J. Ocean Univ. China (Oceanic and Coastal Sea Research)
https://doi.org/10.1007/s11802-019-3858-x
ISSN 1672-5182, 2019 18 (2): 376-382
http://www.ouc.edu.cn/xbywb/
E-mail:xbywb@ouc.edu.cn

Underwater Object Recognition Based on Deep Encoding-Decoding Network

WANG Xinhua[1), 2)], OUYANG Jihong[1), *], LI Dayu[2)], and ZHANG Guang[2)]

1) *College of Computer Science and Technology*, *Jilin University*, *Jilin* 130012, *China*
2) *State Key Laboratory of Applied Optics*, *Changchun Institute of Optics*, *Fine Mechanics and Physics*, *Chinese Academy of Sciences*, *Jilin* 130033, *China*

(Received April 1, 2018; revised June 12, 2018; accepted June 26, 2018)

Abstract Ocean underwater exploration is a part of oceanography that investigates the physical and biological conditions for scientific and commercial purposes. And video technology plays an important role and is extensively applied for underwater environment observation. Different from the conventional methods, video technology explores the underwater ecosystem continuously and non-invasively. However, due to the scattering and attenuation of light transport in the water, complex noise distribution and low-light condition cause challenges for underwater video applications including object detection and recognition. In this paper, we propose a new deep encoding-decoding convolutional architecture for underwater object recognition. It uses the deep encoding-decoding network for extracting the discriminative features from the noisy low-light underwater images. To create the deconvolutional layers for classification, we apply the deconvolution kernel with a matched feature map, instead of full connection, to solve the problem of dimension disaster and low accuracy. Moreover, we introduce data augmentation and transfer learning technologies to solve the problem of data starvation. For experiments, we investigated the public datasets with our proposed method and the state-of-the-art methods. The results show that our work achieves significant accuracy. This work provides new underwater technologies applied for ocean exploration.

Key words deep learning; transfer learning; encoding-decoding; underwater object; object recognition

1 Introduction

Earlier, sonar technology was well known for underwater object detection. However, it fails to meet the current requirement of high-precision underwater tasks due to its low-resolution imaging problem. With the continual increase in underwater video applications, a large amount of high-resolution underwater videos can be obtained, thus leading to an innovation in the generation from coarse object detection to grain object recognition. Video technology plays an important role and is extensively applied for underwater environment observation. It provides a continuous and non-invasive method for investigating the underwater ecosystem and life. In addition, underwater video technology is vital in some automatic discovery tasks including searching and rescuing underwater robots. Moreover, video technology has led to the revolution of some conventional exploration methods and research fields, such as the statistics of fisheries (Cappo *et al.*, 2006), submarine geology (Bonin-Font *et al.*, 2015), and marine biology observation (Struthers *et al.*, 2015). However, different underwater scenes have different properties, thus underwater technology inevitably encounters many great challenges, some of which are as follows. First, the imaging quality is mainly enslaved to the light scattering, absorption and color distortion of the underwater environment. Second, the energy loss during light spreading weakens the light intensity. Third, suspended colloid and granular impurities in the water produce complex noise into images. Furthermore, because of the unrestricted environment for such moving objects as fish, determining their size and pose from the underwater video and image is challenging. To summarize, these difficulties pose new challenges for underwater video and image applications including object detection and recognition. Pixel to pixel networks are proposed to solve the image restoration and denoising problem (Mao *et al*., 2016). In the meantime, a deconvolution network is suggested for semantic segmentation and achieving remarkable performance (Noh *et al*., 2015). The deconvolution operation can be used as a decoding procedure of the convolution operation. The difficulty of object recognition from the underwater images is the noise distribution. Based on the pixel-to-pixel image denoising studies, the deconvolution operation removes most of the noise information from the image. In this paper, we propose a new deep encoding-decoding convolution architecture for alleviating the im-

* Corresponding author. E-mail: ouyj@jlu.edu.cn

pact of noise for recognition. Compared with the conventional handcraft features including HOG and SIFT, the deep convolutional features provide higher robustness for the recognition task in the underwater environment. Fig.1 shows the proposed deep encode-decode convolution network (ED-Net) and learning workflow.

