Automated Marking of Underwater Animals Using a Cascade of Neural Networks

Oleg Iakushkin[1,3(✉)], Ekaterina Pavlova[1,3], Evgeniy Pen[3], Anna Frikh-Khar[3,4], Yana Terekhina[2], Anna Bulanova[2], Nikolay Shabalin[2], and Olga Sedova[1]

[1] Saint-Petersburg University, Saint Petersburg, Russia
o.yakushkin@spbu.ru
[2] Lomonosov Moscow State University, Moscow, Russia
[3] BioGeoHab, Saint Petersburg, Russia
[4] Marine Research Center, Lomonosov Moscow State University, Moscow, Russia
https://www.biogeohab.com

Abstract. In this work, a multifactorial problem of analyzing the seabed state of plants and animals using photo and video materials is considered. Marine research to monitor benthic communities and automatic mapping of underwater landscapes make it possible to qualitatively assess the state of biomes. The task includes several components: preparation of a methodology for data analysis, their aggregation, analysis, presentation of results. In this work, we focused on methods for automating detection and data presentation.

For deep-sea research, which involves the detection, counting and segmentation of plants and animals, it is difficult to use traditional computer vision techniques. Thanks to modern automated monitoring technologies, the speed and quality of research can be increased several times while reducing the required human resources using machine learning and interactive visualization methods.

The proposed approach significantly improves the quality of the segmentation of objects underwater. The algorithm includes three main stages: correction of image distortions underwater, image segmentation, selection of individual objects. Combining neural networks that successfully solve each of the tasks separately into a cascade of neural networks is the optimal method for solving the problem of segmentation of aquaculture and animals.

Using the results obtained, it is possible to facilitate the control of the ecological state in the world, to automate the task of monitoring underwater populations.

Keywords: Few-shot learning · Neural networks · Video analysis · Segmentation

O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12956, pp. 460–470, 2021.
https://doi.org/10.1007/978-3-030-87010-2_34

1 Introduction

Today, marine research is relevant for monitoring benthic communities [22] and automatic mapping of underwater landscapes [9]. The tasks of segmentation and recognition underwater can be solved using machine learning methods [15]. Counting populations is necessary for analyzing the ecological state of the environment.

It is necessary to have labelled sets of various data to analyze data from photo and video observations. The existing databases describing the biological diversity of the deep seabed are scattered and contain little structured information. Also, the problem of creating extensive databases lies in the rarity of many benthic inhabitants, as a result of which the number of images of a certain species of animals can be very small. That is why it is advisable to solve the problem of segmentation of underwater communities with the help of neural networks solving the Few Shot Learning [11] class of problems. The peculiarity of FSL tasks is that with their help it is possible to segment an object having only a few images of each species of animals and plants. We prepared small datasets containing 60 species of marine life, and also analyzed the freely available datasets.

The study of benthic communities can also consist of determining the habitat on the map and calculating the volumes of the objects under study. 3D modelling of plant and animal surface coverage from the video can help localize the main biotopes on maps [24]. Localization coverage counts can be used for further environmental studies.

This paper will present a solution to the problem of counting deep-sea sedentary inhabitants using neural network technologies.

2 Problem Definition

The work aims to create a framework that solves the problem of marking, detecting and identifying sedentary underwater inhabitants on a map. To achieve this goal, it is necessary to solve several tasks, which in turn can be divided into 3 subsets presented in Fig. 1.

The tasks of the first stage are to form a dataset. After several iterations of finding objects, processing them and selecting object types, the main database will be formed. It will serve as a basis containing several photographs of the target objects, with the help of which the segmentation will be performed at the second stage of the work.

At the stage of preprocessing and segmentation, the video sequence undergoes preprocessing, which consists of splitting into frames and improving the image quality [17]. Finding and pre-training few-shot learning neural networks on underwater objects is one of the important tasks of the second stage. This task has the greatest impact on the quality of the result of the entire system.

