DOI 10.31256/UKRAS19.12

Underwater Scene Segmentation by Deep Neural Network

Yang Zhou, Jiangtao Wang, Baihua Li, Qinggang Meng, Emanuele Rocco and Andrea Saiani

***Abstract*— A deep neural network architecture is proposed in this paper for underwater scene semantic segmentation. The architecture consists of encoder and decoder networks. Pre-trained VGG-16 network is used as a feature extractor, while the decoder learns to expand the lower resolution feature maps. The network applies max un-pooling operator to avoid large number of learnable parameters, and, in order to make use of the feature maps in encoder network, it concatenates the feature maps with decoder and encoder for lower resolution feature maps. Our architecture shows capabilities of faster convergence and better accuracy. To get a clear view of underwater scene, an underwater enhancement neural network architecture is described in this paper and applied for training. It speeds up the training process and convergence rate in training.**

I. Introduction

The scene understanding of the underwater environment is an appealing topic among marine researchers and the public too, as underwater and especially undersea domains highly capture its attention. Many applications benefit from underwater scene information such as seafloor survey and marine object detection[1]. Conventional methods for underwater scene understanding fall into multi-sensor data fusion. Castellani et al.[2] proposed to reconstruct 3D underwater environment with the aid of multiple acoustic views given by underwater acoustic sensors, but the trade-off between speed and accuracy limits this method for the real-time use. Moroni et al[1] instead proposed to use both acoustic and stereo camera sensors, but the additional data fusion process for mapping has to be carefully considered. Moreover, it could be difficult to calibrate a stereo camera in underwater environment because the refractive effects lead to non-linear distortion effects that depend on the seawater density and incidence light rate. Furthermore, depth image only cannot provide the straightforward information for object recognition task in the current camera view. To achieve the object recognition task without considering depth information, an alternative approach is to use image semantic segmentation based on monocular camera.

Image semantic segmentation is one of the key fields in computer vision, to which deep learning has been giving many contributions during the past three years[3]. It is successfully used for indoor scene segmentation and outdoor scene segmentation[4]. The development of semantic segmentation benefits an increasing number of applications including autonomous driving, human-computer interaction and augmented reality, to name a few[3]. Compared with conventional semantic segmentation methods such as Markov Random Field (MRF), Conditional Random Field (CFR)[5] and SVM [6] as classifiers, deep neural network can achieve higher accuracy with the ability of learning from high level representations Moreover, deep neural networks enable end-to-end image semantic segmentation with simpler procedures.

Autoencoder is a popular network structure for image semantic segmentation in deep neural network application field [3]. The encoder part is a convolutional neural network (CNN) for generating the feature maps by applying pooling operator. On the other hand, the decoder network is a reverse convolutional neural network based on the un-pooling operator.

As in the successful cases of image semantic segmentation on indoor scene and road scene applications[4], our work focuses on end-to-end underwater scene semantic segmentation by using deep neural network with monocular camera only. In this paper, a network structure is proposed for underwater scene segmentation, which can be used for real-time inference. The neural network architecture proposed in this paper enhances the generalization ability compared with using SegNet[4] when applying to real-time videos and needs less memory during training process compared with U-Net[7]. The underwater data used in this paper are collected by Witted Srl, Italy.

Furthermore, considering the color distortion and underwater optical effects on underwater images, we apply Generative Adversarial Network (GAN) for underwater image enhancement to transform the original bluish images into surface-like ones. This approach aids to speed up the training process as it highlights the boundaries of the objects.

This paper is organized as follow. Sec. 2 describes the deep neural network structure we proposed. Sec. 3 illustrates the experiment of training on underwater images and comparisons with other network architectures. To get a better view of underwater environment for training, Sec. 4 shows the methods we used for underwater image enhancement. Finally, we conclude in Sec. 5.

II. Underwater Scene Understanding

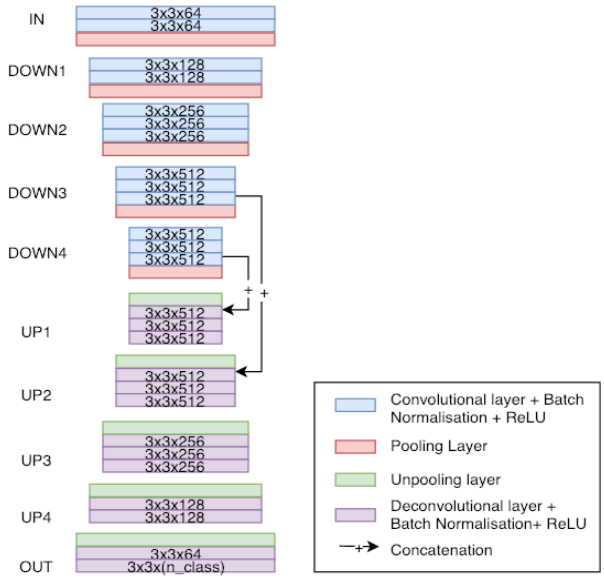
Fully Convolutional Network (FCN) [8] was the first work for pixel-wise semantic segmentation enabled by deep neural network. Instead of using decoder process, FCN applied *backwards convolution* (known as *deconvolution*) to connect coarse output with dense pixels [8]. U-Net [7] introduced instead the decoder network to expand the feature maps for medical image segmentation, adopting the idea of