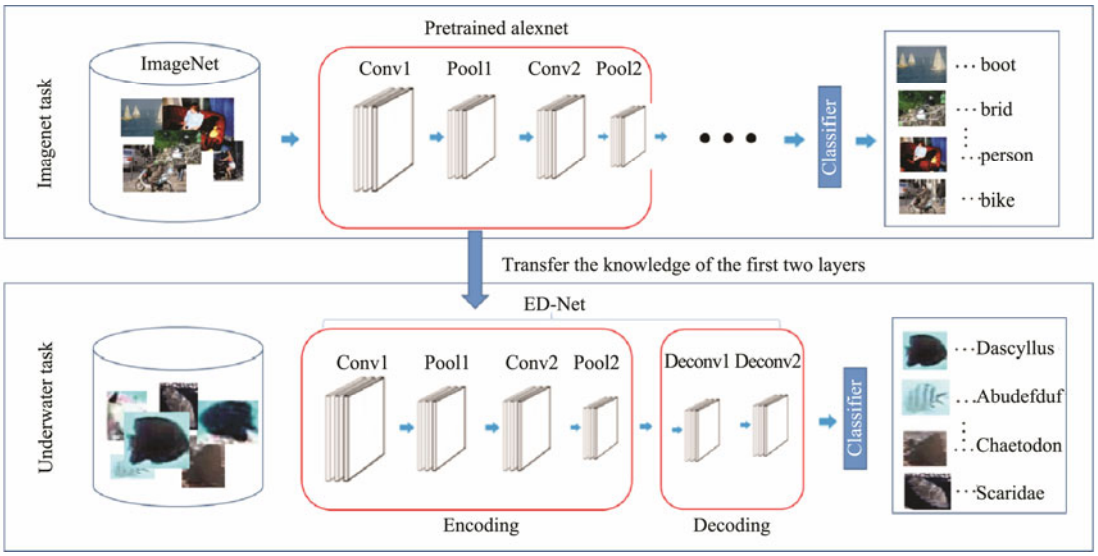

Fig.1 Proposed underwater object recognition framework.

Our contributions are as follows:

1) We proposed a new deep learning architecture with convolutional encoding and deconvolutional decoding for underwater object recognition and proved the effectiveness of our model through visualizing the feature maps.

2) We applied the deconvolution kernel with a matched feature map to solve the problem of dimension disaster and low accuracy.

3) We used data augmentation and transfer learning to solve the problem of data starvation in deep learning.

4) Our proposed network model achieved high accuracy in public underwater datasets.

2 Related Work

2.1 Object Recognition in an Underwater Environment

Marine biologists apply the videos captured by underwater monitor and robots in making the marine ecosystems analysis (Pelletier *et al.*, 2011). The research of underwater vision technology promotes the development of underwater object recognition. Lines *et al*. (2001) proposed underwater fish estimation method based on image analysis, but it only works in special underwater conditions. Spampinato *et al*. (2008) proposed an underwater vision system that facilitated fish detection, recognition, tracking, and counting. In addition, they proposed a fish automatic classification algorithm (Spampinato *et al.*, 2010) in their later research that assisted in marine biologist analysis and understanding fish behavior. Sun *et al*. (2018) introduced a deep learning method for object recognition in low-quality underwater videos and achieved remarkable results. Boom *et al*. (2014) provided continuous underwater videos for underwater object recognition, which is an important tool used by marine biologists for analyzing the submarine ecological environment.

2.2 Convolutional Neural Network

Deep learning has been successfully applied in computer vision. A convolutional neural network (CNN) is the most sutitable network among various networks. In the previous work, because of the limitation of datasets and hardware, training the model with high performance is challenging. The model may not be well fitted and may be overfitting. With the development of GPUs and the approach of big data era, the research of CNNs has garnered significant success, such as AlexNet (Krizhevsky *et al.*, 2012) and VGGnet (Simonyan and Zisserman, 2014). Compared to traditional artificial feature descriptors (*e.g.*, LBP, SIFT, *etc*.), CNN extracts features directly, not intervening manually (Zeiler and Fergus, 2014). We found that the first few layers mainly extracted the low-level features of objects' side and angle by visualizing different layers' feature maps. With a deeper network, the extracting features of the layers are more comprehensive. Therefore, CNNs extract richer features than conventional feature descriptors. CNNs are applied to objects detection (Zhang *et al.*, 2014), object recognition (Simonyan and Zisserman, 2014) and natural language processing (NLP) (Kim, 2014), and have achieved the state-of-art performance. Because of its excellent performance in abstract feature extraction, we used the CNNs as the basic architecture in our study.