After segmentation of objects, the third stage of mapping the bottom and localization of segmented objects begins on the 3D model built based on the video [21]. The constructed 3D model makes it possible to estimate the volume

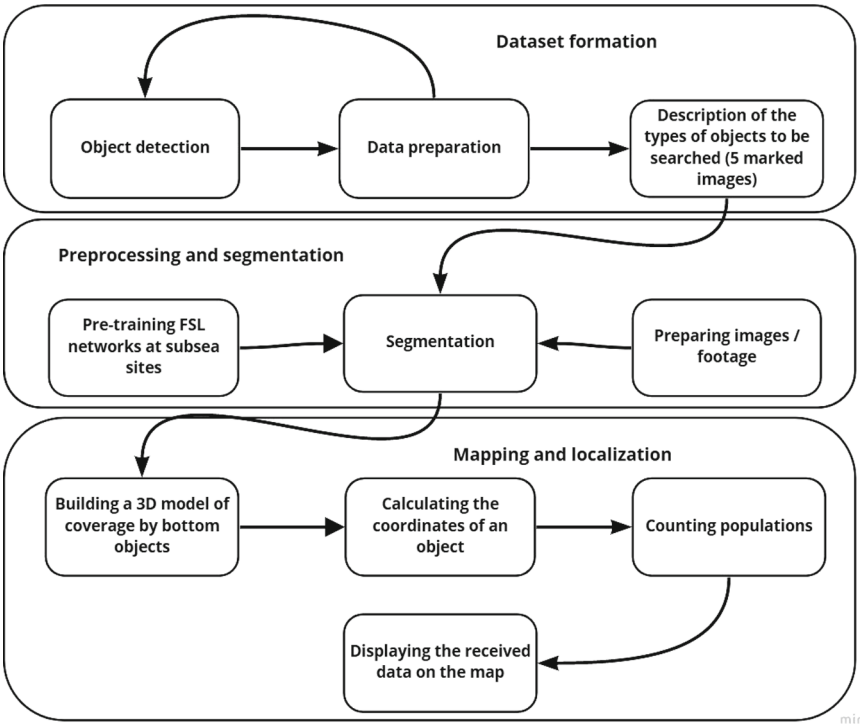

Fig. 1. Steps to complete the task of studying the seabed: dataset formation, preprocessing and segmentation, mapping and localization; and sequence of tasks.

of the object and its coordinates relative to the camera and the investigated point of the bottom [19].

3 Related Work

The article [10] described an approach for automatically detecting and segmenting underwater crabs for more accurate feeding. However, the significant disadvantages of this solution are the small depth of finding the animals, their close location to the camera. During deep-sea surveys, the terrain may not allow you to get so close to the objects of interest.

Our proposed solution is based on the creation of a cascade of neural networks, each of which is capable of solving one of the assigned tasks. The main tasks are: improving the quality of the image, segmentation and counting the coverage of surface individuals. By achieving the best result at every stage [12], it is possible to achieve the set goal. Neural networks must successfully cope with the tasks of unsupervised, semi-supervised or supervised segmentation on video or photos, as well as tasks of super-resolution for data preprocessing.

3.1 Architecture Components

One of the important criteria for choosing the algorithm components was the limitation on the size of the training dataset. Therefore, first of all, attention was paid to models capable of performing the segmentation without large-scale datasets (for example ImageNet and COCO). The model should perform the same task, but over several frames, that is, the number of training examples can be limited to five or less. An example of a model trained on FSS-1000 - a few-short segmentation database and described in the article [7,18]. The easy scalability of the FSS-1000 can also be attributed to the positive characteristics of this approach.

Convolutional neural networks efficiently handle the object segmentation task [20]. For a video stream, such problems are usually solved by separate processing of information about appearance and movement using standard 2D convolutional networks with the subsequent merging of two information sources. However, a new approach - DC-Seg, which is faster and more accurate than the masking described in the article [14], proposes to segment visible objects on video using 3D CNN. This approach has not previously been effectively used to solve computer vision problems.

Difficult terrain or objects that are too close to each other may make it difficult to segment them, and as a result, take them into account when counting. Therefore, the problem of segmentation of background or partially hidden object is also relevant for video monitoring of an underwater biotope. The semi-supervised model [25] aims to segment a specific object across the entire video sequence based on the object mask specified in the first frame. This technique can be useful for segmenting interactive objects.

Due to the nature of the underwater video, a hazy image is one of the most common obstacles to accurately masking an object. Therefore, a method was also tested, when the video sequence is divided into frames, to each of which SISR is applied [14] a generative super-resolution model for increasing the clarity of the frame.

