Control Engineering Practice 118 (2022) 104458

Contents lists available at ScienceDirect

Control Engineering Practice

journal homepage: www.elsevier.com/locate/conengprac

Review on deep learning techniques for marine object recognition: Architectures and algorithms

Ning Wang *, Yuanyuan Wang, Meng Joo Er *

School of Marine Electrical Engineering, Dalian Maritime University, Dalian, 116026, China

ARTICLE INFO

Keywords:
Deep learning
Marine object recognition
Marine vehicles
Learning architecture

ABSTRACT

Due to the rapid development of deep learning techniques, numerous frameworks including convolutional neural networks (CNNs), deep belief networks (DBNs) and auto-encoder (AE), etc., have been established. In this context, advances in marine object recognition have been dramatically boosted, especially in the past decade. In this paper, we exclusively focus on an intensive review on deep-learning-based object recognition for both surface and underwater targets. To facilitate a comprehensive review, key concepts and typical architectures are firstly summarized in a unified framework. Accordingly, popular/benchmark datasets for marine object recognition are thoroughly collected and deep learning methodologies are comprehensively analyzed with intensive comparisons. Moreover, experimental results and futuristic trends in marine object recognition are intensively discussed. Finally, conclusions on state-of-the-art marine object recognition using deep learning techniques are drawn.

1. Introduction

Object recognition is a computer vision technique for identifying objects in images or videos. It includes the identification between two very similar targets, as well as that of one or two and even more kinds of targets in an image. As shown in Fig. 1, when looking at a picture or watching a video, humans can readily spot objects (cruise, sailboat), scenes (seawater, mountains), and visual details. The main goals of object recognition are to recognize objects from an image as humans do, and then to teach a computer to gain a level of understanding of what an image contains. Object can be recognizing from various view point such as from the front view, back view and the side view. The object can also be recognized in various sizes and when they are partly blocked from the viewer (Sharma, Singh, & Khurana, 2016). Various object recognition tasks including recognition of face (Liu et al., 2017), handwriting (Lecun et al., 1989; Lecun, Boser, Denker, Henderson, & Jackel, 1997), speech (Ayadi, Kamel, & Karray, 2011), license plate (Anagnostopoulos, Anagnostopoulos, Ioannis, Loumos, & Kayafas, 2008), lane line (Borkar, Hayes, & Smith, 2012), ship and military objects (Yang et al., 2018; Zabidi, Mustapa, Mokji, Marsono, & Sha'ameri, 2009), fish and underwater creatures (Jin & Hong, 2017; Meng, Hirayama, & Oyanagi, 2018), etc., have been widely investigated in recent years. Roughly two-thirds of the earth is covered by oceans (Yuh, Marani, & Blidberg, 2011), but comparatively not a lot of technologies pertaining to marine research have been thoroughly explored (Liu, Zhang, Yu, & Yuan, 2016). More importantly, marine safety including shipwreck, naval battle, etc., is a serious matter and there is a high demand of marine object recognition technologies for maritime surveillance.

Functionally, marine object recognition consists of two critical steps, namely feature extraction and classification, whereby the former is of more importance. In this context, conventional approaches such as manual classification and recognition, traditional statistical analysis and ocean model simulation, etc. heavily depend on the availability of visual features, and are inefficient and inaccurate for marine big data processing. For instance, in order to recognize surface ships, in Wang and Wang (2011), four geometric features were artificially defined in advance. In Liu, Yu, and Lv (2011), three affine invariant moment features of three-dimensional ships in different attitudes were extracted as features. In Zhu, Hui, Wang, and Guo (2010), Hu's moment invariants for feature extraction were deployed to extract shape and texture features of ship targets from remote sensing images. On one hand, it is computationally expensive to extract features; On the other hand, these types of features are local or low-level results such that the objects cannot be sufficiently represented. Bringing new opportunities to the intelligence of marine big data.

The ability of deep learning to process big data can meet the urgent requirements of fast and accurate analysis of marine big data, and can solve a series of marine problems such as marine disaster prevention and mitigation, ecological environmental protection, emergency rescue, and marine target detection, tracking and recognition.

* Corresponding authors.
E-mail address: n.wang@ieee.org (N. Wang).

[1] http://deeplearning.net/tutorial/lenet.html.

https://doi.org/10.1016/j.conengprac.2020.104458
Received 12 December 2019; Received in revised form 16 April 2020; Accepted 12 May 2020
Available online 23 May 2020
0967-0661/© 2020 Elsevier Ltd. All rights reserved.

Fig. 1. Using object recognition to identify one or multiple objects.

Deep learning methods have been widely applied to marine systems, for example, marine data reconstruction (CNN) (Ducournau & Fablet, 2016), marine data classification and recognition (CNN) (Duo, Wang, & Wang, 2019), marine data prediction (RNN, CNN) (Yang et al., 2017). As the first true multilayer-structure deep learning algorithm, the LeNet-5 (Lecun, Bottou, Bengio, & Haffner, 1998) that consists of two convolutional layers (convolutional and subsampling operators) and three fully-connected layers (two full connections and a Gaussian classification), can learn complex features through feature conversion, and then automatically simulate the characteristics of objects. It should be noted that the convolutional neural networks (CNNs) suffer from significantly computational complexity due to too much hyper-parameters. Meanwhile, traditional neural networks (NNs) (Wang, Er, & Han, 2015a), fuzzy NNs(FNN) (Oh, Kim, Nam, & Yoo, 2013; Wang, Er, & Han, 2015b; Wang, Er, & Meng, 2009; Wang, Sun, & Liu, 2016) and support vector machine (SVM) (Cortes & Vapnik, 1995; Pöyhönen, Arkkio, Jover, & H., 2005) are still actively deployed in the field of object recognition and classification. Since the single-layer restricted Boltzmann machine (RBM) was integrated into a deep neural network by Hinton, Osindero and Teh (2006), Hinton, Osindero, Welling and Teh (2006), deep learning framework has been increasingly established. In this context, high-level abstractions and/or discriminative features can be autonomously learnt from different complex datasets effectively, due to the deep networks with many levels of nonlinearities (Yuan, Huang, Wang, Yang, & Gui, 2018). In addition, powerful GPU with excellent computation capability has greatly boosted the efficiency of deep learning architectures, such as deep belief network (DBN) (Hinton, Osindero and Teh, 2006), CNNs,[1] recurrent neural network (RNN) (Graves, Liwicki, Fernández, Bertolami, & Schmidhuber, 2009) and auto-encoders (AE) (Vincent, Larochelle, Bengio, & Manzagol, 2008), especially in object classification and detection (Bejiga, Zeggada, & Melgani, 2016; Ren, He, Girshick, & Sun, 2015), since the complex features with deep architectures can be sufficiently explored. Moreover, robust training algorithms enable learning information of object representations without the need of artificial features (Lecun, Bengio, & Hinton, 2015). There are four key ideas behind CNNs that utilize the properties of natural signals, namely local connections, shared weights, pooling, and multiple layers (Lecun et al., 2015), each of which constitutes a nonlinear information processing unit. Since the breakthrough success of the AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) on ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 contest (Russakovsky et al., 2015), deep learning techniques have been closely connected to CNN models, including VGGNet (Alberto, Sergio, Sergiu, Victor, & Jose, 2017), GoogLeNet (Szegedy et al., 2015), and ResNet (He, Zhang, Ren, & Jian, 2016). These results show that a large and deep CNN is capable of achieving high accuracy on a large dataset using purely supervised learning.

In 2015, LeCun, Bengio and Hinton published a general review on deep learning (Lecun et al., 2015), which demonstrates that deep neural networks have succeeded in both industrial and academic community. Furthermore, specific object recognition surveys have also been reported in the areas of 3D object recognition (Mian, Bennamoun, & Owens, 2005; Singh, Mittal, & Bhatia, 2019), visual object recognition (Rajurkar, 2015), object segmentation and recognition (Sharma et al., 2016). It is apparent that many deep learning techniques and architectures for marine object recognition have been proposed over the years. Unfortunately, the pros and cons of these techniques are not clear. Moreover, the futuristic trends and challenges have not been identified. In this context, an intensive review on deep-learning-based marine recognition techniques is timely and is of both theoretical and practical significance in the marine engineering community.

The main contributions of this paper are as follows:

(1) An in-depth and comprehensive review on most popular techniques and typical deep network frameworks for marine object recognition for both surface and underwater targets is provided.
(2) The datasets (image or video) for underwater and surface objects from various view point (the front, the back and the side views) are thoroughly collected and analyzed.
(3) Experimental results of various deep learning methods for marine object recognition are comprehensively reviewed and compared.
(4) Futuristic trends and the possible challenges in marine object recognition using deep learning techniques are robustly discussed.

The remainder of this paper is organized as follows. In Section 2, key concepts and architectures pertaining to deep networks for object recognition are presented. Popular datasets are revisited in Section 3. In Section 4, typical deep learning methods together with comprehensive comparisons are systematically provided. A series of potential trends in future works are discussed in Section 5. Conclusions are drawn in Section 6.

2. Key concepts and architectures

As foreshadowed, objects can be from different platforms such as air (e.g. airplanes, birds) (Sommer, Schumann, Muller, Schuchert, & Beyerer, 2017; Vilches, Escobar, Vallejo, & Taylor, 2006), land (e.g. face, vehicles) (Anagnostopoulos et al., 2008; Liu, Wen et al., 2017), surface (e.g. ships, islands and buoys) (Zabidi et al., 2009) and underwater (e.g. fishes, sea cucumbers) (Jin & Hong, 2017; Meng et al., 2018), but can be any task-specific objects such as military objects (Yang et al., 2018). In this paper, we focus on marine object recognition.

To facilitate the understanding of how marine object recognition is done by deep learning techniques, it is necessary to formulate some key concepts and architectures here.

2.1. *Supervised learning*

2.1.1. *AlexNet*

As shown in Fig. 2, the AlexNet (Krizhevsky et al., 2012) is an intuitive structure consisting of five convolutional layers, which are stacked by pooling and/or nonlinear layers followed by three fully connected layers.

Using creative techniques including data augmentation (DA) (He et al., 2016; Simonyan & Zisserman, 2014), dropout (Bell & Koren, 2007), rectified linear units (ReLUs) (Nair & Hinton, 2010; Xavier, Antoine, & Yoshua, 2011), local response normalization (LRN) and overlapping pooling (Krizhevsky et al., 2012), the AlexNet won the championship with the testing error of 15.4% and far outstripped the second place winner with the testing error of 26.2%.

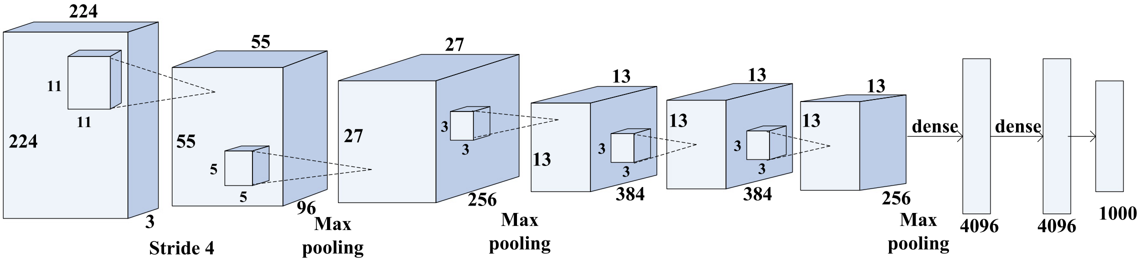

Fig. 2. AlexNet architecture.

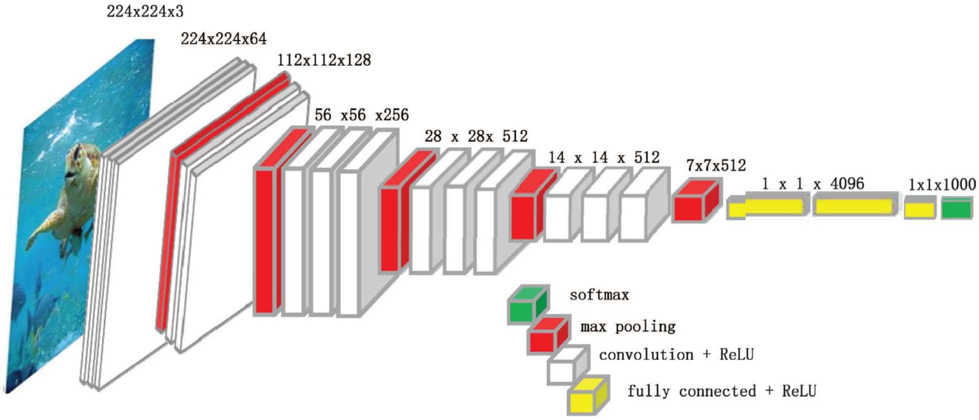

Fig. 3. VGG-16 architecture.

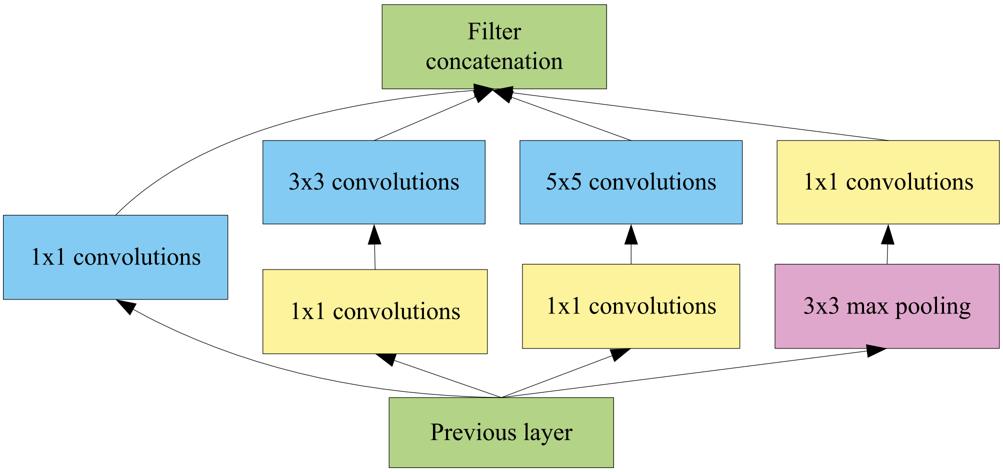

Fig. 4. Inception module with dimensionality reduction.