3 Proposed Methods

We proposed a novel network structure that was applied in underwater object recognition. Fig.2 shows the overall architecture. In addition, we implemented the training method, including transfer learning and data augmentation.

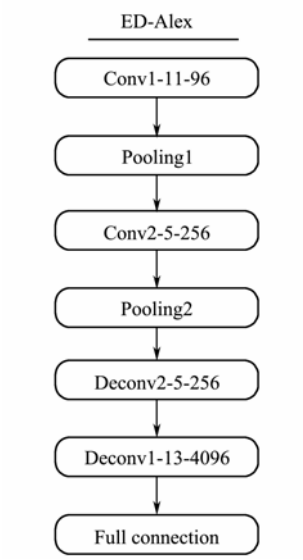

Fig.2 Parameter configuration of ED-Net. Considering conv1-11-96 as an example, conv1 represents the first convolutional layer, 11 indicates the kernel size is 11×11, and 96 denotes the number of output feature maps.

3.1 The Architecture

In 2015, the deconvolution network was first proposed for image segmentation (Noh *et al.*, 2015). First, images are transformed into deep features by the convolution layers, and then the deconvolution layers are used as decoders for refining the images. We try to employ such convolution-deconvolution network for classification. We used the deconvolution layers for feature fine-grain extraction.

Fig.2 shows the proposed convolution and de-convolution network that has symmetrical parameters. The network contains two convolution layers, two de-convolution layers, and pooling layers. In addition, we set the parameter of convolution layers to be the same as that of the AlexNet, which is easy for initializing the network by parameter transferring. The parameters of deconv2 are the same as that of conv2, and the parameters of deconv1 are set according to the feature map size. This network model continuously refined the object features by convolutional encoding and de-convolutional decoding. Moreover, we applied CNN feature descriptors fusion to classification, which will be shown in the following section.

3.1.1 Convolution and feature activation

A cascaded CNN contains many convolutional filters that extract richer feature representations from images. In addition, cascaded convolutional filters fuse an object's local and global features. We applied the active operation following convolutional layers. To formulate:

$$f(x) = \mathrm{Re}LU(\theta x + b), \tag{1}$$

where θ represents the weight matrix of the model; x, input feature map; and b, bias. As shown in Fig.3(a), the convolution is a multi-to-one filter operation, and the convolution layers cascaded assist in extracting the feature representation from local to global.

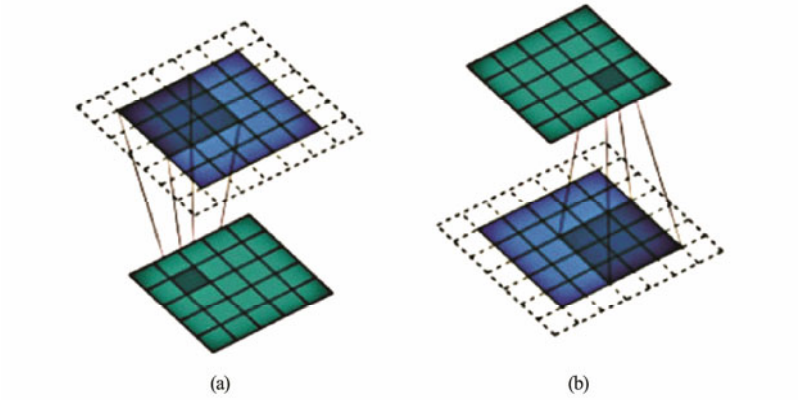

Fig.3 (a) Multi-to-one convolution; (b) one-to-multi deconvolution. The dotted line represents padding in the computation.

3.1.2 Pooling

Pooling obtains new features by downsampling the feature maps. In our model, we applied pooling after the convolution layers, which renders higher robustness. The main advantages of pooling are twofold. First, it reduces the parameters. As the convolution and de-convolution networks have many parameters, computing is challenging for the hardware. With pooling operation, the parameters are reduced and over-fitting is prevented. Second, it adds context information to a network. After convolution computing, the pooling layer fuses the features. Because of the pooling layers, the extracting features experience space invariance.