To calculate the volume of coverage by bottom objects, it is necessary to compile a depth map for the landscape and segmented objects. Since we can only use one camera for underwater photography, we need to solve the Single-view reconstruction problem. The framework described in the article [23] solves this problem using a combination of 3 neural networks: DepthNet, PoseNet and FeatureNet for depth map prediction, egomotion prediction and feature learning respectively.

3.2 Data Preparation

Before solving the segmentation problem, it is necessary to create a set of classes of different species of animals and plants. In Fig. 1, after the object is detected, the data preparation stage begins, without which it is impossible to proceed to the segmentation stage. This is because the neural network needs small datasets with targets to be segmented in the image.

One of the most famous open databases - Fish Recognition Ground-Truth data [1], used in the Fish4Knowledge project to solve the problems of species detection and recognition [2]. However, it contains only 15 species of fish that live near the coral reef. Other datasets [3,4] contain about 500 views, but rectangles are used as object selection, in connection with which often the fins and parts of the body can be cut off.

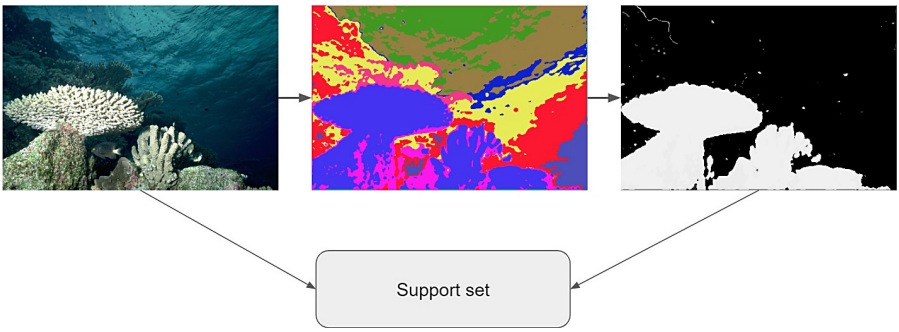

Fig. 2. Data preparation for support set: changing the resolution and size, semantic segmentation, highlighting the necessary masks.

SUIM Dataset for semantic segmentation of underwater imagery [13] could make it easier to create a dataset but contains 8 significantly larger than required classes, such as human divers or aquatic plants and sea-grass.

To solve the problem of segmentation of various marine life, a sample of 6 images for each species was created [16]. For this dataset, the resize the original image to 224 $\times$ 224 pixels and the creation of a black and white image mask of the same size was produced.

The first step in Fig. 1 is the automated creation of image masks and database preparation. With the help of a neural network that solves the semantic segmentation problem, different kinds of objects are segmented on each of the frames. After segmentation, the resulting masks belong to different classes of objects. The process of creating a dataset for its further use as a support set to generalize to unseen samples from a query set is shown in Fig. 2.

4 Implementation

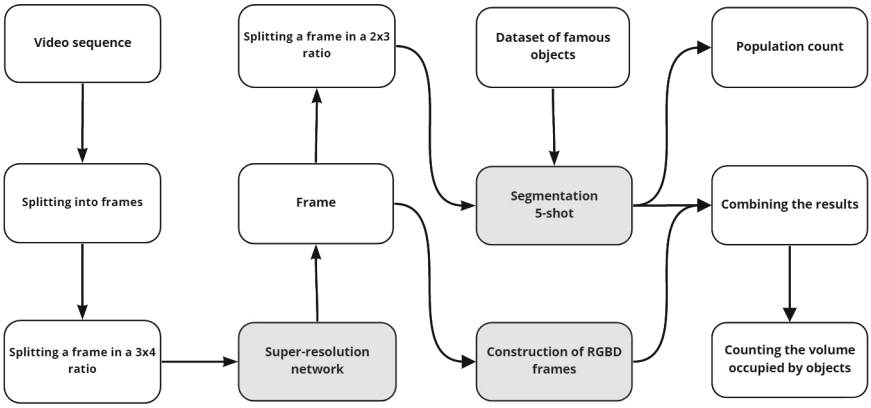

Fig. 3. Stages of video sequence processing.