2.1.2. ZFNet

As a variant of the AlexNet, the ZFNet was proposed by Zeiler and Fergus (2014). The main differences between the ZFNet and AlexNet lie in the visualization of first and second layers: (a) The first-layer filter size was reduced from 11 × 11 to 7 × 7, and (b) The convolutional stride was reduced from 4 to 2, and thereby retaining much more information in those two layers. It should be noted that the visualization operation is facilitated by the deconvolution layer (Zeiler, Taylor, & Fergus, 2012).

2.1.3. VGG-16

As shown in Fig. 3, the Visual Geometry Group Network (VGGNet) (Alberto et al., 2017), is a kind of CNN model developed by the Visual Geometry Group, University of Oxford. As various variants of deep CNNs (Simonyan & Zisserman, 2014), the VGG-16 and VGG-19 are typical models. In particular, the VGG-16 which is composed by 16 weight layers achieved 7.3% testing error, and was first runner-up in the ILSVRC-2013 competition.

To be more specific, it has been illustrated that representation depth and small receptive field, using a conventional CNN architecture (Krizhevsky et al., 2012; Lecun et al., 1989) with substantially increased depth (Simonyan & Zisserman, 2014), benefit the classification accuracy, thereby contributing to remarkable performance on the ImageNet challenge dataset.

2.1.4. GoogLeNet

Motivated by the philosophy of Network in Network (NiN), the GoogLeNet was proposed by Szegedy et al. (2015), whereby 22 layers in the inception framework are involved with learnt parameters. Within

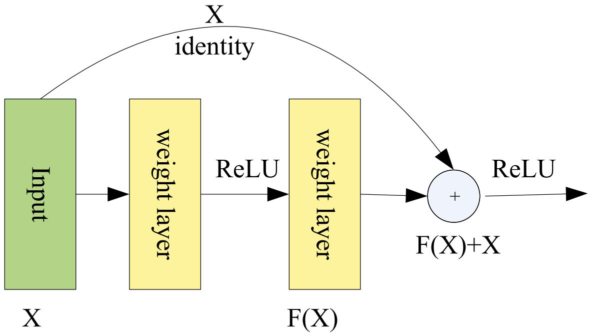

Fig. 5. Residual learning: a building block.

the GoogLeNet, the multi-layer stack of inception-module blocks (see Fig. 4) including pooling operation, large-size and small-size convolution layers are deployed. In order to reduce computational dimensionality, all the aforementioned blocks are computed in parallel together with 1 x 1 convolution operations. In addition, it has been proven that higher accuracy can be achieved by using more convolutional operators and/or deeper layers.

In essence, the GoogLeNet aims to reduce the number of feature filters of each layer, and thereby reducing computational complexity. As a consequence, the GoogLeNet achieved remarkable performance in classification and detection with a top-5 test error rate 6.7% in the ILSVRC-2014 competition.

2.1.5. ResNet

The Residual Neural Network (ResNet) with high depth (152 layers) was proposed by He et al. (2016), and won the ILSVRC-2015 competition with the testing error of 3.57%.

In the ResNet, residual blocks benefit from residual learning (see Fig. 5), whereby the output is composed of the original input **X** and the last-layer output **F**(**X**). The direct connection from the input layer to the output layer is termed as "identity skip connection". It should be noted that the ResNet 50 and ResNet 101 have been widely applied in the areas of detection, segmentation and recognition.

2.1.6. SENet

The Squeeze-and-Excitation Network (SENet) (Hu, Shen, Albanie, Sun, & Wu, 2017) has won the championship with the testing error of 2.251% in the ILSVRC-2017 competition, surpassing the winning entry of the ILSVRC-2016 competition by a significant improvement of 25%.

In essence, the SENet is formulated by stacking a collection of "SE block" (see Fig. 6), whereby the representations can be significantly improved by explicitly modeling the interdependencies between convolutional features. In this context, dynamic channel-wise feature recalibration can be eventually performed. Furthermore, the SE block, can be embedded into other frameworks, i.e., SE-Inception and SE-ResNet modules (Hu et al., 2017). In addition, since the SENets are not restricted to a specific dataset or task, better generalization ability can be ensured.

For clarity, the testing error of these state-of-the-art deep networks on ImageNet, their structures and techniques are shown in Table 1. As foreshadowed, the first three models are constructed by convolutional and fully-connected layers, the GoogLeNet and SENet are designed by inserting inception architecture to conventional CNN, whereas ResNet is derived from formulating the layers as learning residual functions about the inputs layer, instead of learning unreferenced functions. Besides, a series of training techniques, including DA, Dropout, LRN and BN are deployed in models. The DA is applied to all networks here since it can enrich the diversity of data. The dropout and LRN are deployed to reduce overfitting. However, it has been proved that the LRN is not effective (Simonyan & Zisserman, 2014). In this context, the BN that is similar to the normalization can improve the training speed and benefit the comparatively deeper networks. As can be seen from Table 1, with increasing network depth, the training accuracy becomes better. The SENet significantly outperforms past models on the ImageNet challenge dataset. However, the computational cost is much higher than that of the past models. Therefore, a proper model should be chosen based on the tradeoff between speed and accuracy. At present, there are no general guidelines of selecting a perfect deep learning architecture for different recognition tasks, which in turn often depend on the practical applications and datasets.

2.2. Unsupervised learning

Unsupervised deep learning models include DBNs and various stacked variants of AE including denoising AE (DAE) (Vincent et al., 2008), stacked sparse AE (SAE) (Luo & Wan, 2013), and contractive AE (CAE) (Rifai, Vincent, Muller, Glorot, & Bengio, 2011). Generally speaking, parameters in unsupervised learning models are optimized by the greedy layer-wise training method, which is divided into two phases, i.e., pre-training and fine-tuning. In the pre-training phase shown in Fig. 7(a), a series of three-layer (Input, Hidden and Output layers) NNs are trained individually by minimizing the reconstruction error. As shown in Fig. 7(b), to facilitate fine tuning, all hidden layers with parameters trained in the pre-training stage are stacked together in a sequence. In this context, a classifier or classification layer is deployed to be the last layer, and thereby achieving fine-tuning by optimizing the loss function determined by the specific task at hand.

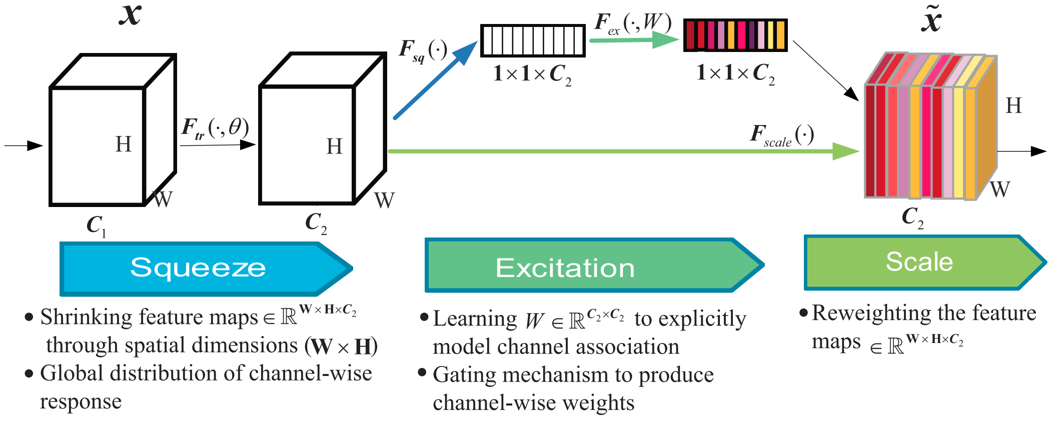
x
H
W
C_1
$F_{tr}(\cdot,\theta)$
H
W
C_2
$F_{sq}(\cdot)$
$1\times1\times C_2$
$F_{ex}(\cdot,W)$
$1\times1\times C_2$
$\tilde{x}$
$F_{scale}(\cdot)$
H
W
C_2
Squeeze
Excitation
Scale
• Shrinking feature maps $\in\mathbb{R}^{W\times H\times C_2}$ through spatial dimensions ($\mathbf{W}\times\mathbf{H}$)
• Global distribution of channel-wise response
• Learning $W\in\mathbb{R}^{C_2\times C_2}$ to explicitly model channel association
• Gating mechanism to produce channel-wise weights
• Reweighting the feature maps $\in\mathbb{R}^{W\times H\times C_2}$

Fig. 6. A Squeeze-and-Excitation block.

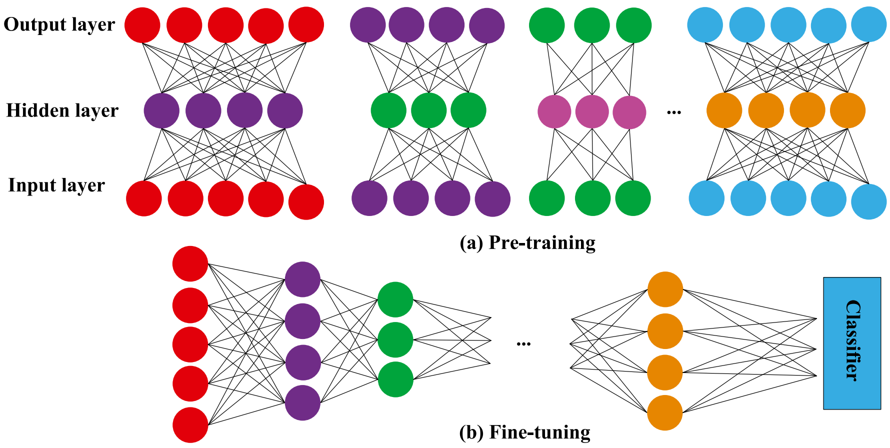

Fig. 7. Unsupervised deep neural networks.

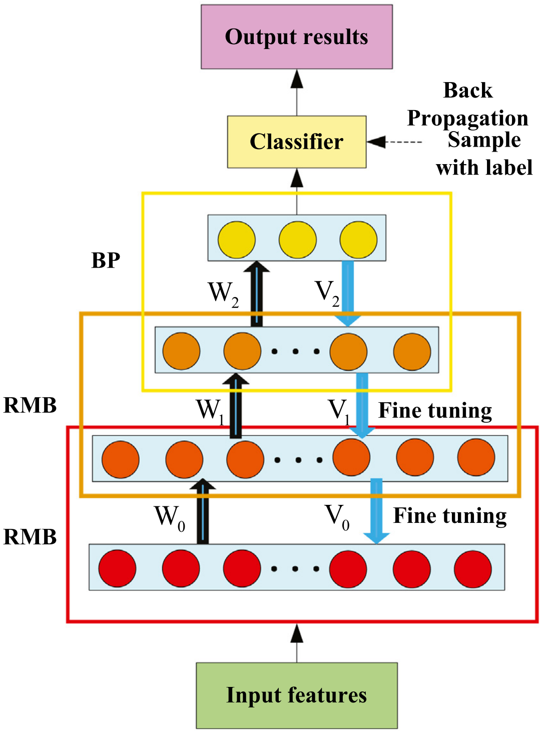

Fig. 8. DBN formed by stacking two RBMs.

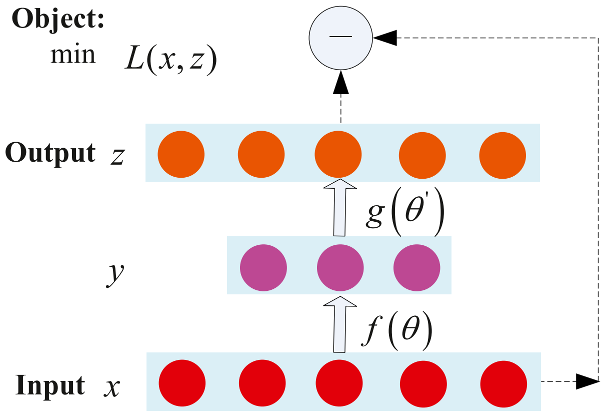

Fig. 9. Training process of AE.

```
Algorithm 1: Auto-Encoder
Encoder: Input vector $x$, the corresponding output $y$: $y = f_\theta(x) = s(Wx + b)$
Decoder: Reverse mapping from $y$ to the reconstruction vector $z$:
$z = g_{\theta'}(y) = s(W' y + b')$
Objective: Minimize the reconstruction error function:
$\arg_{\theta,\theta'} \min E_{q^0(x)} \left[\log p(x|z = g_{\theta'}(f_\theta(x)))\right]$.
```

2.2.1. Deep belief network

As shown in Fig. 8, the DBN formulates a multi-layer generative model with joint probability distribution, whereby multiple RBMs and a back-propagation layer are stacked in order.

In the entire DBN structure, similar to the feedforward NNs, two adjacent layers comprise full connections, while there do not exist connections between neurons in the same layer. To identify the weights layer by layer in order, a greedy algorithm for the RBM was proposed by Hinton, Osindero and Teh (2006), since the outputs/activations of the previous layer serve as the inputs of successive layers within the RBM.

2.2.2. Auto-encoder

As shown in Fig. 9, the training process of the AE can be divided into encoder and decoder steps. As summarized in Algorithm 1, the essence of AE is to reconstruct the input signals from the outputs, i.e., the output z is as close as possible to the input x by minimizing the reconstruction error function.

2.3. Transfer learning

Transfer learning (Pan & Yang, 2009; Yang et al., 2018) is a model transformation technique for small datasets, i.e., transferring the learning parameters of the model well pre-trained by a large-scale dataset with labels (e.g. ImageNet) to a new model. Alternatively, even if the available dataset is large enough, it is also helpful to begin with pre-trained weights instead of randomly initialized ones. In this context, the weights of pre-trained model are fine tuned to satisfy the new model.

Yosinski, Clune, Bengio, and Lipson (2014) proved that the model with transferring weights performs better than those using random initialization. It should be noted that the transferring ability of features decreases as the difference between the source and target tasks increases. Besides, transfer learning technique only regards partial parameters of the pre-trained model as initial values of the new model. Moreover, architectural constraints have to be met in the pre-trained model, e.g., properly choosing layers. Therefore, it is not general to

Table 1
Comparisons of various architectures on ImageNet.