3.1.3 Deconvolution kernel

Deconvolution is generally applied to image denoising (Mao *et al.*, 2016), image segmentation (Noh *et al.*, 2015), and visualization of deep networks (Zeiler *et al.*, 2011). The purpose of deconvolution to underwater object detection is that applying de-convolution refines the features after convolution for classifying an object. As shown in

Fig.2(b), deconvolution is the opposite process of convolution; thus, the model reverts more feature details through the one-to-multi relationship. In the convolution and deconvolution symmetry model, the input image retains the two-dimensional feature map after deconvolution. Moreover, the dimension of feature map is very high after deconvolution. For example, the output of deconv1 is 227×227×x (x>1), which can bring in dimension disaster and lower accuracy. Therefore, we proposed using the deconvolution kernel with the matched feature map for extracting linear features.

According to the principle of convolution and deconvolution computation, we summarize the relationship of network propagating as follows:

$$output_w = \left\lfloor \frac{input_w + 2pad - \ker nel_{\text{size}}}{stride} \right\rfloor + 1, \quad (2)$$

$$output_h = \left\lfloor \frac{input_h + 2pad - \ker nel_{\text{size}}}{stride} \right\rfloor + 1, \quad (3)$$

where $input_w$, $input_h$, $output_w$, and $output_h$ represent the width and the height of the input feature maps and those of the output ones, respectively. And ker nel_{size} is the size of convolution kernel, and *pad*, the padding we add to feature maps.

In this network, we did not use the full connection layer to pull the high-dimensional feature maps into one-dimensional vectors. Instead, we reduced the feature dimension by convolution kernel matching the input feature maps. Compared to fully connection layers, the model maintained the context of the features to the utmost. For example, if the output of deconv2 in our network was 13×13×256, the deconv1 applied the 13×13×256×m (m is the output feature maps' number) deconvolutional kernel to de-convoluted the output of deconv2. It not only rendered a linear feature and maintained context, but also solved the dimension disaster. The model exhibited high accuracy in the test stage.

3.2 Strategies for Optimizing the Model

Deep learning has high performance in computer vision because the network contains millions of parameters. Thus, huge training data was required for fitting our model. However, there is limited underwater training data. For parameters optimization, we proposed some methods to solve the data starvation problem.

3.2.1 Transfer learning

In the machine learning task, we assume that the training data and test data obey the same distribution. If not, the hypothesis function cannot accurately predict the test data well. However, the expired training data in most cases and limited label make the assumed function ineffective. In transfer learning, the knowledge learned in an environment is used to assist the learning tasks in the new environment (Pan and Yang, 2010; Sun *et al*., 2018). Therefore, we applied the model trained by a large dataset to facilitate training the model that lacked data. It is helpful in solving the problem of limited label.

AlexNet is a fully trained model that can be applied to various vision tasks (Krizhevsky *et al*., 2012). We designed the convolution and de-convolution networks based on AlexNet. Moreover, we transfer the trained parameters to our network, that is, the part of the convolution shown in Fig.1. Due to parameters transfer, we trained the model faster and obtained higher accuracy, as shown in Table 1.

3.2.2 Data augmentation

For our experiments, we investigated a public dataset Fish4Knowledge project (Boom *et al*., 2014). It has 23 classes of fishes which are labeled by the marine biologist manually (Boom *et al*., 2012). Moreover, we applied data augmentation for increasing the training data. For data augmentation, we used a horizontal mirror, crop, down-sampling, and affine transformation, as suggested by reference (Sun *et al*., 2018). We mirrored the original image horizontally for simulating the different swim direction of the target. Then, we extracted the original picture on the left and right sides two-third for simulating the target appearing in the camera in different directions. In addition, the target was smaller for simulating the distance of the target distance from the camera by down sampling the image. For simulating the different gestures of the object more abundantly, we used the horizontal affine transformation, as shown in Eq. (4):

$$x' = \begin{cases} x + \sin\alpha\left(x - \frac{w}{2}\right), x \geq \frac{w}{2} \\ x - \sin\alpha\left(x - \frac{w}{2}\right), x < \frac{w}{2} \end{cases}, \quad (4)$$

where w is the width of the image and x is any one pixel in the image. Moreover, α presents the angle of affine transformation. We used different angles, $\alpha = -20^\circ$, $\alpha = -10^\circ$, $\alpha = 10^\circ$ and $\alpha = 20^\circ$.