Thus, at the first stage of dataset formation using the DC-Seg model, various kinds of objects are segmented on video materials, which will subsequently be marked up and formed into classes of 5 or more images (Fig. 4). At the second stage, video preprocessing takes place - splitting into frames and applying the super-resolution SISR model. With the help of a dataset prepared in advance at stage 1, the video sequence divided into frames is segmented using a 5-shot model pre-trained on the FSS-1000 dataset. The third stage begins by building a depth map for the incoming footage using the framework described above. The resulting RGBD frames are combined with previously segmented objects, which makes it possible not only to count the number of objects but also to calculate their occupied volume. The resulting solution is shown in Fig. 3.

The video sequence of underwater shooting is divided into frames, which, in turn, undergo preprocessing and, using a neural network, increase the image resolution. Segmentation of objects from the database performs on frames that have passed preprocessing. A more detailed description of the neural networks used is presented in the Table 1.

Unsupervised video segmentation task consists of segmentation of all pixels belonging to a "salient" object, while the categories of objects are not known a priori. The solution to this problem is necessary for the automatic collection of underwater databases. The 5-shot model uses elements of a previously created database as images of targets for segmentation. With SISR, each image frame is enlarged by 4 times.

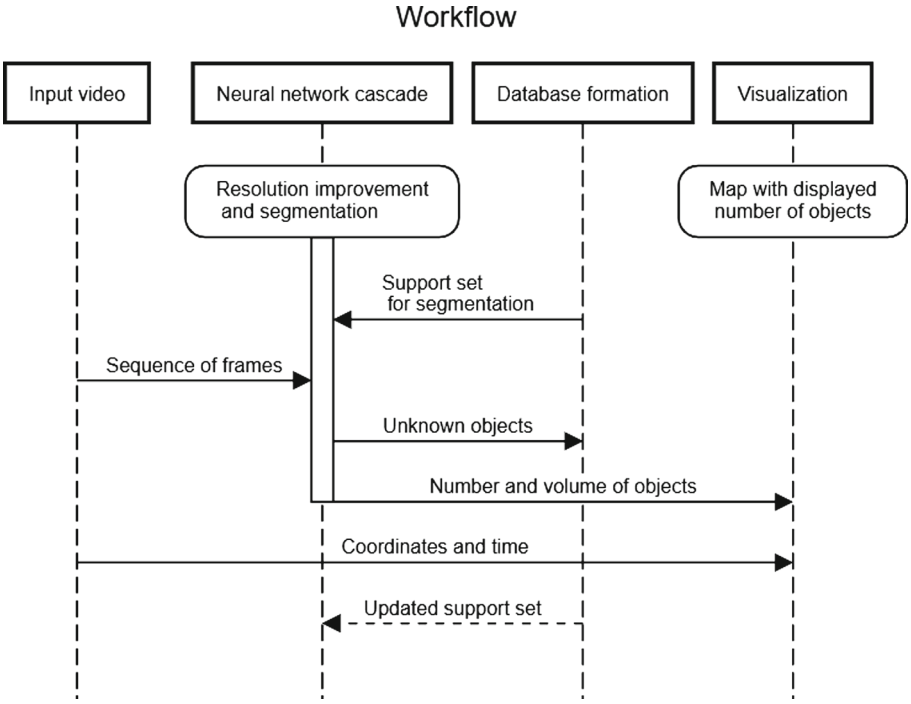

Fig. 4. Processing diagram of the incoming video sequence. The video is split into frames, each frame is processed by neural networks. Previously unknown objects are stored in the dataset. Coordinates, time and counting results are synchronized and visualized on the map.

Table 1. The main considered neural networks.

Model	Parameters	Pre-train model	Dataset	Applying segmentation
3DC-SEG	Total params: 103,244,936 Trainable params: 103,244,936	+	COCO, YouTubeVOS, DAVIS'17	Video segmentation (unsupervised task)
5-shot	Total params: 32,833,025 Trainable params: 32,833,025	+	FSS-1000	FSL segmentation
SISR	Total params: 1,944,579 Trainable params: 1,940,227	−	USR-248	Single Image Super-resolution

5 Visualization

The graphical display of results is generated using the open-source large-scale geolocation data sets [8] visualization application -Kepler.gl [5]. To build a relief on a world map you need to get high-rise models. We used global datasets [6] elaborated by NASA and NGA hosted on Amazon S3 and SRTM 90m Digital Elevation Database v4.1 elaborated by CGIAR-CSI. Using open libraries based on GDAL, the raster file is converted into a format suitable for Kepler.Gl.