	AlexNet (2012)	ZFNet (2013)	VGG-16 (2014)	GoogLeNet (2015)	ResNet (2016)	SENet (2017)
Top-5 testing error	16.4%	11.2%	7.32%	6.66%	3.57%	2.251%
Layers	8	8	19	22	152	154
Convolution Layer	5	5	16	21	151	153
FC layer (Softmax)	3	3	3	1	1	1
Filter size	11,5,3	7,5,3	3	7,1,3,5	7,1,3,5	3
FC layer size	4096,4096,1000	4096,4096,1000	4096,4096,1000	1000	1000	1000
Inception (NiN)	✗	✗	✗	✓	✗	✓
DA	✓	✓	✓	✓	✓	✓
Dropout	✓	✓	✓	✓	✓	✓
LRN	✓	✓	✗	✓	✗	✗
BN	✗	✗	✗	✗	✓	✓

Note: LRN (Local Response Normalization), FC (Full-Connection), DA (Data Augmentation), BN (Batch Normalization). "✓" and "✗" denote whether the operation is involved or not, i.e., "✓" yes, "✗" no.

train a new architecture from scratch with a small training dataset, but to reuse the existing network architectures.

2.4. Data augmentation

The DA (Krizhevsky et al., 2012) is also used in small and unbalanced datasets. It typically consists of a set of transformations in either data or feature spaces. In essence, the DA aims to generate more training samples to broaden the data dimension and/or diversity. As shown in Fig. 10, the transformation techniques of DA include parallel mirror, warps, scales, crops and rotations, etc. Various tools, including ImageMagick,[2] Keras Image Augmentation API,[3] Augmentor[4] in OpenCV, have been flexibly applied to data transformations.

In this context, DA can effectively prevent overfitting, and synthetically produce new samples which are more representative for the tasks at hand.

Obtaining sufficient data is one of the core issues in deep learning, making the learning model effective and preventing overfitting. However, the training images are usually small for image recognition, or the size and quality are not even. To the best of our knowledge, the corresponding tackling techniques that are used in the literatures are DA and transfer learning, especially the latter, since some researches have shown that CNN models trained by ImageNet can be regarded as the generalized feature extractors, which are capable of producing powerful high-level features for many new related tasks (Sun et al., 2017).

3. Datasets

A large-scale and trainable dataset is a core modality for marine object recognition and is essential for training deep learning frameworks. Here, we firstly list typical large-scale image datasets which include many kinds of objects and can be deployed to be source domain datasets in the pre-training phase of transfer learning. To facilitate the following discussion, we have collected the underwater and surface object datasets for marine object recognition, and the popular datasets are listed in Table 2.

Currently, the most popular large-scale datasets used for object recognition, including ImageNet,[5] Pascal VOC,[6] standard Caltech-101 dataset,[7] Open Image,[8] and COCO,[9] etc., can be freely downloaded from websites.

The large-scale image datasets tend to own sufficient images to train a complex deep model. However, in practice, it is difficult for the marine engineering community including both surface and underwater communities to acquire high-resolution training images and/or videos due to complex marine environments. In this context, as summarized in Table 2, the existing marine datasets suffer from data and/or label starvation or distribution unbalance.

The public datasets for underwater object recognition consist of ground-truth dataset: Fish4 Knowledge project (Fish4K)[10] (Hinton, Osindero and Teh, 2006; Lines et al., 2001), Kyutech 10K (Lu et al., 2018) which is the first deep-sea marine organism dataset and provided by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC), etc.

From different views, marine ship datasets can be divided into two types: overhead and front-view images. The overhead images are obtained by the aerial equipments in general, including radars, satellites, and unmanned aerial vehicles (UAVs). For instance, the existing high-resolution optical satellite images consist of SPOT-5 (Corbane, Najman, Pecoul, Demagistri, & Petit, 2010; Corbane et al., 2008; Proia & Page, 2010; Tang, Deng, Huang, & Zhao, 2014; Yang, Li, Ji, Gao, & Xu, 2013), WorldView-2 (Yokoya & Iwasaki, 2015), QuickBird (Liu et al., 2013), Venezuelan Remote Sensing Satellite (VRSS-1) (Zou & Shi, 2016), GaoFen-1 (Qi, Ma, Lin, Li, & Tian, 2015; Zou & Shi, 2016), and Google Earth (Guo, Xia, & Wang, 2014; Shi, Yu, Jiang, & Li, 2014; Xu, Sun, Zhang, & Fu, 2014; Yu, Guan, & Zheng, 2015; Zhang, Yao, Zhang, Feng, & Zhang, 2016). In particular, ship images from Google Earth are widely utilized in the literatures, e.g. "High Resolution Ship Collection 2016 (HRSC2016)[11] (Liu, Yuan et al., 2017)", which covers not only bounding-box labeling way, but also rotated bounding box way with three-level classes including ships, ship categories and types. The datasets for ship recognition in Guo et al. (2014) and Guo and Xia (2017) are randomly selected from Google map's aerial images. In addition, the front view images refer to the side views of ships. For instance, the "MARitime VEsseLs (MARVEL)[12] (Erhan et al., 2016)" is the largest fine-grained visual recognition dataset. The "FleetMon[13] (Zhao et al., 2016)" owns more than 0.5 million ship images and provides the ships from all over the world with name, the international maritime organization (IMO)or maritime mobile service identity (MMSI) numbers, flag state, length and vessel types. The "E2S2-Vessel" is created by collecting the high-resolution images from online maritime vessel photos and website ShipSpotting[14] (130 000 images, 35 classes), whereby the images of maritime vessels with various annotations are uploaded by the hobby photographers, on-ship and on-shore cameras. Besides, remote sensing images of ships are provided by 2017 CCF BDCI (Big Data & Computing Intelligence Contest), which can be downloaded

[2] http://www.imagemagick.org/.
[3] https://machinelearningmastery.com/image-augmentation-deep-learning-keras/.
[4] https://augmentor.readthedocs.io/en/master/index.html.
[5] http://www.image-net.org/.
[6] http://pjreddie.com/projects/pascal-voc-dataset-mirror/.
[7] http://www.vision.caltech.edu/Image_Datasets/Caltech101/101_ObjectCategories.tar.gz.
[8] https://github.com/openimages/dataset.
[9] http://images.cocodataset.org/.
[10] http://groups.inf.ed.ac.uk/f4k/GROUNDTRUTH/RECOG/.
[11] http://www.escience.cn/people/liuzikun/DataSet.html.
[12] https://github.com/avaapm/marveldataset2016.
[13] https://www.fleetmon.com/vessels/.
[14] www.shipspotting.com.

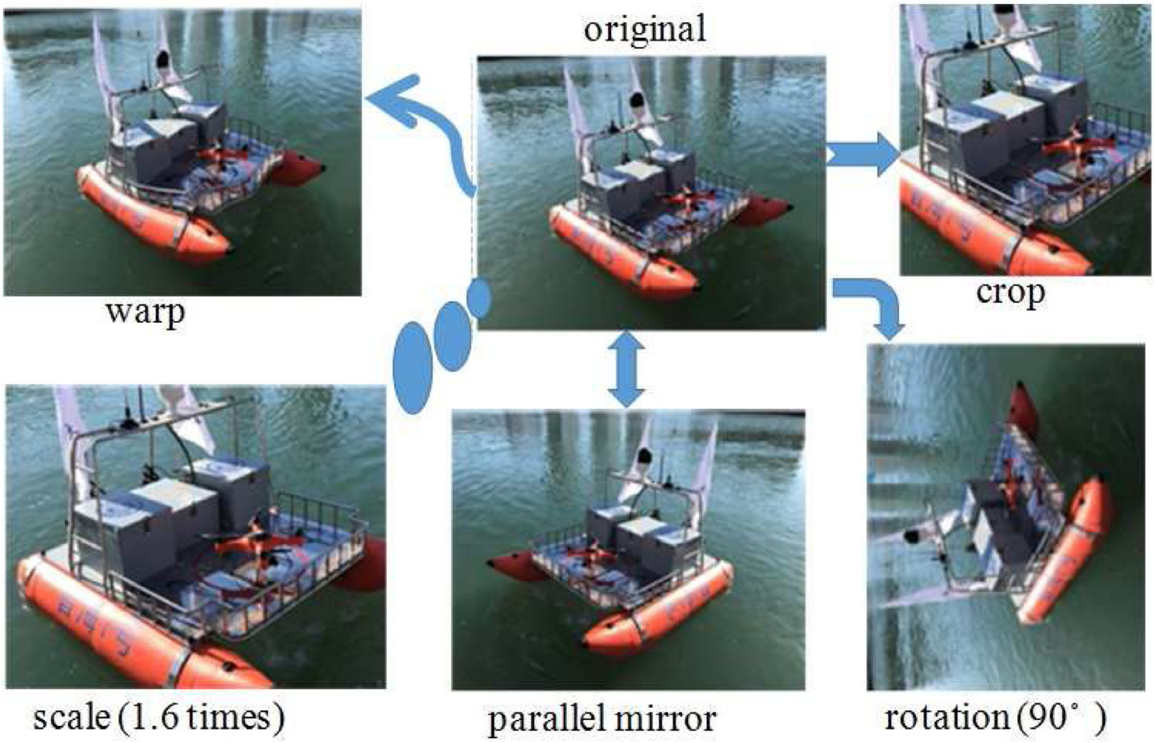

Fig. 10. Transformation techniques of DA.

Table 2
Popular underwater and surface object recognition datasets.

Dataset	Target	Purpose	Class	Type	Resolution (m) or pixels	Samples
Fish4K (Lines et al., 2001)	Fish	Under-water	17	Stereo pair videos	384 × 288	26 + 34
Kyutech-10K (Lu et al., 2018)	Marine organism	Under-water	7	Images+Videos	480 × 640	10 728 + 1489
QuickBird (Liu et al., 2013)	Ship	Surface	–	Satellite imagery	0.61 m or 200 × 200–800 × 800	88 + 37
SPOT-5 (Corbane, Pecoul, Demagistri, & Petit, 2008)	Ship	Surface	–	Satellite imagery	5 m or 3000 × 3000	79 + 37
HRSC2016 (Liu, Yuan, Weng and Yang, 2017)	Ship	Surface	3	Satellite images	0.4 m–2 m or 300 × 300–1500 × 900	1207 + 541
R2VV-p (Lang & Wu, 2017)	Ship	Surface	3	SAR images	15 m	75 + 75
VAIS (Zhang, Choi, Daniilidis, Wolf, & Kanan, 2015)	Ship	Surface	6	Visible–Infrared images	1024 × 768	544 + 107
E2S2-Vessel (Daoduc, Xiaohui, & Morère, 2015)	Ship	Surface	35	Visible images	256 × 256	103 000 + 26 000
FleetMon (Zhao, Wang, & Yuan, 2016)	Ship	Surface	5	Images	>200 × 200	500 + 310
MARVEL (Erhan, Berkan, Veysel, & Aykut, 2016)	Ship	Surface	26	Visible images	1024 × 768	212 992 + 26 624
CCF-BDCI (Xiong, Ding, Deng, Fang, & Gong, 2018)	Ship	Surface	3	Remote sense images	1024 × 1024	13 668 + 3417

Note: "–" denotes unknown term. In the last column. "A+B" denotes "Sum of training and testing samples". "R2VV-p" denotes "Radarsat-2 standard-mode VV-polarization images".

from the website,[15] including nine coarse-grained categories, namely buoy, cruise ship, ferry boat, freight boat, gondola, kayak, paper boat and sailboat. In Xiong et al. (2018), the database has been used in the aforementioned contest as experimental samples for surface ships recognition, whereby 17085 images are divided into 13668 training samples and 3417 testing samples.

The infrared images produced by long-wavelength infrared (LWIR) cameras are capable of capturing ships during the day and night, and can measure thermal emissions in the environment. However, the number of images in the existing datasets are usually small. In Teutsch and Krüger (2010), an infrared dataset consisting of 2205 open water images has been created to determine if an image includes a ship, clutter, or an irrelevant object. In Withagen, Schutte, Vossepoel, and Breuers (1999), a ship dataset including 136 infrared images has been built. This dataset consists of six kinds of vessels, and are deployed to recognize the individual ships rather than ship categories. In Zhang et al. (2015), the paired visible and infrared ship imagery (VAIS[16]) is the only existing database of "Visible–infrared" paired ship images, whereby 2865 images (1623 visible and 1242 IR) are collected.

Airborne and spaceborne synthetic aperture radar (SAR) images are also particularly suitable for object classification and recognition, since they can be obtained in all-weather day-and-night conditions and have very high resolution. The "Moving and Stationary Target Acquisition and Recognition (MSTAR)" benchmark dataset (Chen, Wang, Feng, & Jin, 2016) is the most typical dataset, and has been widely used to test and analyze SAR automatic target recognition (SAR-ATR) algorithms with intensive comparisons. In Ødegaard, Knapskog, Cochin, and Louvigne (2016), the simulated dataset has been used as SAR images for ships classification. In Kumlu and Jenkins (2013), synthetic and real images are used for learning and validation, respectively. In addition, electro-optical images have also been suggested for ship recognition (van den Broek et al., 2014).

Apart from the aforementioned datasets comprising of single-view images, 360-degree panoramic images getting multi-views information have attracted more attention. An underwater drone equipped with fisheye lenses has been designed in Meng et al. (2018), whereby the 360-degree images can be generated with a 360-degree panoramic camera by using an image generation algorithm.

[15] https://www.datafountain.cn/datasets/.
[16] http://vcipl-okstate.org/pbvs/bench/.