3.2.3 Visualization of each layer

To demonstrate the effect of the deep features, we visualized the feature maps during the test procedure, as shown in Fig.4. The first convolutional layer focuses on the texture features. The feature maps of the pooling2 layer exhibit some discrete blocks, which indicates the features are much abstract than the former. In addition, the feature maps from deconv2 are much better and abstract. The first convolutional layers tend to learn features that resemble edges, lines, corners and shapes. The latter layers are closer to the outline of the object, which is important for the classification problem.

4 Experiments

4.1 Evaluation on Dataset

For verifying the effectiveness of the proposed method, we conducted experiments with the fish images dataset from the Fish4Knowledge project (Krizhevsky *et al*., 2012; Lines *et al*., 2001). We run all experiments in a

Python environment, and the configuration of our hardware was Inter(R) Xeno (R) CPU E5-2620 2.10 GHz, with NVIDIA GeForce GTX 980. We used the Caffe toolkit for training and testing the proposed ED-Net model.

4.2 Evaluation on Dataset

We applied ten-fold cross-validation for training and testing the proposed model. First, we divided the dataset into 10 copies. Then, we chose one copy as the test dataset, while others as train dataset. We calculated the average classification accuracy after experimenting 10 times. Due to the severity of the imbalance among categories, we listed the classification accuracy of each category and average accuracy of all categories for comparing with other methods. As shown in Table 1, we applied SVM (ED-Net-SVM) and softmax (ED-Net-Soft) for classifying an object. We compared our method with the state-of-the-art work DeepFish (Qin *et al*., 2015), UW-*CNN* (Sun *et al*., 2018), and our method with SVM performed best. The reason may be that linear SVM performs well with small categories.

From Table 1, we observe that, in most cases, the results of DeepFish (Qin *et al*., 2015) and UW-*CNN* (Sun *et al*., 2018) are much lower than those of the proposed method. Moreover, the proposed method achieves the best accuracies in nine categories. For the rest fourteen categories, the performance of our method is slightly lower than others.

Another challenge is training the network efficiently. For this, we transferred the parameters from AlexNet. Table 2 lists the results of the transferring part of the parameters from AlexNet. ED-Net-C1, ED-Net-C2, ED-Net-C1-C2 and ED-Net-FF denote the parameters transferred for conv1, conv2, conv1-conv2 and all layers. We observe that transferring more knowledge as initial values yields better performance. For example, the accuracy on category Canthigaster valentine increases from 73.33% to 98.88%. This knowledge learned from a big dataset is crucial for our special application.

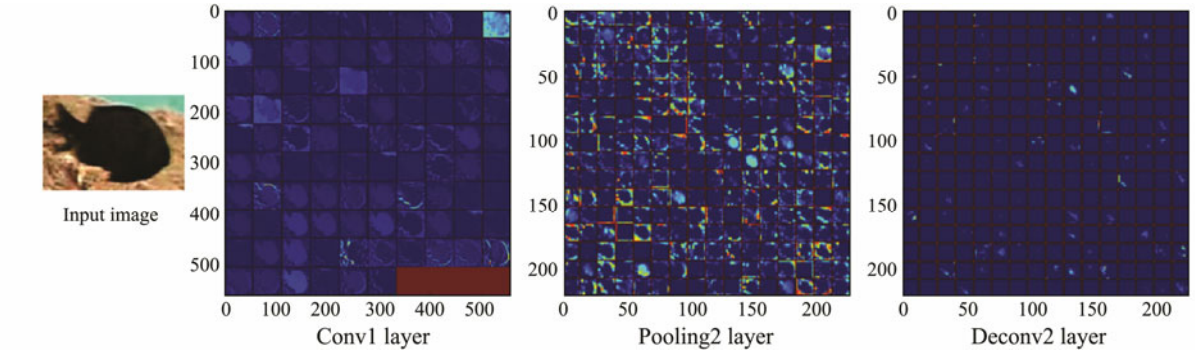

Fig.4 Visualization of feature maps for layers.