Yang Zhou is with Loughborough University, Loughborough, The United Kingdom, LE11 3TU (e-mail: Y.Zhou5@lboro.ac.uk).
Jiangtao Wang is with Loughborough University, Loughborough, The United Kingdom, LE11 3TU (e-mail: J.Wang4-18@student.lboro.ac.uk).
Baihua Li is with Loughborough University, Loughborough, The United Kingdom, LE11 3TU (e-mail: B.Li@lboro.ac.uk).
Qinggang Meng is with Loughborough University, Loughborough, The United Kingdom, LE11 3TU (e-mail: Q.Meng@lboro.ac.uk).
Emanuele Rocco is with Witted Srl, Piazza Manifattura, 1 38068 Rovereto (Tn) Italy (e-mail: emanuele@witted.it).
Andrea Saiani is with Witted Srl, Piazza Manifattura, 1 38068 Rovereto (Tn) Italy (e-mail: andrea@witted.it).

concatenating the feature maps from encoder to decoder. However, the number of parameters in the deconvolution operator and of gradients in the back propagation, both generated by the feature map concatenation, slows down the training process. To reduce the number of parameters, SegNet [4] applies a new up-sampling method without learnable filters when expanding the feature maps, and removes the concatenation to reduce the gradients to be calculated during the back propagation.

A. *Network architecture*

Typical neural networks for classification task such as LeNet [9] and AlexNet [10] take fixed-sized inputs because of the requirement of the fully-connected layers. However, both U-Net [7] and SegNet [4] remove these layers so that the input can be of any size. To this extent, the architecture we propose for underwater scene segmentation removes the fully-connected layers and consist of an encoder network followed by a decoder network. This architecture is showed in figure 1.



Network architecture for underwater scene segmentation.

Our encoder network includes 13 convolutional layers from VGG-16 [11], pre-trained on the ImageNet dataset, where each layer is followed by a batch normalization layer[12] and a ReLU layer [13] to speed up the training process. These layers are grouped in 5 blocks, each ending with a max-pooling layer to reduce the size of feature maps. The decoder network is the mirror reverse structure of our encoder network, being made of 5 blocks each with one un-pooling layer followed by deconvolutional layers.

B. *Un-pooling layer*

Current research [4][3] has shown that, instead of using learnable filters with large number of learnable parameters, un-pooling operators without learnable parameters can achieve similar performances. Our architecture follows this approach, and, it applies the max un-pooling method which records the max value indices of the pooling layers to the decoder network. The corresponding un-pooling layers use these indices to define the output position for their inputs, while leaving zero to the undefined positions. With this approach, the architecture keeps the spatial information of the feature maps.

C. *Concatenation*

As from U-Net [7], in our architecture shown in Figure 1, we use the concatenation operator to transfer the feature map from the encoder to the decoder network. Furthermore, to reduce the size of the needed memory during training, we keep such concatenations for final two blocks of the encoder network, responsible for generating the dense feature maps.

Concatenation helps preserves the information learned from encoder, and lets the decoder directly learn from the feature maps. For deep neural network, the concatenation operator also alleviates the gradient vanishing problem [14].

III. EXPERIMENT

We use PyTorch to implement our architecture. The training dataset were selected from the frames of two underwater videos recorded by Witted Srl; 70 images were manually labeled with 4 categories: seagrass, rocks, sand and seawater. The images were at 1920×1080 resolution, resized at 640×360 for training to reduce the number of parameters. Over the training dataset, the numbers of pixels in each category are as follow: seagrass of 6.5M, rocks of 53.2M, sand of 22.2M and seawater of 63M. As the seagrass and sand mostly appear in the central of the view, we use augmentation methods to increase pixels numbers for sand and seagrass and the dataset size. After augmentation, the number of training data increases to 140 images by randomly rotating, and cropping. As for testing, we applied the model to those two raw underwater videos which include the 70 frames for training. These two raw underwater videos consist of thousands of frames, but some of the frames are new to the model.

During training, to increase speed, we do not alter the pre-trained weights of the VGG-16 pretrained model used as a feature extractor and we initialize the decoder parameters as describe in [14]. The network is trained by stochastic gradient descent (SGD) algorithm with learning rate as 0.01 and momentum as 0.9. The learning rate schedule is based on step decay with 0.1 decay rate for every 100 epochs.