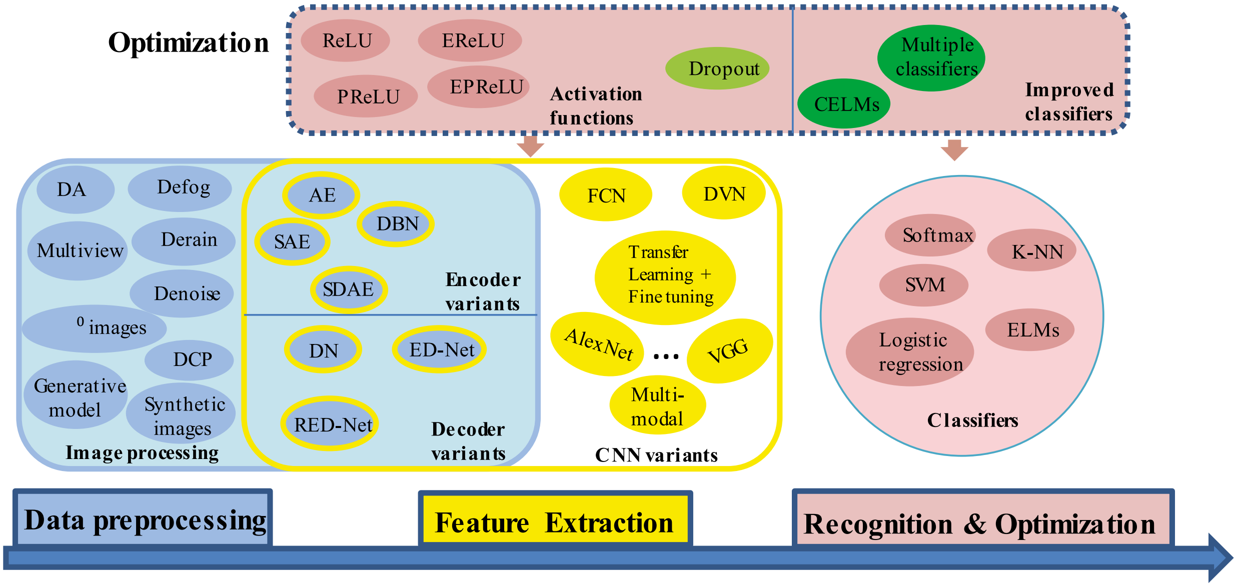

Fig. 11. Classification of the reviewed methods.

Remark 1. For the marine object recognition task, the information from different modalities, such as images (visual (Erhan et al., 2016), radars (Chen et al., 2016), satellites (Corbane et al., 2008), infrared (Teutsch & Krüger, 2010), visible, electro-optical images (van den Broek et al., 2014), 3D (RGB-D) (Richard, Brody, Bharath, Christopher, & Andrew, 2012) and videos (Wang, Ouyang, Li, & Zhang, 2019), can be fused together to achieve a more discriminant network. For example, paired visible–infrared images (Zhang et al., 2015).

4. Methodology

As shown in Fig. 11, marine object recognition can be divided into three modules, i.e., data preprocessing, feature extraction, and recognition and model optimization.

Significantly different from the object recognition tasks in the ground and static environments, the marine environments become rather challenging on the sea or underwater. It is well known that deep learning models need a large amount of high-quality images or videos as training samples. In this context, the quality of datasets which is mainly affected by complex marine environments (e.g. rain, snow and fog, the cameras and underwater noises) becomes vital. Therefore, image preprocessing plays a very pivotal role in marine object recognition. Of course, some deep network structures own the capability of image preprocessing to some extent, i.e., supervised deconvolutional network (Zeiler, Krishnan, Taylor, & Fergus, 2010), deep encoding–decoding network (Wang et al., 2019) and residual encoder–decoder network (Mao, Shen, & Yang, 2016), and unsupervised DBN (Hinton, Osindero and Teh, 2006) and AE (Vincent et al., 2008). Image preprocessing methods will be reviewed in Section 4.1.

Considering the difficulties of collecting images in complex marine environments, transfer learning method is usually used in classic deep networks for object recognition and classification tasks. Various variants of commonly used deep model structures are further proposed to improve precision. Feature extraction methods will be presented in Section 4.2.

Under the high-resolution images condition, classic models have been applied with great success in object recognition. However, to improve the recognition accuracy and training speed, a series of improved algorithms have been proposed. The recognition and model optimization works are summarized in Section 4.3.

4.1. Data preprocessing

Image preprocessing is a vital step for marine object recognition within deep learning models since images or videos are the precondition and application foundation of deep learning methods. In this context, image preprocessing techniques need to be incorporated into deep learning frameworks for marine object recognition.

For the images of which the main scene consists of cloud, rain and islands, a discriminant method based on the dark channel prior (DCP) theory has been proposed in Xiong et al. (2018), such that the original images can be automatically classified into two categories, i.e., clear and fuzzy images. Fuzzy images need to be tackled by the DCP defogging algorithm such that the image quality can be enhanced. Afterwards, all clear images are utilized to train the pre-trained AlexNet. Differing from the surface images, underwater object images or videos captured by underwater cameras are usually deteriorated by underwater noises, i.e., pulse and Gaussian noises. An improved median filter has been designed in Jin and Hong (2017), whereby only those pixels polluted by impulse noises, rather than all pixels, are processed. Considering low-contrast and low-resolution videos in the low illumination environment, in Sun et al. (2017), a real-time object recognition system for underwater videos has been designed by transferring the knowledge and parameters from the pre-trained AlexNet to the target domain, rather than processing the low-contrast images directly.

Different from image preprocessing techniques, deep convolutional network (ConvNet) architectures and their variants have been shown to be remarkably efficient for low-quality images or videos, since the deep networks can tolerate some disturbances or noises. Deconvolutional Network (DN) proposed by Zeiler et al. (2010) is such a typical framework (see Fig. 12) that permits unsupervised construction of hierarchical image representation, and can be used for low-level tasks such as denoising and providing features for marine object recognition.

The deep encoding–decoding network (ED-Net) (Wang et al., 2019) was designed by using the DN such that the discriminative features can be directly extracted from the noisy low-light underwater images rather than high-resolution images. Moreover, deconvolutional layers (see Fig. 18) instead of full-connection layers were applied since this operation can be used as a decoding procedure of the convolution operation, and thereby removing most noises from the images.

Another approach is the fusion of image preprocessing and deep learning methods. A filtering deep convolutional network (FDCNet) (Lu

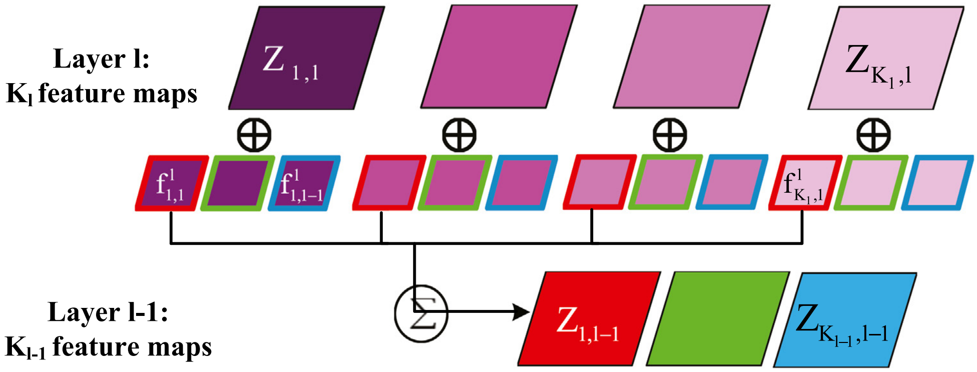

Fig. 12. A single Deconvolutional Network layer. (For clarity, only the connectivity for a single input map is shown here.)

et al., 2018) has been constructed to classify deep-sea objects. As shown in Fig. 13, underwater dark channel prior (UDCP) estimator is applied to calculate disparity which is the key point for underwater descattering. Meanwhile, the unary and pairwise super-pixel in a unified deep CNN are learnt. In this context, the disparities of CNN and UDCP are combined by joint bilateral filtering to achieve the classification task.

As foreshadowed, marine datasets suffer from the starvation of accessible training data due to complex marine environments. In this context, the DA and transfer learning as well as deep learning generative models are deployed. In addition, in Pei et al. (2017), a multiview deep learning framework was employed for SAR-ATR. The multiview SAR images are obtained through a given ground target from different aspect angles in different intervals by the SAR platform, without using too many raw SAR images. It should be noted that the newly generated data and the raw images cannot be treated equally. In this context, a parallel network topology with multiple inputs was required such that the features of SAR images from different views can be extracted and fused layer by layer progressively. In particular, the 4-Views DCNN (Pei et al., 2017) achieved a better recognition rate than the existing SAR-ATR methods under the standard operating condition.

Similar to the aforementioned multiview images, in Meng et al. (2018), the 360-panoramic images are generated by correcting the images which are taken by an underwater drone with a 360-panoramic camera. The generated high-quality images of underwater objects benefit fish recognition and has achieved an accuracy of 87% by deep learning techniques.

4.2. Feature extraction

4.2.1. Transfer learning and fine-tuning techniques

Recently, object recognition through transfer learning on CNNs has achieved considerable improvements (Pan & Yang, 2009). Due to the difficulty in obtaining the images dataset on the sea or underwater, it usually assumes that a pre-trained model has already learnt features that are useful for the recognition task. In this context, transfer learning together with fine-tuning has been an important technique for object recognition and has been successfully applied in various areas (Malmgren-Hansen et al., 2017; Sun et al., 2017; Yang et al., 2018).

As illustrated in Fig. 14, by transferring the parameters and structure of the AlexNet to the target domain as its initialization, a method to recognize objects from the underwater videos has been proposed in Sun et al. (2017). In addition, the new underwater dataset augmented by the DA is deployed to supervised fine-tuning of these parameters in the training process of target domain. Finally, a suitable model can be obtained and is named “UW-CNN”.

Different from Sun et al. (2017), in the training process of target domain, only partial parameters in the entire layers are required to be updated. As shown in Fig. 15, in Yang et al. (2018), the last three layers are retrained while parameters of the remaining layers are preserved according to the fact that the deeper layer can learn semantic features. The operation is different from the conventional transfer learning and is named “deep transfer learning”.

4.2.2. Deep convolutional network variants

In the aspect of network structures, a question about the significance of depth: “Is learning better for networks with deeper layers?” has been discussed in Szegedy et al. (2015). Results show that the deeper models, e.g., ResNet101 and ResNet1001, have achieved better results. However, sharp disadvantages, including increasing number of parameters and reduction of convergence speed, are also not to be ignored. The existing and well-known models, including VGG-16, GoogLeNet, and ResNet mentioned above, are variants of the AlexNet framework in different tasks and/or aims.

It has been found that a large portion of trainable parameters in deep convolutional networks are induced by the fully-connected layers (Alberto et al., 2017). As shown in Table 3, roughly 60 million parameters are deployed in the entire AlexNet architecture, astonishingly, wherein 96.2% is derived from the last three fully-connected layers. Therefore, it is high time to find ways to reduce the computational burden. Fortunately, the fully convolutional and deconvolutional operations have been proposed by removing the fully-connected layers based on the existing and well-known CNNs in the sequel.

(a) Fully Convolutional Network

The Fully Convolutional Network (FCN) proposed by Long, Shelhamer, and Darrell (2015) firstly applied for semantic segmentation task in an end-to-end manner to solve structured pixel-wise labeling problem. As shown in Fig. 16, the entire framework is composed of only convolutional layers and a classifier, by transforming fully-connected layers into convolutional layers rather than simply removing them. However, the FCN without fully-connected layers resembles architecture of the Fukushima’s neocognitron (Fukushima, 1980) which is regarded as the origin of the ConvNets, such that the computation is much more complex than ConvNets. It should be noted that convolution outputs are spatial maps rather than feature vectors such that the FCN is not affected by the limitation of fix-size input images. More importantly, the FCN tremendously reduces the number of free parameters. It has been applied for aerial image semantic labeling (Sherrah, 2016) and large-scale remote sensing classification (Fu, Liu, Zhou, Sun, & Zhang, 2017). However, to our best knowledge, the FCN architecture has not been applied for object recognition tasks.

Similar to the aforementioned architecture of the FCN without fully-connected layers but sparse connection or convolution, all-convolutional networks (A-ConvNets) whose neurons of layers are arranged in three dimensions: width, height and depth are, used for automatic target recognition of SAR (SAR-ATR). As shown in Fig. 17, in Krizhevsky et al. (2012), dropout regularization typically used in fully-connected layers was added to higher convolution layers, and

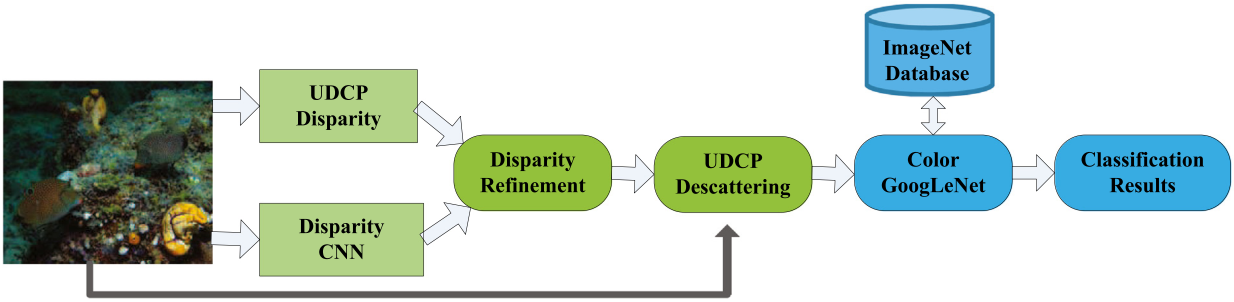

Fig. 13. FDCNet.