Table 1 Accuracy of different models in the Fish4Knowledge dataset

No.	Categories	ED-Net-SVM	ED-Net-Soft	DeepFish	UW-CNN
1	Dascyllus reticulatus	**100.00%**	0.00%	99.31%	99.78%
2	Plectroglyphidodon dickii	**99.44%**	96.81%	97.13%	98.79%
3	Chromis chrysura	99.54%	98.02%	98.64%	99.75%
4	Amphiprion clarkii	98.83%	92.12%	100.0%	99.97%
5	Chaetodon lunulatus	99.94%	99.75%	100.0%	100.0%
6	Chaetodon trifascialis	99.95%	99.73%	92.59%	100.0%
7	Myripristis kuntee	96.85%	86.79%	98.44%	100.0%
8	Acanthurus nigrofuscus	**99.74%**	58.69%	64.52%	89.05%
9	Hemigymnus fasciatus	91.92%	80.30%	100.0%	98.15%
10	Neoniphon sammara	**100%**	97.16%	100.0%	100.0%
11	Abudefduf vaigiensis	99.59%	95.91%	92.86%	100.0%
12	Canthigaster valentini	98.88%	83.33%	95.24%	100.0%
13	Pomacentrus moluccensis	**99.16%**	90.00%	100.0%	96.09%
14	Zebrasoma scopas	**99.41%**	100.00%	84.62%	85.06%
15	Hemigymnus melapterus	97.33%	61.33%	66.67%	100.0%
16	Lutjanus fulvus	91.30%	65.21%	96.55%	100.0%
17	Scolopsis bilineata	**100.00%**	96.87%	85.71%	100.0%
18	Scaridae	**100.00%**	91.83%	100.0%	86.67%
19	Pempheris vanicolensis	**100.00%**	98.07%	100.0%	100.0%
20	Zanclus cornutus	94.73%	94.73%	66.67%	100.0%
21	Neoglyphidodon nigroris	**100.00%**	92.30%	50.00%	84.62%
22	Balistapus undulatus	63.63%	40.90%	83.33%	95.45%
23	Siganus fuscescens	97.61%	95.23%	100.0%	100.0%
	Avg	**99.36%**	95.83%	90.10%	97.10%

Table 2 Accuracy of different fine-tuning layers in the Fish4Knowledge dataset

No.	Categories	ED-Net-C1	ED-Net-C2	ED-Net-C1-C2	ED-Net-FF
1	Dascyllus reticulatus	0.00%	0.00%	25.00%	100.00%
2	Plectroglyphidodon dickii	97.35%	93.95%	97.51%	99.44%
3	Chromis chrysura	97.64%	96.84%	91.32%	99.54%
4	Amphiprion clarkii	94.80%	94.95%	92.00%	98.83%
5	Chaetodon lunulatus	99.64%	99.70%	99.48%	99.94%
6	Chaetodon trifascialis	99.69%	99.91%	99.73%	99.95%
7	Myripristis kuntee	84.90%	77.98%	77.98%	96.85%
8	Acanthurus nigrofuscus	90.93%	72.29%	85.64%	99.74%
9	Hemigymnus fasciatus	77.77%	83.33%	79.79%	91.92%
10	Neoniphon sammara	91.03%	90.56%	90.09%	100%
11	Abudefduf vaigiensis	93.06%	93.87%	91.02%	99.59%
12	Canthigaster valentini	73.33%	76.66%	80.00%	98.88%
13	Pomacentrus moluccensis	91.66%	97.50%	95.00%	99.16%
14	Zebrasoma scopas	99.41%	98.83%	98.25%	99.41%
15	Hemigymnus melapterus	57.33%	21.33%	64.00%	97.33%
16	Lutjanus fulvus	71.73%	65.21%	69.56%	91.30%
17	Scolopsis bilineata	97.91%	97.91%	97.39%	100.00%
18	Scaridae	83.67%	55.10%	89.79%	100.00%
19	Pempheris vanicolensis	90.38%	55.76%	50.00%	100.00%
20	Zanclus cornutus	84.21%	100.00%	73.68%	94.73%
21	Neoglyphidodon nigroris	61.53%	76.92%	100.00%	100.00%
22	Balistapus undulatus	36.36%	54.54%	27.27%	63.63%
23	Siganus fuscescens	92.85%	92.85%	45.23%	97.61%
	Avg	96.67%	94.62%	95.51%	**99.36%**

5 Conclusions

Video technology is of great importance and is extensively applied for underwater environment observation. It remarkably promotes the research of marine science. Different from the conventional methods, video technology explores the underwater ecosystem continuously and non-invasively. Complex noise and low-light condition pose critical challenges for underwater video applications, which result from serious scattering and attenuation of light transport in water. In this paper, we proposed a new deep encode-decode convolution network for underwater object detection. First, it extracted the deep discriminative features from the noisy low-light underwater images. Second, we applied the deconvolutional layers to learn fine-grain deep features. Moreover, we used data augmentation and transfer learning for solving the problem of 'data starvation'. The experimental results showed that the proposed method achieved remarkable accuracy. To conclude, this work focused on contributing some new technologies to the research of marine science.