Data derived from underwater survey videos, such as depth, latitude, longitude, surface area and timestamp, are built on top of global terrain digital models. Also, when rendered, the data is transformed from the dataset's own coordinate reference system to CRS84. The creation process is shown in Fig. 5.

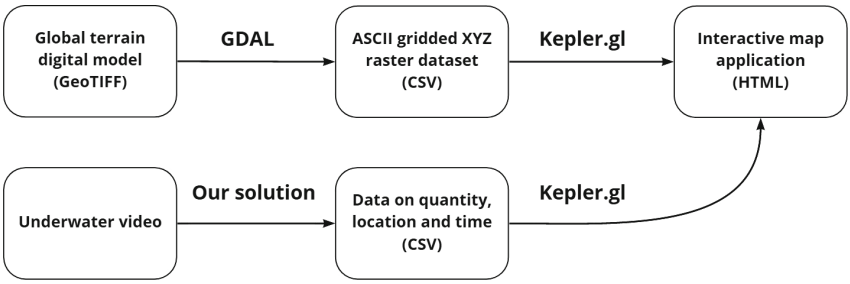

Fig. 5. The process of visualization of the depth map and the results obtained.

6 Results

One of the main problems we faced was the lack of underwater datasets for the segmentation of a large number of different objects. The solution to this problem has become part of the architecture of our system, with the help of which it is possible to form the necessary dataset from the original video sequence.

The lack of a large number of computational resources and the limited data contributed to the choice in favour of pre-trained models or models that do not require long training. However, the weights of the super-resolution model were obtained independently and published in the public domain.

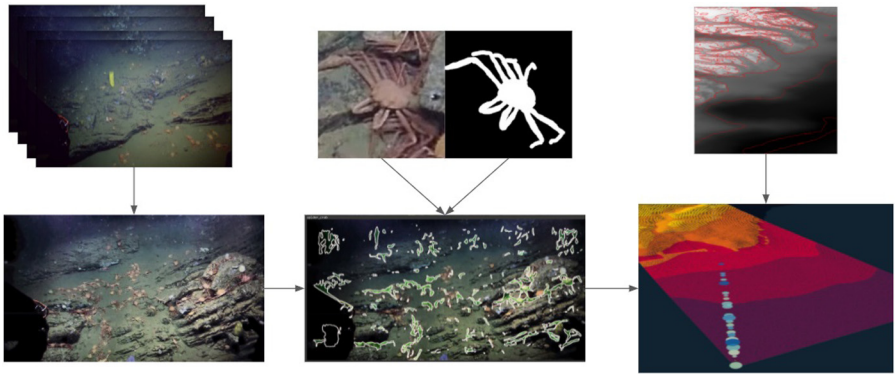

Fig. 6. Flowchart showing the results of each stage.

Today, there are a large number of solutions to the semantic segmentation problem, but underwater videos have a number of features, such as turbidity of the water, lack of light, partial visibility, etc. This imposes additional restrictions and makes it difficult to choose the architecture of the model.

Table 2. Video sequence processing time.

Method	Video	SISR	5-shot	Total time	Result
Our solution	144 s (29,97 fps)	1230 s (14 fps)	2723 s	3953 s	csv + interactive map
	1 frame	40–50 s	9–10 s	50–60 s	
Human based approach	1 frame			∼60 s	Number of objects

The resulting solution is shown in Fig. 6. At each stage of the solution, the most modern methods and models are used. The neural networks and tools used are able to solve the assigned tasks quickly. The Table 2 shows the approximate operating time of the obtained automatic system in comparison with classical methods of bottom researches.

7 Conclusion

With the help of the presented architecture, it is possible to successfully solve the problem of marking and detecting sedentary underwater inhabitants.

New algorithms for image enhancement and segmentation of underwater objects can be used to automate monitoring and population counting. Compared to traditional methods of studying marine life, an automated approach can significantly improve the quality of data and reduce the number of wasted resources.