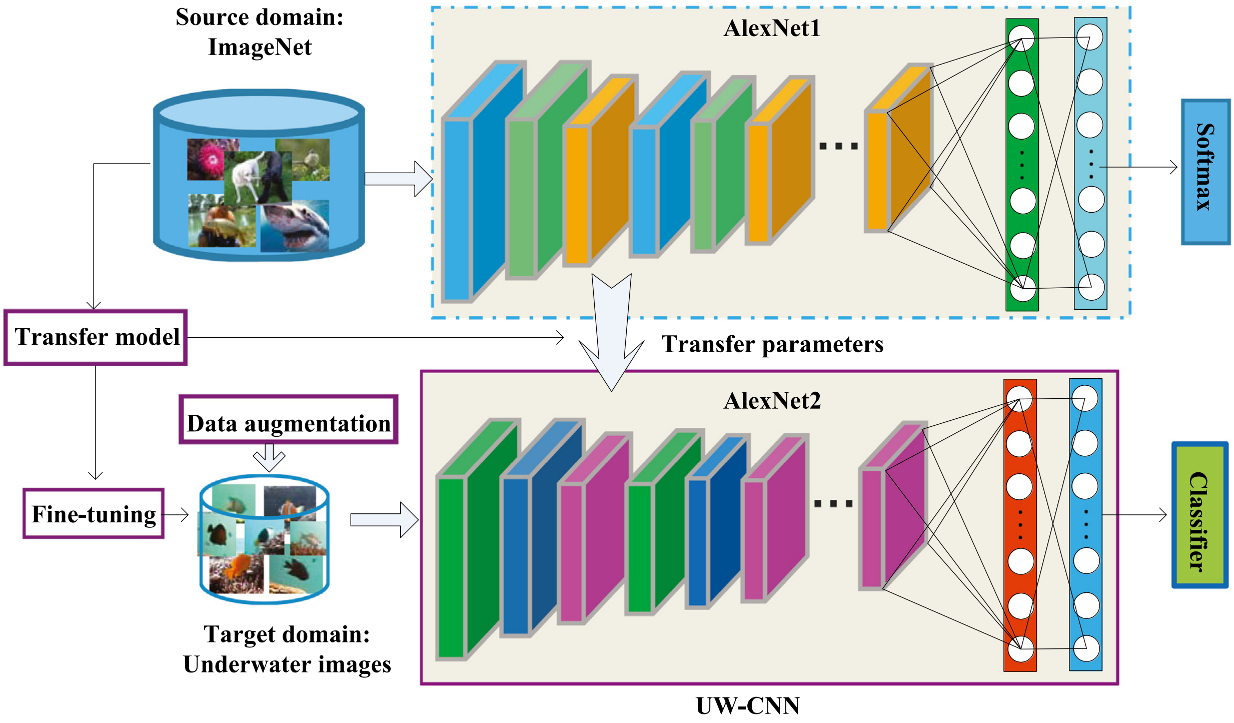

Fig. 14. Framework of the transferring network for underwater object recognition. (First, the AlexNet is pre-trained on a large source domain (ImageNet) which learns rich feature representations. Then, the parameters are transferred to the entire AlexNet structure as initialization on target domain.)

Table 3
Parameter statistics of AlexNet configuration.

	Convolutional kernel	Hyperparameters	Hyperparameters rate
Conv1	11 × 11 × 3, 96	11 × 11 × 3 × 96 + 96 = 34,944	
Conv2	5 × 5 × 48, 256	(5 × 5 × 48 × 128 + 128) × 2 = 307,456	
Conv3	3 × 3 × 256, 384	3 × 3 × 256 × 384 + 384 = 885,120	3.8% (2,334,080)
Conv4	3 × 3 × 192, 384	(3 × 3 × 192 × 192 + 192) × 2 = 663,936	
Conv5	3 × 3 × 192, 256	(3 × 3 × 192 × 128 + 128) × 2 = 442,624	
FC6	1 × 1, 4096	(6 × 6 × 128 × 2) × 4096 + 4096 = 37,752,832	
FC7	1 × 1, 4096	4096 × 4096 + 4096 = 16,781,312	96.2% (58,631,144)
FC8	1 × 1, 4096	4096 × 1000 + 1000 = 4,097,000	
Total	–	60,965,224	100%

can further reduce overfitting. In Chen et al. (2016), in the end-to-end SAR-ATR system, the first step is to detect potential targets and isolate the regions out from a complex background (e.g., sea surface). Secondly, it is used to feed these isolated image chips to a classifier, and ultimately declare the recognized target type. The A-ConvNets have achieved better classification accuracy than traditional ConvNets for SAR images. In addition, the antinoise performance of A-ConvNets has been approved.

(b) Deconvolutional operation

Affected by the AE reconstructing for input signals, in order to capture robust mid-level image cues which spontaneously emerge from image data, the encoder–decoder operation or deconvolutional structure was proposed by Zeiler et al. (2010). The DN is widely used in the procedures of object recognition, including unsupervised feature learning of mid and high-level image representations (Zeiler et al., 2012), visualization stage (Zeiler & Fergus, 2014), image generation,

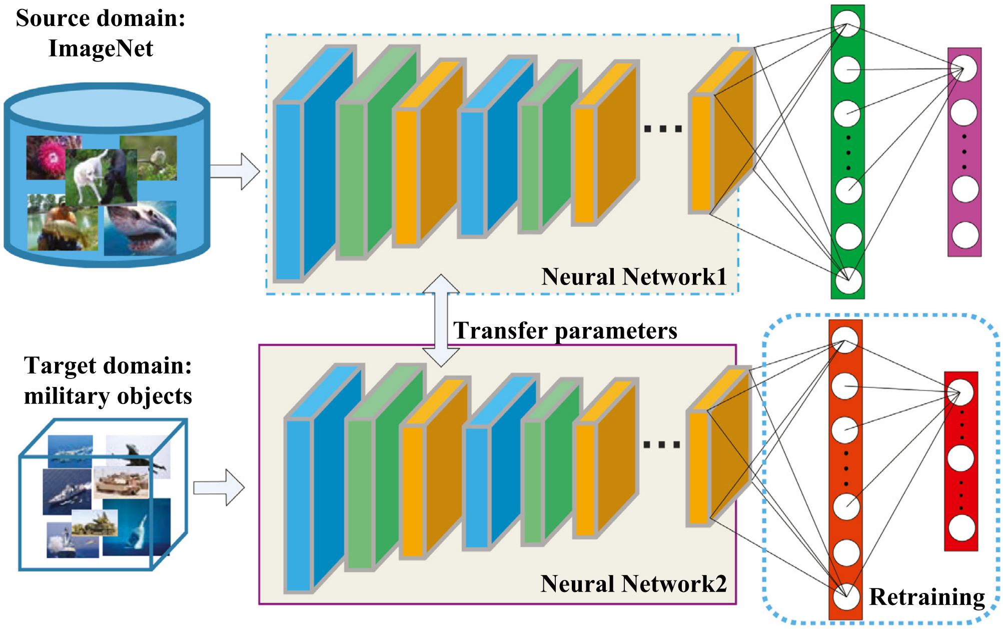

Fig. 15. Framework of the deep transferring network for military object recognition using small training set.

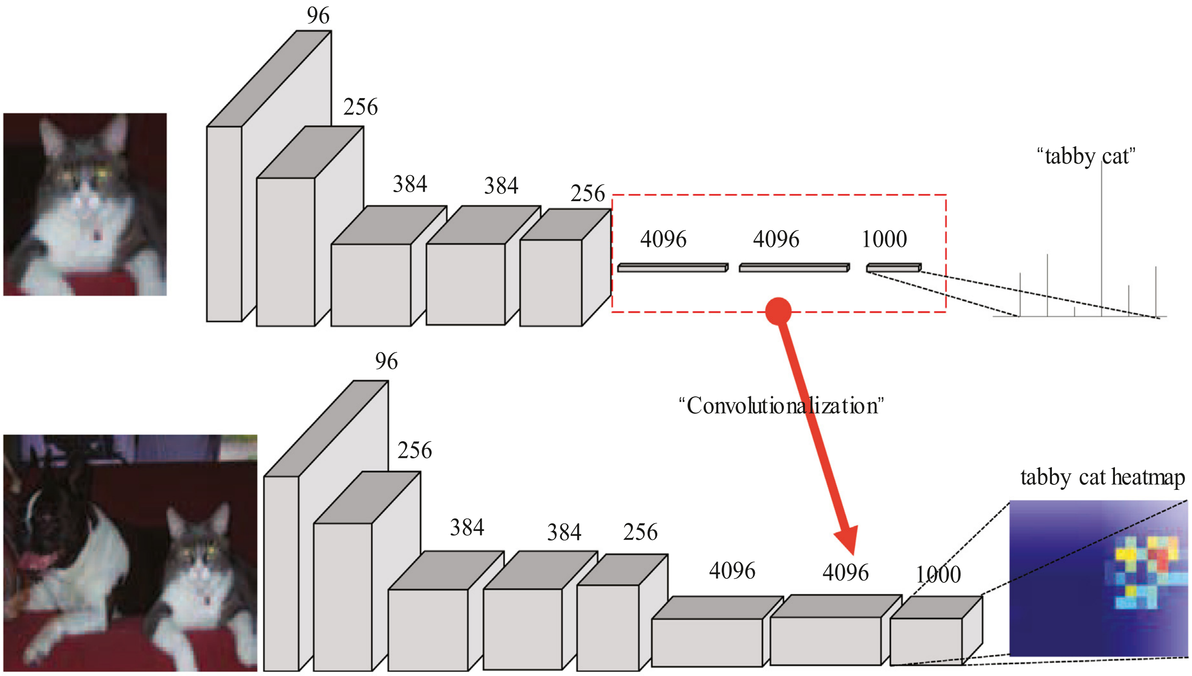

Fig. 16. FCN structure. (Transforming a classification-purposed CNN to produce spatial heatmaps by replacing fully-connected layers with convolutional ones.)

and image segmentation in the upsampling stage of the FCN (Long et al., 2015).

To solve the dimension disaster and low accuracy, ED-Net (Wang et al., 2019) aforementioned was designed for underwater object recognition by applying the deconvolution kernel with a matched feature map, instead of full connection. As shown in Fig. 18, underwater images or videos are transformed into deep features by two convolution layers, and then the deconvolution layers are used as decoders for refining the images or videos. In addition, the DA and transfer learning techniques were introduced to solve the problem of data starvation. The results show that ED-Net achieved significantly higher accuracy than the state-of-the-art methods UW-CNN (Sun et al., 2017), DeepFish (Qin, Xiu, Jian, Peng, & Zhang, 2016).

In Mao et al. (2016), the very deep residual encoder–decoder network (RED-Net) was proposed by incorporating the ED-Net into the residual network. As shown in Fig. 19, the RED-Net is composed of symmetric link of convolutional and deconvolutional layers, whereby convolutional layers act as feature extractors which preserve or encode the primary components of objects while eliminate corruptions in an image, and deconvolutional layers play a role in recovering the image content details by decoding the image abstraction. More importantly,

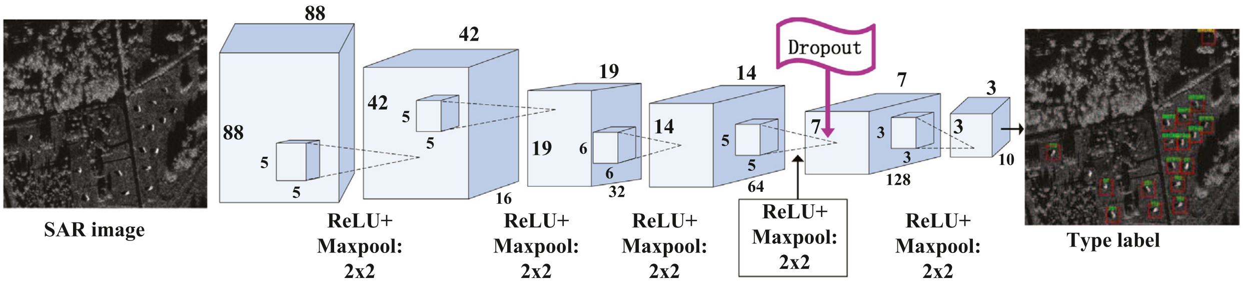

Fig. 17. Architecture of the A-ConvNets for SAR image object recognition. (Dropout is used in the fourth convolution layer.)

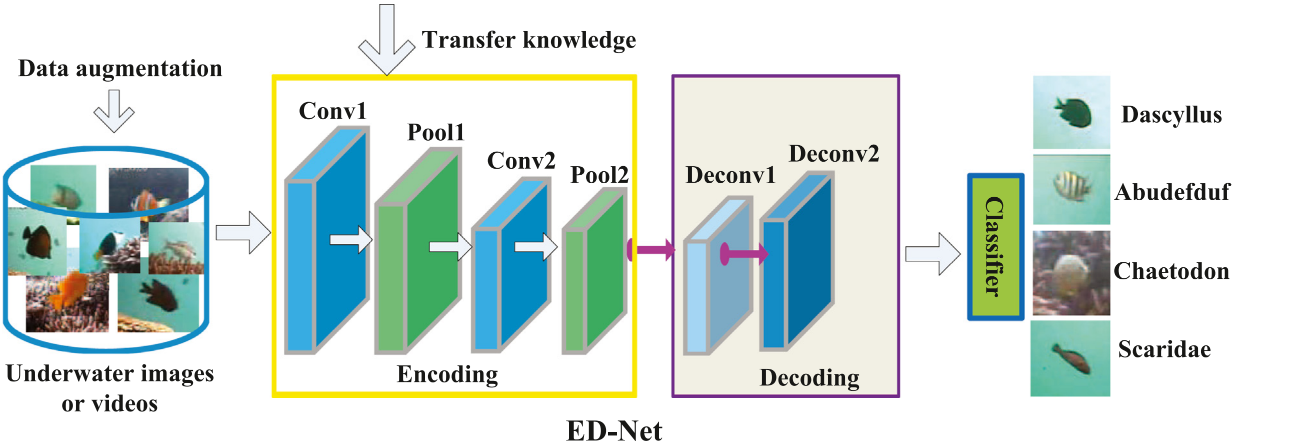

Fig. 18. Architecture of ED-Net for underwater object recognition.

Table 4
The recognition results of unsupervised methods on underwater acoustic data.

Model	PNN	SAE-Softmax	GRNN	SVM	DBN	SDAE
Accuracy %	92.5	94.12	94.2	96.2	96.8	98.2

the skip connections between convolutional and corresponding deconvolutional layers were built to backpropagate the signal to bottom layers directly. In this context, comprehensive image information was conveyed to the top layers, thereby benefiting in recovering the original images.

(c) DBN variants

As shown in Fig. 8, the DBN is constructed by stacking multiple RBMs and a classifier, and it is a classic hierarchical generative model that can be trained in a purely unsupervised manner.

Unlike the CNNs based on image signal features, the DBN model is good at extracting features from underwater radiated noises to recognize underwater targets. The radiated noises need to be built or modeled through a simulation model combining continual spectrum with line spectrum. In this context, the signal features extracted by the DBN act as input data of classifier. In Chen and Xu (2017), the DBN training is divided into two stages: pre-training and fine-tuning. In the pre-training stage, the entire DBN network is decomposed into multiple RBMs, and the unsupervised greedy algorithm is used. In the fine-tuning phase, the data with label are used to carry out back-propagation (BP) adjustment. As shown in Table 4, compared with the SVM, general regression neural network (GRNN), and probabilistic neural network (PNN), the DBN achieved the highest recognition accuracy than the first three methods.

