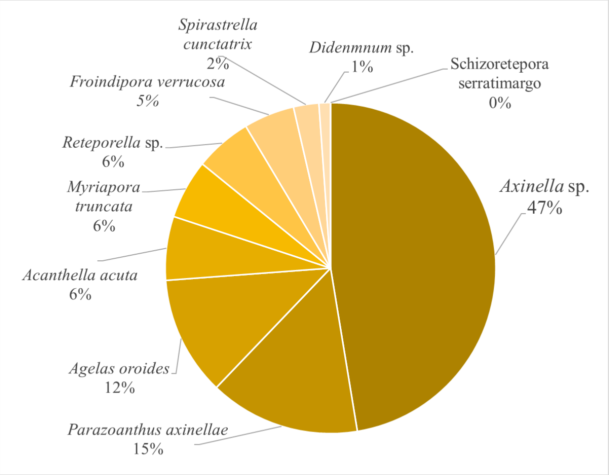


Figure 7. Relative abundance of species.

3.5 Concluding remarks

In conclusion, the use of deep learning algorithms for analyzing underwater images is proving to be a powerful tool for monitoring marine biodiversity, particularly in vulnerable and difficult-to-access habitats such as underwater caves. In this work, we proposed a methodology based on video recordings of the area under study following no particular patterns. From the video, an algorithm to extract non-overlapped frames was applied, and the true size of each image was also automatically estimated from reference cords. The images were processed by a semantic segmentation algorithm to automatically detect specimens, their species, size, and shape. We tested several algorithms for this segmentation task and found the Mask R-CNN model with Xception101 backbone achieved the best accuracy, with an average segmentation accuracy of 82%. This methodology could be applied to monitoring marine biodiversity over time, identifying potential threats to the ecosystem, such as the influence of environmental variables or anthropogenic stress, and in particular, the impact of recreational diving. This will allow for making more informed decisions for the conservation and protection of these delicate habitats. With the increasing availability of underwater cameras and the development of more efficient algorithms, the potential for using deep learning techniques in marine biodiversity monitoring is enormous.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work was supported by the Organismo Autónomo de Parques Nacionales, Ministerio para la Transición Ecológica y el Reto Demográfico, Gobierno de España, Project VirtualMAR (codes 2470-S/2017 and 2493-S/2017); by IP INTEMARES project (LIFE15 IPE/ES/000012); by the R+D project PID2019-107270RB-C21 (funded by MCIN/AEI/10.13039/501100011033); by Plan Nacional de I+D+I; and by Instituto de Salud Carlos III (ISCIII), Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Ciencia, Innovación y Universidades.

Reference

1. Gerovasileiou V, Bianchi CN. Mediterranean marine caves: A synthesis of current knowledge. Boca Raton: CRC Press; 2021. p. 87.
2. Navarro-Barranco C, Ambroso S, Gerovasileiou V, *et al*. Conservation of dark habitats. In: Espinosa F (editor). Anonymous coastal habitat conservation. Cambridge: Academic Press; 2023. p. 147–170.
3. Gerovasileiou V, Voultsiadou E. Marine caves of the Mediterranean Sea: A sponge biodiversity reservoir within a biodiversity hotspot. PLoS One 2012; 7(7): e39873. doi: 10.1371/journal.pone.0039873.
4. Montefalcone M, De Falco G, Nepote E, *et al*. Thirty year ecosystem trajectories in a submerged marine cave under changing pressure regime. Marine Environmental Research 2018; 137: 98–110. doi: 10.1016/j.marenvres.2018.02.022.
5. Gerovasileiou V, Trygonis V, Sini M, *et al*. Three-dimensional mapping of marine caves using a handheld echosounder. Marine Ecology Progress Series 2013; 486: 13–22. doi: 10.3354/meps10374.
6. Quiles-Pons C, Baena I, Calvo-Manazza M, *et al*. Monitoring the complex benthic habitat on semi-dark underwater marine caves using photogrammetry-based 3D reconstructions. In: Proceedings of 3rd Mediterranean Symposium on the Conservation of the Dark Habitats; 2022 Sep 21–22; Genoa. Palma De Mallorca: Centro Oceanográfico de Baleares; 2022.
7. Dimarchopoulou D, Gerovasileiou V, Voultsiadou E. Spatial variability of sessile benthos in a semi-submerged marine cave of a remote Aegean Island (eastern Mediterranean Sea). Regional Studies in Marine Science 2018; 17: 102–111. doi: 10.1016/j.rsma.2017.11.015.
8. Er MJ, Chen J, Zhang Y, Gao W. Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: A review. Sensors 2023; 23(4): 1990. doi: 10.3390/s23041990.
9. Mohamed H, Nadaoka K, Nakamura T. Automatic semantic segmentation of benthic habitats using images from towed underwater camera in a complex shallow water environment. Remote Sensing 2022; 14(8): 1818. doi: 10.3390/rs14081818.
10. Abad-Uribarren A, Prado E, Sierra S, *et al*. Deep learning-assisted high-resolution mapping of vulnerable habitats within the Capbreton Canyon System, Bay of Biscay. Estuarine, Coastal and Shelf Science 2022; 275: 107957. doi: 10.1016/j.ecss.2022.107957.
11. Pierce JP, Rzhanov Y, Lowell K, Dijkstra JA. Reducing annotation times: Semantic segmentation of coral reef survey images. In: Proceedings of Global Oceans 2020: Singapore–U.S. Golf Coast; 2020 Oct 5–30; Biloxi. New York: IEEE; 2020. p. 1–9.
12. Stobart B, Díaz D, Álvarez F, *et al*. Performance of baited underwater video: Does it underestimate abundance at high population densities? PLoS One 2015; 10(5): e0127559. doi: 10.1371/journal.pone.0127559.
13. Zhang S, Zhao S, An D, *et al*. Visual SLAM for underwater vehicles: A survey. Computer Science Review 2022; 46: 100510. doi: 10.1016/j.cosrev.2022.100510.
14. Lindeberg T. Scale invariant feature transform. Scholarpedia 2012; 7(5): 10491. doi: 10.4249/scholarpedia.10491.
15. Moreno-Barea FJ, Jerez JM, Franco L. Improving classification accuracy using data augmentation on small data sets. Expert Systems with Applications 2020; 161: 113696. doi: 10.1016/j.eswa.2020.113696.
16. Han F, Yao J, Zhu H, Wang C. Underwater image processing and object detection based on deep CNN method. Journal of Sensors 2020; 2020. doi: 10.1155/2020/6707328.
17. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A (editors). Proceedings of Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference; 2015 Oct 5–9; Munich. Cham: Springer International Publishing; 2015. p. 234–241.
18. Zhang H, Wu C, Zhang Z, *et al*. ResNeSt: Split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops; 2022 Jun 19–20; New Orleans. New York: IEEE; 2022. p. 2736–2746.

19. Liu Z, Mao H, Wu C, *et al*. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans. New York: IEEE; 2022. p. 11976–11986.
20. Gulati M. How to choose evaluation metrics for classification models [Internet]. Gurgaon: Analytics Vidhya; 2020 [updated 2020 Oct 11]. Available from: https://www.analyticsvidhya.com/blog/2020/10/how-to-choose-evaluation-metrics-for-classification-model/.
21. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice. New York: IEEE; 2018. p. 2961–2969.