The resulting solution can be used as a basis for creating extensive databases for the subsequent training of static segmentation models.

Acknowledgments. The authors would like to acknowledge the Reviewers for the valuable recommendations that helped in the improvement of this paper.

References

1. Fish Recognition Ground-Truth data. http://groups.inf.ed.ac.uk/f4k/groundtruth/recog. Accessed 20 Mar 2021
2. Fish Species Recognition. http://www.perceivelab.com/datasets. Accessed 20 Mar 2021

3. Ozfish. https://aims.github.io/ozfish. Accessed 21 Mar 2021
4. Fish Dataset. https://wiki.qut.edu.au/display/raq/Fish+Dataset. Accessed 22 Mar 2021
5. Kepler.gl. https://github.com/keplergl/kepler.gl. Accessed 22 Mar 2021
6. LP DAAC - SRTMGL1. https://lpdaac.usgs.gov/products/srtmgl1v003. Accessed 22 May 2021
7. Azad, R., Fayjie, A.R., Kauffmann, C., Ben Ayed, I., Pedersoli, M., Dolz, J.: On the texture bias for few-shot CNN segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2674–2683 (2021)
8. Bakiev, M., Khasanov, K.: Comparison of digital elevation models for determining the area and volume of the water reservoir. Int. J. Geoinform. **17**(1), 37–45 (2021)
9. Benjamin, J., et al.: Aboriginal artefacts on the continental shelf reveal ancient drowned cultural landscapes in northwest Australia. PLoS ONE **15**(7), e0233912 (2020)
10. Cao, S., Zhao, D., Sun, Y., Liu, X., Ruan, C.: Automatic coarse-to-fine joint detection and segmentation of underwater non-structural live crabs for precise feeding. Comput. Electron. Agric. **180**, 105905 (2021)
11. Dong, N., Xing, E.P.: Few-shot semantic segmentation with prototype learning. In: BMVC, vol. 3 (2018)
12. Ghorbani, M.A., Deo, R.C., Kim, S., Hasanpour Kashani, M., Karimi, V., Izadkhah, M.: Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia. Soft Comput. **24**(16), 12079–12090 (2020). https://doi.org/10.1007/s00500-019-04648-2
13. Islam, M.J., et al.: Semantic segmentation of underwater imagery: dataset and benchmark. arXiv preprint arXiv:2004.01241 (2020)
14. Islam, M.J., Enan, S.S., Luo, P., Sattar, J.: Underwater image super-resolution using deep residual multipliers. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 900–906. IEEE (2020)
15. Jian, M., Liu, X., Luo, H., Lu, X., Yu, H., Dong, J.: Underwater image processing and analysis: a review. Sig. Process. Image Commun., 116088 (2020)
16. Jung, A.B., et al.: imgaug (2020). https://github.com/aleju/imgaug. Accessed 1 Feb 2020
17. Li, C., Anwar, S., Porikli, F.: Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recogn. **98**, 107038 (2020)
18. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: FSS-1000: a 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2869–2878 (2020)
19. Liu, S., Yu, J., Ke, Z., Dai, F., Chen, Y.: Aerial-ground collaborative 3D reconstruction for fast pile volume estimation with unexplored surroundings. Int. J. Adv. Robot. Syst. **17**(2), 1729881420919948 (2020)
20. Miao, J., Wei, Y., Yang, Y.: Memory aggregation networks for efficient interactive video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10366–10375 (2020)
21. Nocerino, E., Menna, F., Chemisky, B., Drap, P.: 3D sequential image mosaicing for underwater navigation and mapping. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **43**, 991–998 (2020)
22. Roach, T.N., et al.: A field primer for monitoring benthic ecosystems using structure-from-motion photogrammetry. JoVE (J. Vis. Exp.) **170**, e61815 (2021)

23. Shu, C., Yu, K., Duan, Z., Yang, K.: Feature-metric loss for self-supervised learning of depth and egomotion. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12364, pp. 572–588. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_34
24. Urbina-Barreto, I., et al.: Quantifying the shelter capacity of coral reefs using photogrammetric 3D modeling: from colonies to reefscapes. Ecol. Ind. **121**, 107151 (2021)
25. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 332–348. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_20