(d) AE variants

The sparse AE is a symmetrical neural network that is used to learn invariant features from a data set in an unsupervised manner by minimizing the error between original inputs and reconstructed signals. Although we can obtain efficient recognition results with sparse AE, the extracted features are still low level due to the shallow network. In Hinton, Osindero and Teh (2006), to generate more complex and invariant features, similar to the DBN, the SAE is constructed by layer-by-layer stacking of input and hidden layers in an unsupervised greedy method. A softmax classifier is trained with the deep features extracted by the SAE model, and the SAE-Softmax framework (Xu, Zhang, Yang, & Niu, 2016) is proposed to classify underwater targets in an unsupervised manner. Finally, a fine-tuning technique with labeled data is utilized in the entire system to further improve model performance. The stacked DAE (SDAE) (Chen & Xu, 2017) was established by stacking the DAE whose input samples formed by adding noises into input samples can be reconstructed from the damaged ones. It is also used to recognize the underwater targets according to features extracted from underwater target radiated signals with the DBN. As shown in Table 4, the SDAE gained a higher recognition rate.

(e) Multideep models

In Aziz and Bouchara (2018), a multimodal architecture that makes use of joint synchronized visible and long-wave IR stream was proposed for ship classification. As shown in Fig. 20, two CaffeNets are separately applied for training the VIS and IR images, and the final fully-connected features among VIS and IR are fused to the classification layer.

The deep variance network (DVN) (Li, Song, Qin, & Hao, 2018) is constructed by incorporating a hierarchical Bayesian model into the powerful CNN framework, where Bayesian network is used to augment the distribution of small-scale datasets, and the CNN is used to learn deep features. This technique is effective for data starvation by transferring the joint feature distribution from complete training dataset for some objects to incomplete training dataset in other objects in an iterative way. As illuminated in Table 5, on ImageNet100 dataset, the DVN furnished VGGNet is superior to other DVN furnished deep networks including ResidualNet50, GoogLeNet, VGG19 and AlexNet.

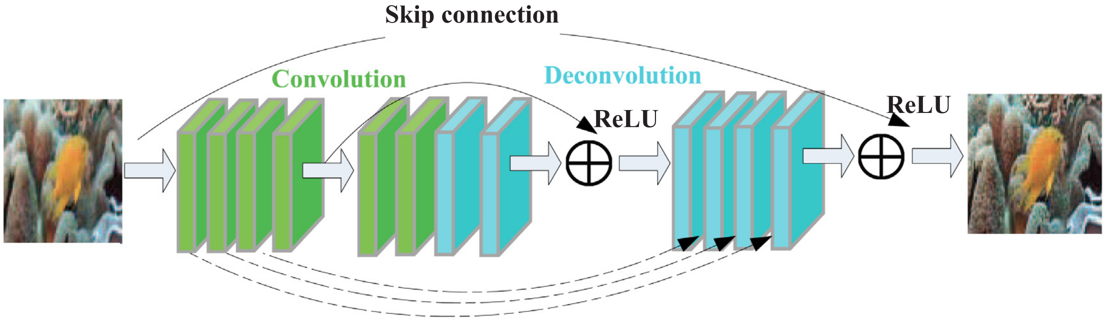

Fig. 19. RED-Net. (The cuboid in green and blue denote convolution and deconvolution respectively. ⊕ denotes element-wise sum of feature maps.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

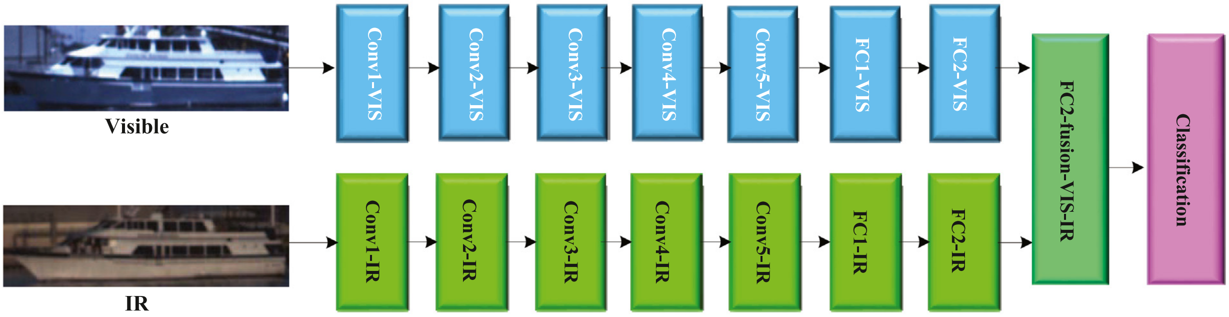

Fig. 20. Multimodal CNN for Visible and IR stream. (The top branch operates on Visible images and the bottom branch operates on IR images.)

Table 5
Performance comparisons of DVN-furnished networks on ImageNet100 dataset.

ImageNet100	ResidualNet50	GoogLeNet	VGG19	AlexNet
Original	94.21%	92.36%	98.29%	79.61%
DVN	96.73%	94.52%	98.49%	85.20%

More importantly, there is a significant improvement in the top-5 accuracy of the original deep networks. The model has exhibited excellent performance in various object recognition applications.

The fusion of statistical model and deep learning can be a good feature extractor. Applying Zemike moment and deep learning CNN for extracting the feature of ship image that is pre-processed by gray algorithm, denoised by wavelet transform and segmented by the Morphological Watershed algorithm, paper (Cao, Gao, Chen, & Wang, 2019) effectively identified 3 types of ships with an average detection accuracy of 87%. Based on the deep non-negative matrix factorization model, a multimodal structure was proposed in Guo and Chen (2017) by using the Poisson Gamma Belief network (PGBN) for feature selection in the field of SAR images and the Naive Bayes rule for acquiring label information in the training stage. The model termed NBPGBN, in comparison with other feature extraction methods such as PGBN, RBM, SDBN, DBN, SVM and KSVM, achieved a higher recognition accuracy of 92.67%.

In Lang and Wu (2017) and Lang, Wu, and Xu (2018), 11 naive geometric features extracted from vessel images are fused, and thereby several classifiers including the K-nearest neighbor (KNN) and cluster have been tested for ship recognition. In Zhao et al. (2016), combining the depth of ship image extracted by CNN with edge and color information gained by HOG and HSV algorithm, an ideal result has been achieved with average recognition accuracy of 93.6%. In Waxman et al. (1995), different types of image feature information from three categories of ship targets including SAR, IR and visible images are integrated to CNN or DNN for ship recognition.

The fusion recognition of multi-band images was proposed in Liu, Shen, Ma, Zhang (2017), whereby an improved AlexNet was designed to concurrently extract ship target features of three wave band images, including visible light, middle wave infrared (MWIR) and LWIR bands. In this context, feature selection and determination of fusion eigenvectors were performed by sorting the importance of concatenated feature eigenvectors. Moreover, the recognition accuracies of three fusion methods including early, middle and late fusion are discussed. As a result, the recognition rate from middle fusion is higher than that of the others, and all of them outperformed the single-band recognition in the same application scenes. In addition, in Qiu, Gong, Ma, and Sun (2018), 5 medium-wave infrared images are collected to form an image dataset. Firstly, Dense SIFT feature of each infrared image was extracted. Secondly, principal component analysis (PCA) was applied to each SIFT feature to reduce dimension. The spatial and spectral position information of each SIFT feature was then integrated into the feature vector. Based on the Gaussian mixture model, the feature vectors of the multi-spectral images were encoded to obtain the Fisher vector representing the target. Finally, the linear SVM classifier was used to identify Fisher vector and further to recognize the ship target. It achieved a high recognition rate of 97%.

4.3. Recognition and model optimization

Apart from inserting deconvolutional layers and replacing fully-connected layers, the sophisticated network designing techniques consist of selecting activation functions and classifiers. Some of them can avoid overfitting.

The DA (Krizhevsky et al., 2012) can be used to enlarge dataset, and transfer learning is a model transformation method. In fact, all of them play a role in reducing overfitting. Besides, “dropout (Bell & Koren, 2007)” is also an important way to avoid overfitting for object recognition model and it is mostly applied in the fully-connected layer and sometimes in the convolutional layer.

The selection of non-saturating nonlinear activation functions including sigmoid, softmax, tanh, ReLU and variants, and radial basis function (Karlik & Olgac, 2011), is also a crucial step for a deep network. In general, softmax is often-used in the fully-connected layers, while ReLU and its variants are used in feature extraction. The training speed of the DNNs together with ReLU is usually much faster than their equivalents with saturated nonlinearity, which easily suffer from vanishing gradient. In this context, some variants of the ReLU including Elastic ReLU (EReLU), Parametric ReLu (PReLU) (He, Zhang, Ren, & Sun, 2015), and Elastic parametric ReLU (EPReLU) were proposed to further improve the performance of networks in He et al. (2015). It is verified that these new activation functions based on the AlexNet performed better than ReLU on ImageNet 2012 dataset.

In order to make classification and detection suitable for object recognition, the corresponding and vital attributions such as powerful network structures (Liu et al., 2017), good training strategies and effective techniques against overfitting should be further focused. Training strategies, including online sequential learning (Wang et al., 2009), stochastic gradient descent (SGD) (Lecun et al., 1998), non-saturating nonlinear activation function and BN (Ioffe & Szegedy, 2015), have proven to be effective ways of training deep networks.

The selection of classifiers is a vital step for object recognition. The most common classifiers have gained excellent results, for instance, KNN, softmax, SVM and extreme learning machine (ELM) (Huang, Zhou, Ding, & Zhang, 2012; Wang, Han, Dong and Er, 2014; Zhang, He, & Liu, 2017) and variants (e.g. complex-valued ELM (CELM) (Zhang, Wang, Xu, Wang, & Xu, 2018), parsimonious ELM (Wang, Er and Han, 2014)), and logistic regression classifiers (Ma, Khatibisepehr, & Huang, 2015; Zou et al., 2015). In addition, as can be seen from Table 6, the UW-CNN with classifier SVM, in comparison of that with Softmax and the model fine-tuned in last layer, gain the highest recognition accuracy under the same other conditions. While the comparisons of ED-Net with classifier Softmax and SVM, the latter performed better. In Ji, Xing, Chen, Zou, and Chen (2013), the ship classification methods based on classifiers combination (SVMs, BPs) and individual classifiers (KNN, Bayes, and BP neural network classifier) were investigated on TerraSAR-X SAR ship images. The recognition accuracy of combination strategies was superior to that of individual classifiers, in particular, the SVM combination strategy achieved a high classification rate.

5. Discussion

In this section, recognition results of mostly considered methods have been summarized in Table 6. Noted that "I-Net", "A-Net", "G-Net", "TL", "Soft", "4-VDCNN", "2C2P1F", "PFA" and "NB" denote "ImageNet", "AlexNet", "GoogLeNet", "Transfer Learning", "Softmax function", "4-Views Deep CNN", "layers of two Convolutional, two Pooling, one FC", "Poisson Factor Analysis" and "Naive Bayes rule", respectively.

It can be seen from Table 6, regardless of whether it is underwater or surface target recognition, deep learning methods including supervised and unsupervised learning can be used. The main difference is the environments capturing object images and videos, wherein the rain, snow and fog affect the clarity of the surface image, and the tremor of shooting equipment caused by wind and waves, whereas low-light and high-noise scenarios pose challenges for underwater video analysis.

Samples available in small amounts has been a common issue for marine object recognition. The DA and transfer learning with fine-tuning are use-often in supervised CNNs to relieve data starvation, especially DA. However, some models without transfer learning can achieve a high recognition accuracy, such as supervised learning methods used in Aziz and Bouchara (2018), Chen et al. (2016), Guo and Chen (2017), Pei et al. (2017) and Zhao et al. (2016), and unsupervised learning methods including DBN, SAE and SDAE, etc., as listed in Table 6. In addition, differ from surface object recognition which is mostly based on images, underwater objects can be recognized based on videos, target radiated (Chen & Xu, 2017) and acoustic noises (Xu et al., 2016). It is noted that noise model need to be well constructed to extract the target signal features. In addition, applying only deep learning architectures for marine object recognition is ineffective to some extent, some auxiliary prior knowledge are needed for better recognition accuracy, especially for subclass or fine-grained recognition.

In order to validate the recognition performances of the state-of-the-art models on the same marine objects dataset, we have given the experimental results about Inceptionv3 and Mobilnetv1 on the boats type recognition dataset provided by CCF 2016, due to the limitation of paper length (see Fig. 21).

Remark 2. The ability of deep learning to process big data can meet the urgent requirements of fast and accurate analysis of marine big data, except for marine object, it can solve a series of marine problems such as marine disaster prevention and mitigation, ecological environmental protection, marine object detection and tracking, and emergency rescue.

6. Future works and challenges

Despite the rapid development of deep learning for marine object recognition, some outstanding issues still exist. Based on the literature review, we list the issues and challenges here for completeness.

6.1. *Public marine dataset*

At present, there are no widely accepted and open datasets as well as label sets that can be used to detect and recognize the marine objects. As foreshadowed, it is common to train a large-scale database (e.g. ImageNet, CIFAR) in the pre-training stage, and subsequently using the new datasets collected by authors to fine tun the pre-trained model that is obtained in the first step. Moreover, practical datasets are private and vary from different works even for the same task. Therefore, it is very cumbersome to compare the performance of various methods without a public dataset and the creation of a public marine dataset is highly desired.

6.2. *Necessity of pre-training*

Currently, many works that use the deep learning methods to tackle the issues of marine object recognition are based on "transfer learning", i.e., applying the classic image database such as ImageNet, COCO to train the network, and then putting the new dataset of marine objects into the pre-trained network and properly tuning it so that it can be applied in practice. However, the operation is time consuming and difficult to perform, especially in the processing of pre-training. Therefore, a question of whether it is essential to pre-train with the classic database was raised and the answer can be focused in He, Girshick, and Dollár (2019). ImageNet pre-training can speed up convergence early in training, but in essence, does not necessarily provide regularization or improve the final target task accuracy. These findings encourage to rethink the necessity of pre-training and fine-tuning.