To make full use of dataset, we use cross-validation methods to train the model: the whole dataset of 140 images are divided into 28 segments by batch size as 5. Then, only 1 segment is selected as validation data while another 27 segments are used for training. Each epoch, the validation segment is sequentially selected to statistically balance the whole dataset. That means every segment has the same probability to be elected as validation segment. As the task for semantic segmentation is a classification problem, each pixel in label image is a one-hot vector and cross-entropy loss is chosen. We calculate the overall accuracy as the number of correctly predicted pixels over the total pixel number for each epoch.

Table I illustrates the loss and overall accuracy on training images during training process and shows that our network makes the training process convergent faster than U-Net and SegNet. The training results show also that our network outperforms U-Net with a large gap of loss and accuracy and is slightly better than SegNet

TABLE I. TRAINING PROCESS

• Network	• Training (%)					
	100 epochs		*200 epochs*		*300 epochs*	
	Loss	*Acc*[a]	*Loss*	*Acc*	*Loss*	*Acc*
U-Net	0.215	0.930	0.080	0.971	0.078	0.972
SegNet	0.116	0.957	0.055	0.979	0.053	0.980
Ours	**0.100**	**0.961**	**0.045**	**0.983**	**0.036**	**0.986**

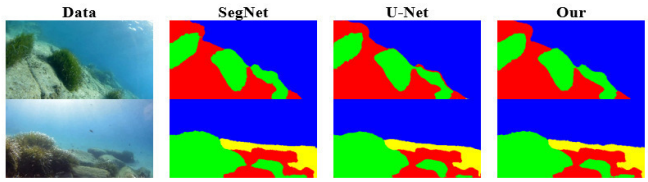
a. Training accuracy (%)

For validation process, we used the metrics of mean accuracy and mean intersection over union (mIoU). The mean accuracy is the mean of predictive accuracy over all classes in the dataset which is slightly different from overall accuracy because it considers the balance among the accuracies between different classes. Mean intersection over union (mIoU) is the metric used in [15] and penalizes the fault predictions. Table II presents the results of validation after 300 training epochs. It shows that our network outperforms U-Net and is slightly better than SegNet too. However, for memory use, SegNet is more efficient than ours when training, as the concatenation operators of our network require more memory during back-propagation, although our network shares the same number of trainable parameters with SegNet.

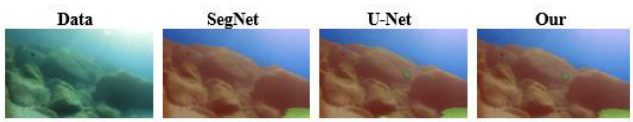
TABLE II. VALIDATION PROCESS

• Network	• Validation				
	Parameters	*Memory used*	*Accuracy (%)*	*Mean Accuracy (%)*	*mIoU (%)*
U-Net	16.08M	8G	0.976	0.718	0.670
SegNet	**14.7M**	**4G**	0.985	0.742	0.705
Ours	**14.7M**	7G	**0.994**	**0.744**	**0.707**

Figure 2 shows the example results predicted from these three networks. It shows that the three networks all work well with on the dataset.



Predictions on different network architectures



Real-time video frame testing

For real-time video testing, our model can achieve nearly 25 frames per second, which is close to the standard real-time frame rate. The results are shown in figure 3, our work and U-Net have a better generalization ability to recognizing unseen scenes on new videos than SegNet[4]. E.g. The seagrass in the centre can be recognized by U-Net and our architecture. This improved performance is given by the concatenation operation in U-Net and in the last two blocks in encoder network of our architecture.

IV. UNDERWATER IMAGE ENHANCEMENT

The undersea images are of a blue green tinge, mostly blurry and unclear because of the light absorption in water and diffusion due to suspended particles [16]. Moreover, the color distortion and blur effects change during seasons. In this situation, the visual model trained with raw images may not perform well. Hence, we consider an image enhancement process to standardize all images in a clear view.

The algorithms for underwater image enhancement can be classified into two categories: physics-based technique and deep learning technique. For example, the work of Luz [16] applies an energy minimization formulation using a Markov Random Field. Deep learning models instead, such as WaterGAN [17], UGAN and UGAN-P [18] used Generative Adversarial Network (GAN) to enhance the underwater images.

A. *Method*

We use GAN architecture as well, which consists of a generator network and a discriminator network. The U-Net structure is used for the generator network and it is responsible for learning the image style by matching blurry images to clear ones. The discriminator network uses instead the same module described in PatchGAN [19] with four convolution layers and calculates the loss from enhanced images and clear images. During inference, the generator network predicts clear images from blurry images as input.