Acknowledgements

The study is supported by the Jilin Science and Technology Development Plan Project (Nos. 20160209006GX, 20170309001GX and 20180201043GX).

References

Bonin-Font, F., Oliver, G., Wirth, S., Massot, M., Negre, P. L., and Beltran, J. P., 2015. Visual sensing for autonomous underwater exploration and intervention tasks. *Ocean Engineering*, **93**: 25-44.

Boom, B. J., He, J., Palazzo, S., Huang, P. X., Beyan, C., Chou, H. M., Lin, F. P., Spampinato, C., and Fisher, R. B., 2014. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. *Ecological Informatics*, **23**: 83-97.

Boom, B. J., Huang, P. X., He, J., and Fisher, R. B., 2012. Supporting ground-truth annotation of image datasets using clustering. *21st International Conference on Pattern Recognition*. Tsukuba, Japan, 1542-1545.

Cappo, M., Harvey, E., and Shortis, M., 2006. Counting and measuring fish with baited video techniques – An overview. *Australian Society for Fish Biology Workshop Proceedings*. Hobart, Australia, 101-114.

Kim, Y., 2014. Convolutional neural networks for sentence classification. *Eprint Arxiv*, No. 1408.5882.

Krizhevsky, A., Sutskever, I. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. California, USA, 1097-1105.

Lines, J., Tillett, R., Ross, L., Chan, D., Hockaday, S., and McFarlane, N., 2001. An automatic image-based system for estimating the mass of free-swimming fish. *Computers and Electronics in Agriculture*, **31**: 151-168.

Mao, X. J., Shen, C., and Yang, Y. B., 2016. Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections. *Eprint Arxiv*, No. 1603.09056.

Noh, H., Hong, S., and Han, B., 2015. Learning deconvolution network for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*. Santiago, Chile, 1520-1528.

Pan, S. J., and Yang, Q., 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**: 1345-1359.

Pelletier, D., Leleu, K., Mou-Tham, G., Guillemot, N., and Chabanet, P., 2011. Comparison of visual census and high de-

finition video transects for monitoring coral reef fish assemblages. *Fisheries Research*, **107**: 84-93.

Qin, H., Li, X., Liang, J., Peng, Y., and Zhang, C., 2015. DeepFish: Accurate underwater live fish recognition with a deep architecture. *Neurocomputing*, **187**: 49-58.

Simonyan, K., and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *Eprint Arxiv*, No. 1409.1556.

Spampinato, C., Chen-Burger, Y. H., Nadarajan, G., and Fisher, R. B., 2008. Detecting, tracking and counting fish in low quality unconstrained underwater videos. *The 3th International Conference on Computer Vision Theory and Applications*, **2**: 514-519.

Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y. H., Fisher, R. B., and Nadarajan, G., 2010. Automatic fish classification for underwater species behavior understanding. *Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*. Firenze, Italy, 45-50.

Struthers, D. P., Danylchuk, A. J., Wilson, A. D., and Cooke, S. J., 2015. Action cameras: Bringing aquatic and fisheries research into view. *Fisheries*, **40**: 502-512.

Sun, X., Shi, J., Liu, L., Dong, J., Plant, C., Wang, X., and Zhou, H., 2018. Transferring deep knowledge for object recognition in low-quality underwater videos. *Neurocomputing*, **275**: 897-908.

Zeiler, M. D., and Fergus, R., 2014. Visualizing and understanding convolutional networks. *European Conference on Computer Vision*. Zurich, Switzerland, 818-833.

Zeiler, M. D., Taylor, G. W., and Fergus, R., 2011. Adaptive deconvolutional networks for mid and high level feature learning. *2011 IEEE International Conference on Computer Vision*. Barcelona, Spain, 2018-2025.

Zhang, N., Donahue, J., Girshick, R., and Darrell, T., 2014. Part-based R-CNNs for fine-grained category detection. *European Conference on Computer Vision*. Zurich, Switzerland, 834-849.

(Edited by Chen Wenwen)