6.3. *The need of a unified framework*

The existing deep learning models based on images, for instance, AlexNet, GoogLeNet and VGG, mostly need a large amount of high-quality images and/or videos as their inputs, since the high-quality images or videos usually carry more discriminant features. However, the influence of widely varying environmental conditions, such as fog, lighting, rain, wave occlusions, sophisticated background, as well as view angle and range is the most challenging issue for obtaining high-quality images and/or videos (Ma, Wen, & Liang, 2013). Thus, subsequent works to construct a unified framework or stream are highly needed, by integrating three steps: image preprocessing, feature extraction and classification for marine object recognition task such that all the images obtained by camera or other equipments can directly be fed into models. It is conjectured that the new framework will significantly reduce the requirements for images.

Table 6
Comparisons of marine object recognition methods.

Methods	Target	Source domain	Model pretrain	Dataset	TL	DA	Mean accuracy
UW-CNN-Last (Sun et al., 2017)	Fish	I-Net	A-Net	Fish 4K	✓	✓	89.5%
DeepFish (Qin et al., 2016)	Fish	I-Net	A-Net	Fish 4K	✓	✓	90.1%
ED-Net-Soft (Wang et al., 2019)	Fish	I-Net	A-Net	Fish 4K	✓	✓	95.83%
UW-CNN-Soft (Sun et al., 2017)	Fish	I-Net	A-Net	Fish 4K	✓	✓	97.1%
ED-Net-SVM (Wang et al., 2019)	Fish	I-Net	A-Net	Fish 4K	✓	✓	99.36%
UW-CNN-SVM (Sun et al., 2017)	Fish	I-Net	A-Net	Fish 4K	✓	✓	99.68%
FDCNet (Lu et al., 2018)	Fish	I-Net	G-Net	Kyutech10K	✓	✗	92.0%
CNN (Meng et al., 2018)	Fish	–	A-Net	Panoramic images(360)	✗	✓	87.0%
SAE-Soft (Xu et al., 2016)	Under-water	–	AE	Acoustic signal	✗	✗	94.12%
DBN (Chen & Xu, 2017)	Under-water	–	DBN	Target radiated noise	✗	✗	96.8%
SDAE (Chen & Xu, 2017)	Under-water	–	AE	Target radiated noise	✗	✗	98.2%
Improved AlexNet (Liu, Shen et al., 2017)	Ship	I-Net	A-Net	Self-built	✓	✗	90.1%
CNN(2C2P1F) (Zhao et al., 2016)	Ship	–	–	FleetMon	✗	✗	93.55%
Multimodal CNN (Aziz & Bouchara, 2018)	Ship	–	Two CaffeNets	VAIS	✗	✓	86.67%
PGBN (Guo & Chen, 2017)	Ship	–	PFA+DBN	MSTAR	✗	✓	89.16%
NBPGBN (Guo & Chen, 2017)	Ship	–	PGBN+NB	MSTAR	✗	✓	93.85%
4-VDCNN(4C3P1F) (Pei et al., 2017)	Ship	–	–	MSTAR	✗	✗	98.52%
A-ConvNets(5C3P) (Chen et al., 2016)	Ship	–	–	MSTAR	✗	✗	99.13%
DVN-furnished VGG (Li et al., 2018)	Ship	I-Net	VGG19	Imagenet100	✓	✗	98.49%
CNN-Last3(6C1P) (Yang et al., 2018)	Ship	I-Net	CNN	Military objects	✓	✓	91.1%

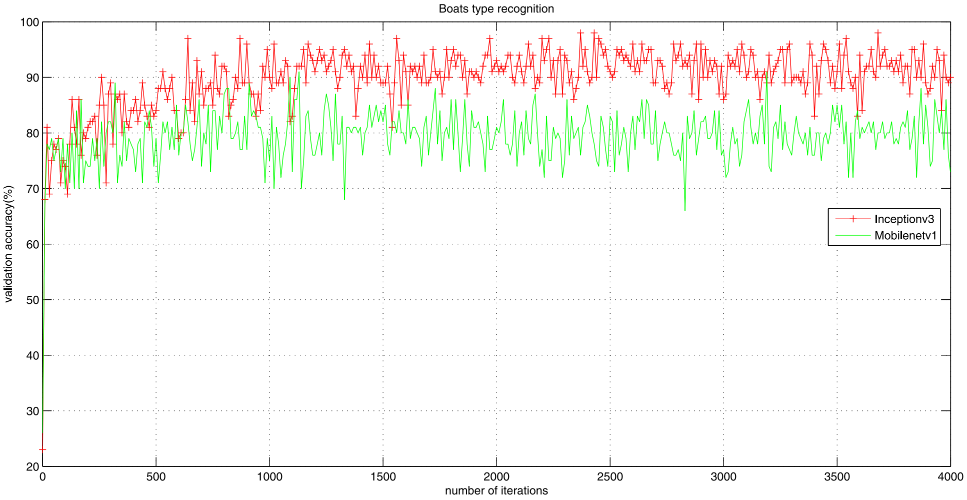

Fig. 21. Validation accuracy for different models on boats type recognition dataset.

6.4. Fusion of multi-sources features

For effective recognition of marine objects, images can be obtained by various platforms such as shore, shipboard, airborne or spaceborne, and the object image features can be extracted from Radar, Infrared, Electronic Support Measure (ESM) and Automatic Identification System (AIS) information. One critical issue is how to effectively combine different image representations for accurate object recognition.

Existing works consist of fusion of visible and infrared image features (Aziz & Bouchara, 2018), which have achieved a good recognition result. Therefore, we can take advantage of these sensor platforms and multi-sources features (e.g. radar, infrared, electronic detection and AIS) to comprehensively recognize marine targets. In addition, this approach can solve the issue of multi-source spatio-temporal matching association and heterogeneous feature fusion for the same target and overcome the uncertainty of characteristics of single sensor target.

Multiple kernel learning (MKL) (Bucak, Jin, & Jain, 2013) is also useful tool for object recognition, where each image is represented by multiple sets of features and MKL is applied to combine different feature sets.

On top of this, the multi-view, different kinds and sizes of one or more targets acquired by two or more of these platforms can also be fused, MKL may be applied and finally realize high-accuracy recognition of marine targets.

6.5. Fusion of multi-deep model

Some multi-model mentioned in this paper have been achieved excellent results in Section 4.2.2(e) for marine ships recognition. Others applied in the other area can be extend to the marine object recognition, for example, the convolutional-recursive deep learning model (Richard et al., 2012) for raw RGB-D images object recognition maybe well used in marine targets.

6.6. Sub-class recognition

Sub-class recognition is not uncommon in natural image domain, for instance, fine-grained categorization. Fine-grained ship recognition can be applied for recognition of the domestic and foreign military objects, and it is conjectured that the research results will provide valuable guidelines for future naval battles.

6.7. The general model structure

At present, there are no general theories of selecting a perfect deep learning architecture for different recognition tasks, which in turn often have a close relationship with the specific applications and datasets.

Many intelligent deep learning methods which have been successfully deployed in object detection, semantic segmentation and some land object recognition (e.g. animals, humans, planes, cars) can also be extended to the recognition task of marine targets, for instance, 3D object recognition (Ouadiay, Zrira, Bouyakhf, & Himmi, 2016; Xiang et al., 2016). In addition, some effective and powerful tools on control learning systems, i.e., fuzzy logic methods, fuzzy-affine-model (Wei, Qiu and Karimi, 2018; Wei, Qiu, Shi and Ligang, 2018), fault-tolerant control methods (Qin, Chen and Sun, 2019; Qin, Chen, Sun and Chen, 2019; Qin, Yu, Zhu, & Deng, 2020) can used in future works and expecting results.

7. Conclusions

Due to powerful learning ability and efficiency in data preprocessing, deep-learning-based marine object recognition has been a research hotspot in recent years. A detailed review for learning-based marine object recognition on most popular techniques and typical deep network frameworks are provided by three parts including image preprocessing, feature extraction, and recognition and model optimization. Various subproblems in marine object recognition have been comprehensively reviewed and compared, including resolution (low, moderate and high) problem of images, samples starvation in image and video data, complex marine environmental factors, different degrees of model architectures, and optimizations based on the supervised and unsupervised learning models. Finally, several outstanding issues that need to be solved and meaningful works in the futuristic marine object recognition landscape have been identified and discussed, providing valuable insights and guidelines for the researchers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Editor-in-Chief, Associate Editor and anonymous referees for their invaluable comments. This work is supported by the National Natural Science Foundation of P.R. China (under Grants 51009017 and 51379002), the Fund for Liaoning Innovative Talents in Colleges and Universities, China (under Grant LR2017024), the Fund for Dalian Distinguished Young Scholars, China (under Grant 2016RJ10), the Liaoning Revitalization Talents Program (under Grant XLYC1807013), the Stable Supporting Fund of Science and Technology on Underwater Vehicle Laboratory, China (SXJQR2018WDKT03), and the Fundamental Research Funds for the Central Universities, China (under Grant 3132019344).

References

Alberto, G. G., Sergio, O. E., Sergiu, O., Victor, V. M., & Jose, G. R. (2017). A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857.

Anagnostopoulos, C. N., Anagnostopoulos, I., Ioannis, P., Loumos, V., & Kayafas, E. (2008). License plate recognition from still images and video sequences: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *9*(3), 377–391.

Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587.

Aziz, K., & Bouchara, F. (2018). Multimodal deep learning for robust recognizing maritime imagery in the visible and infrared spectrums. In *International conference image analysis and recognition* (pp. 235–244).

Bejiga, M. B., Zeggada, A., & Melgani, F. (2016). Convolutional neural networks for near real-time object detection from UAV imagery in avalanche search and rescue operations. In *IGARSS 2016 - 2016 IEEE international geoscience and remote sensing symposium* (pp. 1–4).

Bell, R. M., & Koren, Y. (2007). Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, *9*(2), 75–79.

Borkar, A., Hayes, M., & Smith, M. T. (2012). A novel lane detection system with efficient ground truth generation. *IEEE Transactions on Intelligent Transportation Systems*, *13*(1), 365–374.

van den Broek, S. P., Bouma, H., Hollander, R. J. M. D., Veerman, H. E. T., Benoist, K. W., & Schwering, P. B. W. (2014). Ship recognition for improved persistent tracking with descriptor localization and compact representations. In *Electro-optical and infrared systems: Technology and applications XI* (p. 92490N).

Bucak, S. S., Jin, R., & Jain, A. K. (2013). Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1354–1369.

Cao, X., Gao, S., Chen, L., & Wang, Y. (2019). Ship recognition method combined with image segmentation and deep learning feature extraction in video surveillance. *Multimedia Tools and Applications*, *1*.

Chen, S., Wang, H., Feng, X., & Jin, Y. Q. (2016). Target classification using the deep convolutional networks for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(8), 1–12.

Chen, Y., & Xu, X. (2017). The research of underwater target recognition method based on deep learning. In *2017 IEEE international conference on signal processing, communications and computing (ICSPCC)* (pp. 1–5).

Corbane, C., Najman, L., Pecoul, E., Demagistri, L., & Petit, M. (2010). A complete processing chain for ship detection using optical satellite imagery. *International Journal of Remote Sensing*, *31*(22), 5837–5854.

Corbane, C., Pecoul, E., Demagistri, L., & Petit, M. (2008). Fully automated procedure for ship detection using optical satellite imagery. In *Remote sensing of inland, coastal, and oceanic waters 10* (pp. 1–13).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Daoduc, C., Xiaohui, H., & Morère, O. (2015). Maritime vessel images classification using deep convolutional neural networks. In *Proceedings of the sixth international symposium on information and communication technology* (pp. 276–281).

Ducournau, A., & Fablet, R. (2016). Deep learning for ocean remote sensing: An application of convolutional neural networks for super-resolution on satellite-derived SST data. In *9TH workshop on pattern recognition in remote sensing* (pp. 1–6).

Duo, Z., Wang, W., & Wang, H. (2019). Oceanic mesoscale eddy detection method based on deep learning. *Remote Sensing*, *11*(16), 1921.

Erhan, G., Berkan, S., Veysel, Y., & Aykut, K. (2016). MARVEL: A large-scale image dataset for maritime vessels. In *Asian conference on computer vision* (pp. 165–180).

Fu, G., Liu, C., Zhou, R., Sun, T., & Zhang, Q. (2017). Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sensing*, *9*(5), 498–519.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868.

Guo, D., & Chen, B. (2017). SAR image target recognition via deep Bayesian generative network. In *2017 international workshop on remote sensing with intelligent processing* (pp. 1–4).

Guo, W., & Xia, X. (2017). A ship recognition method of variational inference-based probability generative model using optical remote sensing image. *Optik*, *145*, 365–376.

Guo, W., Xia, X., & Wang, X. (2014). A remote sensing ship recognition method based on dynamic probability generative model. *Expert Systems with Applications*, *41*(14), 6446–6458.

He, K., Girshick, R., & Dollár, P. (2019). Rethinking imagenet pre-training. In *Proceedings of the IEEE international conference on computer vision* (pp. 4918–4927).

He, K., Zhang, X., Ren, S., & Jian, S. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G. E., Osindero, S., Welling, M., & Teh, Y. W. (2006). Unsupervised discovery of nonlinear structure using contrastive backpropagation. *Cognitive Science*, *30*(4), 725–731.

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, G., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *42*(2), 513–529.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Ji, K., Xing, X., Chen, W., Zou, H., & Chen, J. (2013). Ship classification in TerraSAR-X SAR images based on classifier combination. In *2013 IEEE international geoscience and remote sensing symposium-IGARSS* (pp. 2589–2592).

Jin, L., & Hong, L. (2017). Deep learning for underwater image recognition in small sample size situations. In *OCEANS 2017 - Aberdeen* (pp. 1–4).

Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural network. *International Journal of Artificial Intelligence and Expert System*, *1*, 111–122.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*(2), 1097–1105.

Kumlu, D., & Jenkins, B. (2013). Autonomous ship classification using synthetic and real color images. In *Image processing: Machine vision applications VI* (p. 86610M).

Lang, H., & Wu, S. (2017). Ship classification in moderate-resolution SAR image by naive geometric features-combined multiple kernel learning. *IEEE Geoscience and Remote Sensing Letters*, *14*(10), 1765–1769.

Lang, H., Wu, S., & Xu, Y. (2018). Ship classification in SAR images improved by AIS knowledge transfer. *IEEE Geoscience and Remote Sensing Letters*, *15*(3), 439–443.

Lecun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lecun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Lecun, Y., Boser, B., Denker, J. S., Henderson, D., & Jackel, L. D. (1997). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, *2*(2), 396–404.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, S., Song, W., Qin, H., & Hao, A. (2018). Deep variance network: An iterative, improved CNN framework for unbalanced training datasets. *Pattern Recognition*, *81*, 294–308.