B. *Dataset*

The training of the enhancing GAN requires training data of paired clear and unclear images. Such pairs are collected by using clear images from ImageNet [20] and using unclear synthetic images generated by the UGAN of [18] from the clear ones.

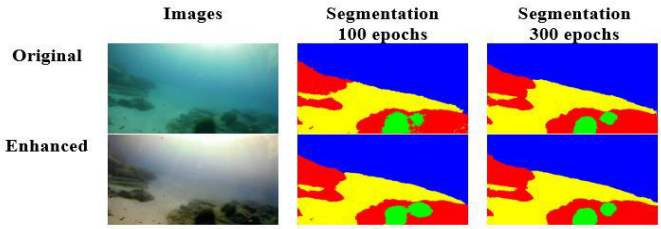
C. *Result*

Underwater image enhancement result (left column shows original images, righ column shows enhanced iamges)

After training, the enhancing GAN architecture is verified with our underwater image dataset. A sample result is showed in figure 4, where the left image shows the original frame before processing, while the right column shows the enhanced one. Not only the enhanced image is not bluish anymore, but also the scene details like edges of stone, sand and sea grass are better defined. However, the color of sand and stones in enhanced image is more yellowish.

D. *Segmentation experiment with enhanced images*

To verify if the image enhancing method helps the training process of our scene segmentation architecture, we separately train our network with two datasets: one model with the original underwater images; the second one with the enhanced ones. Figure 5 shows sample results from such dual training processes, while the recorded loss and overall accuracy are showed in Table III. The model trained on the enhanced image dataset is more accurate at 100 epochs than the original one with a faster convergence rate.



Compared image segmentation result with enhanced iamges

TABLE III. ENHANCEMENT

• Dataset	• Training (%)					
	50 epochs		*100 epochs*		*200 epochs*	
	Loss	*Acc*[a]	*Loss*	*Acc*	*Loss*	*Acc*
Original	0.211	0.925	0.148	0.929	**0.041**	**0.984**
Dehazed	**0.185**	**0.932**	**0.064**	**0.976**	0.043	**0.984**

a. Training accuracy (%)

In conclusion, GAN architecture shows potentials for underwater image enhancement from blurred saturation, recovering the images into ground-like images and helping to convergence of the segmentation training process.

V. CONCLUSION

This paper shows deep neural networks can be effective for underwater semantic segmentation and underwater image enhancement. Our proposed segmentation network achieves better performances according to different metrics than U-Net [7] and is slightly better than SegNet [4] too. The tested GAN architecture for image enhancement is showed to help the training convergence rate of our segmentation architecture. In future our segmentation method could be extended to used jointly with depth information for further underwater vision tasks.

ACKNOWLEDGMENT

The authors are grateful to the EPSRC Centre for Doctoral Training in Embedded Intelligence under grant reference EP/L014998/1 for financial support sponsored by Witted Srl, Italy.

REFERENCES

[1] D. Moroni, M. A. Pascali, M. Reggiannini, and O. Salvetti, "Underwater scene understanding by optical and acoustic data integration," in *Proceedings of Meetings on Acoustics*, 2013, vol. 17, no. 1, p. 070085.

[2] U. Castellani, A. Fusiello, and V. Murino, "Registration of multiple acoustic range views for underwater scene reconstruction," *Comput. Vis. Image Underst.*, vol. 87, no. 1–3, pp. 78–89, 2002.

[3] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation," 2017.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[5] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *Journal of Visual Communication and Image Representation*, vol. 34. Academic Press, pp. 12–27, 01-Jan-2016.

[6] M. Thoma, "A Survey of Semantic Segmentation," 2016.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9351, pp. 234–241.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3431–3440, 2015.

[9] Y. Le Cun *et al.*, "Handwritten Digit Recognition with a Back-Propagation Network," *Dermatologic SurgeryAdvances Neural Inf. Process. Syst. 2 (NIPS 1989)*, 1990.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks Alex," *Proceeding NIPS'12 Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, pp. 1097–1105, 2012.

[11] K. Simonyan and A. Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," *CORR*, 2014.

[12] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015.

[13] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Network," *Proc. 14th Int. Con- ference Artif. Intell. Stat.*, vol. 15, 2011.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

[15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge – a Retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[16] L. A. Torres-Méndez and G. Dudek, "Color Correction of Underwater Images for Aquatic Robot Inspection," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2005, pp. 60–73.

[17] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "WaterGAN: Unsupervised Generative Network to Enable Real-time Color Correction of Monocular Underwater Images," 2017.

[18] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing Underwater Imagery using Generative Adversarial Networks," 2018.

[19] A. Radhakrishnan, C. Durham, A. Soylemezoglu, and C. Uhler, "Patchnet: Interpretable Neural Networks for Image Classification," 2017.

[20] D. Jia, D. Wei, S. Richard, L. Li-Jia, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun-2009.