Lines, J. A., Tillett, R., Ross, L. G., Chan, D., Hockaday, S., & McFarlane, N. (2001). An automatic image-based system for estimating the mass of free-swimming fish. *Computers and Electronics in Agriculture*, *31*(2), 151–168.

Liu, F., Shen, T., Ma, X., & Zhang, J. (2017). Ship recognition based on multi-band deep neural network. *Optics and Precision Engineering*, *25*(11), 2939–2946.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11–26.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). SphereFace: Deep hypersphere embedding for face recognition. In *The IEEE conference on computer vision and pattern recognition* (pp. 212–220).

Liu, X., Yu, J., & Lv, J. (2011). Ship image recognition method based on the affine invariant moments. *Journal of Naval Aeronautical and Astronautical University*, *26*(6), 687–690.

Liu, Z., Yuan, L., Weng, L., & Yang, Y. (2017). A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods* (pp. 324–331).

Liu, Z., Zhang, Y., Yu, X., & Yuan, C. (2016). Unmanned surface vehicles: An overview of developments and challenges. *Annual Reviews in Control*, *41*, 71–93.

Liu, G., Zhang, Y., Zheng, X., Sun, X., Fu, K., & Wang, H. (2013). A new method on inshore ship detection in high-resolution satellite images using shape and context information. *IEEE Geoscience and Remote Sensing Letters*, *11*(3), 617–621.

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

Lu, H., Li, Y., Uemura, T., Ge, Z., Xu, X., He, L., et al. (2018). FDCNet: filtering deep convolutional network for marine organism classification. *Multimedia Tools and Applications*, *77*(17), 21847–21860.

Luo, Y., & Wan, Y. (2013). A novel efficient method for training sparse auto-encoders. In *2013 6th international congress on image and signal processing* (pp. 1019–1023).

Ma, M., Khatibisepehr, S., & Huang, B. (2015). A Bayesian framework for real-time identification of locally weighted partial least squares. *AIChE Journal*, *61*(2), 518–529.

Ma, Z., Wen, J., & Liang, X. (2013). Video image clarity algorithm research of USV visual system under the sea fog. In *International conference in swarm intelligence* (pp. 436–444).

Malmgren-Hansen, D., Kusk, A., Dall, J., Nielsen, A. A., Engholm, R., & Skriver, H. (2017). Improving SAR automatic target recognition models with transfer learning from simulated data. *IEEE Geoscience and Remote Sensing Letters*, *14*(9), 1484–1488.

Mao, X., Shen, C., & Yang, Y. (2016). Image restoration using very deep fully convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems* (pp. 2802–2810).

Meng, L., Hirayama, T., & Oyanagi, S. (2018). Underwater-drone with panoramic camera for automatic fish recognition based on deep learning. *IEEE Access*, *6*, 17880–17886.

Mian, A. S., Bennamoun, M., & Owens, R. (2005). 3D model-based free-form object recognition - A review. *Sensor Review*, *25*(2), 148–154.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807–814).

Ødegaard, N., Knapskog, A. O., Cochin, C., & Louvigne, J.-C. (2016). Classification of ships using real and simulated data in a convolutional neural network. In *2016 IEEE radar conference (RadarConf)* (pp. 1–6).

Oh, J., Kim, G., Nam, B. G., & Yoo, H. J. (2013). A 57 mW 12.5 μJ/Epoch embedded mixed-mode neuro-fuzzy processor for mobile real-time object recognition. *IEEE Journal of Solid-State Circuits*, *48*(11), 2894–2907.

Ouadiay, F. Z., Zrira, N., Bouyakhf, E. H., & Himmi, M. M. (2016). 3D object categorization and recognition based on deep belief networks and point clouds. In *Proceedings of the 13th international conference on informatics in control, automation and robotics* (pp. 311–318).

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Pei, J., Huang, Y., Huo, W., Zhang, Y., Yang, J., & Yeo, T.-S. (2017). SAR automatic target recognition based on multiview deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(4), 2196–2210.

Pöyhönen, S., Arkkio, A., Jover, P., & H., H. (2005). Coupling pairwise support vector machines for fault classification. *Control Engineering Practice*, *13*(6), 759–769.

Proia, N., & Page, V. (2010). Characterization of a Bayesian ship detection method in optical satellite images. *IEEE Geoscience and Remote Sensing Letters*, *7*(2), 226–230.

Qi, S., Ma, J., Lin, J., Li, Y., & Tian, J. (2015). Unsupervised ship detection based on saliency and S-HOG descriptor from optical satellite images. *IEEE Geoscience and Remote Sensing Letters*, *12*(7), 1451–1455.

Qin, H., Chen, H., & Sun, Y. (2019). Distributed finite-time fault-tolerant containment control for multiple ocean bottom flying nodes. *Journal of the Franklin Institute*.

Qin, H., Chen, H., Sun, Y., & Chen, L. (2019). Distributed finite-time fault-tolerant containment control for multiple ocean bottom flying node systems with error constraints. *Ocean Engineering*.

Qin, H., Xiu, L., Jian, L., Peng, Y., & Zhang, C. (2016). DeepFish: Accurate underwater live fish recognition with a deep architecture. *Neurocomputing*, *187*, 49–58.

Qin, H., Yu, X., Zhu, Z., & Deng, Z. (2020). An expectation-maximization based single-beacon underwater navigation method with unknown ESV. *Neurocomputing*, *378*, 295–303.

Qiu, R., Gong, J., Ma, X., & Sun, C. (2018). Ship target recognition based on multi-spectral infrared images. In *Tenth international conference on digital image processing* (p. 108060Q).

Rajurkar, R. C. (2015). Visual object recognition using image mining approach: A review. *International Journal of Computer and Communication Engineering Research*, *3*(2), 28–30.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Richard, S., Brody, H., Bharath, B., Christopher, D. M., & Andrew, Y. N. (2012). Convolutional-recursive deep learning for 3D object classification. In *Advances in neural information processing systems 25*.

Rifai, S., Vincent, P., Muller, X., Glorot, X., & Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning* (pp. 833–840).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sharma, A., Singh, P. K., & Khurana, P. (2016). Analytical review on object segmentation and recognition. In *2016 6th international conference - Cloud system and big data engineering (Confluence)* (pp. 524–530).

Sherrah, J. (2016). Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv preprint arXiv:1606.02585.

Shi, Z., Yu, X., Jiang, Z., & Li, B. (2014). Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(8), 4511–4523.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *IEEE international conference on learning representations* (pp. 730–734).

Singh, R. D., Mittal, A., & Bhatia, R. K. (2019). 3D convolutional neural network for object recognition: A review. *Multimedia Tools and Applications*, *78*(12), 15951–15995.

Sommer, L., Schumann, A., Muller, T., Schuchert, T., & Beyerer, J. (2017). Flying object detection for automatic UAV recognition. In *2017 14th IEEE international conference on advanced video and signal based surveillance* (pp. 1–6).

Sun, X., Shi, J., Liu, L., Dong, J., Plant, C., Wang, X., et al. (2017). Transferring deep knowledge for object recognition in low-quality underwater videos. *Neurocomputing*, *275*, 897–908.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Tang, J., Deng, C., Huang, G., & Zhao, B. (2014). Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(3), 1174–1185.

Teutsch, M., & Krüger, W. (2010). Classification of small boats in infrared images for maritime surveillance. In *2010 international waterside security conference* (pp. 1–7).

Vilches, E., Escobar, I. A., Vallejo, E. E., & Taylor, C. E. (2006). Data mining applied to acoustic bird species recognition. In *International conference on pattern recognition* (pp. 400–403).

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on machine learning* (pp. 1096–1103).

Wang, N., Er, M. J., & Han, M. (2014). Parsimonious extreme learning machine using recursive orthogonal least squares. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(10), 1828–1841.

Wang, N., Er, M. J., & Han, M. (2015a). Generalized single-hidden layer feedforward networks for regression problems. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(6), 1161–1176.

Wang, N., Er, M. J., & Han, M. (2015b). Large tanker motion model identification using generalized ellipsoidal basis function-based fuzzy neural networks. *IEEE Transactions on Cybernetics*, *45*(12).

Wang, N., Er, M. J., & Meng, X. (2009). A fast and accurate online self-organizing scheme for parsimonious fuzzy neural networks. *Neurocomputing*, *72*(16–18), 3818–3829.

Wang, N., Han, M., Dong, N., & Er, M. J. (2014). Constructive multi-output extreme learning machine with application to large tanker motion dynamics identification. *Neurocomputing*, *128*, 59–72.

Wang, X., Ouyang, J., Li, D., & Zhang, G. (2019). Underwater object recognition based on deep encoding-decoding network. *Journal of Ocean University of China*, *18*(2), 376–382.

Wang, N., Sun, J., & Liu, Y. (2016). Direct adaptive self-structuring fuzzy control with interpretable fuzzy rules for a class of nonlinear uncertain systems. *Neurocomputing*, *173*, 1640–1645.

Wang, C., & Wang, L. (2011). Ship targets recognition algorithm based on features. *Application Research of Computers*, *28*(6), 2352–2354.

Waxman, A. M., Seibert, M. C., Gove, A., Fay, D. A., Bernardon, A. M., Lazott, C., et al. (1995). Neural processing of targets in visible, multispectral IR and SAR imagery. *Neural Networks*, *8*(7–8), 1029–1051.

Wei, Y., Qiu, J., & Karimi, H. R. (2018). Fuzzy-affine-model-based memory filter design of nonlinear systems with time-varying delay. *IEEE Transactions on Fuzzy Systems*, *1*.

Wei, Y., Qiu, J., Shi, P., & Ligang, W. (2018). A piecewise-markovian lyapunov approach to reliable output feedback control for fuzzy-affine systems with time-delays and actuator faults. *IEEE Transactions on Cybernetics*, *48*(9), 2723–2735.

Withagen, P. J., Schutte, K., Vossepoel, A. M., & Breuers, M. G. J. (1999). Automatic classification of ships from infrared (FLIR) images. In *Signal processing, sensor fusion, and target recognition VIII* (pp. 180–187).

Xavier, G., Antoine, B., & Yoshua, B. (2011). Deep sparse rectifier neural networks. In *Proceedings of the 14th international conference on artificial intelligence and statistics* (pp. 315–323).

Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., et al. (2016). Objectnet3D: A large scale database for 3D object recognition. In *European conference on computer vision* (pp. 160–176).

Xiong, Y., Ding, S., Deng, C., Fang, G., & Gong, R. (2018). Ship detection under complex sea and weather conditions based on deep learning. *Journal of Computer Applications*.

Xu, J., Sun, X., Zhang, D., & Fu, K. (2014). Automatic detection of inshore ships in high-resolution remote sensing images using robust invariant generalized hough transform. *IEEE Geoscience and Remote Sensing Letters*, *11*(12), 2070–2074.

Xu, C., Zhang, X., Yang, Y., & Niu, L. (2016). Deep learning-based recognition of underwater target. In *2016 IEEE international conference on digital signal processing* (pp. 89–93).

Yang, Y., Dong, J., Sun, X., Lima, E., Mu, Q., & Wang, X. (2017). A CFCC-LSTM model for sea surface temperature prediction. *IEEE Geoscience and Remote Sensing Letters*, *15*(2), 207–211.

Yang, G., Li, B., Ji, S., Gao, F., & Xu, Q. (2013). Ship detection from optical satellite images based on sea surface analysis. *IEEE Geoscience and Remote Sensing Letters*, *11*(3), 641–645.

Yang, Z., Yu, W., Liang, P., Guo, H., Xia, L., Zhang, F., et al. (2018). Deep transfer learning for military object recognition under small training set condition. *Neural Computing and Applications*, *31*(10), 6469–6478.

Yokoya, N., & Iwasaki, A. (2015). Object detection based on sparse representation and hough voting for optical remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(5), 2053–2062.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Yu, Y., Guan, H., & Zheng, J. (2015). Rotation-invariant object detection in high-resolution satellite imagery using superpixel-based deep hough forests. *IEEE Geoscience and Remote Sensing Letters*, *12*(11), 2183–2187.

Yuan, X., Huang, B., Wang, Y., Yang, C., & Gui, W. (2018). Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE. *IEEE Transactions on Industrial Informatics*, *14*(7), 3235–3243.

Yuh, J., Marani, G., & Blidberg, D. R. (2011). Applications of marine robotic vehicles. *Intelligent Service Robotics*, *4*(4), 221–231.

Zabidi, M. M. A., Mustapa, J., Mokji, M. M., Marsono, M. N., & Sha'ameri, A. Z. (2009). Embedded vision systems for ship recognition. In *TENCON 2009-2009 IEEE region 10 conference* (pp. 1–5).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 2528–2535).

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2012). Adaptive deconvolutional networks for mid and high level feature learning. In *2011 international conference on computer vision* (pp. 2018–2025).

Zhang, M. M., Choi, J., Daniilidis, K., Wolf, M. T., & Kanan, C. (2015). VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums. In *The IEEE conference on computer vision and pattern recognition* (pp. 10–16).

Zhang, L., He, Z., & Liu, Y. (2017). Deep object recognition across domains based on adaptive extreme learning machine. *Neurocomputing*, *239*, 194–203.

Zhang, H., Wang, Y., Xu, D., Wang, J., & Xu, L. (2018). The augmented complex-valued extreme learning machine. *Neurocomputing*, *311*, 363–372.

Zhang, R., Yao, J., Zhang, K., Feng, C., & Zhang, J. (2016). S-CNN-based ship detection from high-resolution remote sensing images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *41*, 423–430.

Zhao, L., Wang, X. F., Yuan, Y. T., & University, S. M. (2016). Research on ship recognition method based on deep convolutional neural network. *Ship Science and Technology*, *38*(8), 119–123.

Zhu, C., Hui, Z., Wang, R., & Guo, J. (2010). A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(9), 3446–3456.

Zou, Z., & Shi, Z. (2016). Ship detection in spaceborne optical image with SVD networks. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(10), 5832–5845.

Zou, F., Wang, Y., Yang, Y., Zhou, K., Chen, Y., & Song, J. (2015). Supervised feature learning via l2-norm regularized logistic regression for 3D object recognition. *Neurocomputing*, *151*, 603–611.