IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 30, 2020, accepted May 14, 2020, date of publication May 18, 2020, date of current version June 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995390

Deep Gabor Neural Network for Automatic Detection of Mine-Like Objects in Sonar Imagery

HOANG THANH LE[1,4], SON LAM PHUNG[1], (Senior Member, IEEE), PHILIP B. CHAPPLE[2], ABDESSELAM BOUZERDOUM[1,3], (Senior Member, IEEE), CHRISTIAN H. RITZ[1], (Senior Member, IEEE), AND LE CHUNG TRAN[1], (Senior Member, IEEE)

[1]School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2522, Australia
[2]Defence Science and Technology, Liverpool, NSW 1871, Australia
[3]Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[4]Faculty of Information Technology, Nha Trang University, Nha Trang, Vietnam

Corresponding author: Hoang Thanh Le (tlh857@uowmail.edu.au)

This work was supported in part by the NSW Defence Innovation Network and NSW State Government, and in part by the Australian Research Council.

ABSTRACT With the advances in sonar imaging technology, sonar imagery has increasingly been used for oceanographic studies in civilian and military applications. High-resolution imaging sonars can be mounted on various survey platforms, typically autonomous underwater vehicles, which provide enhanced speed and improved data quality with long-range support. This paper addresses the automatic detection of mine-like objects using sonar images. The proposed Gabor-based detector is designed as a feature pyramid network with a small number of trainable weights. Our approach combines both semantically weak and strong features to handle mine-like objects at multiple scales effectively. For feature extraction, we introduce a parameterized Gabor layer which improves the generalization capability and computational efficiency. The steerable Gabor filtering modules are embedded within the cascaded layers to enhance the scale and orientation decomposition of images. The entire deep Gabor neural network is trained in an end-to-end manner from input sonar images with annotated mine-like objects. An extensive experimental evaluation on a real sonar dataset shows that the proposed method achieves competitive performance compared to the existing approaches.

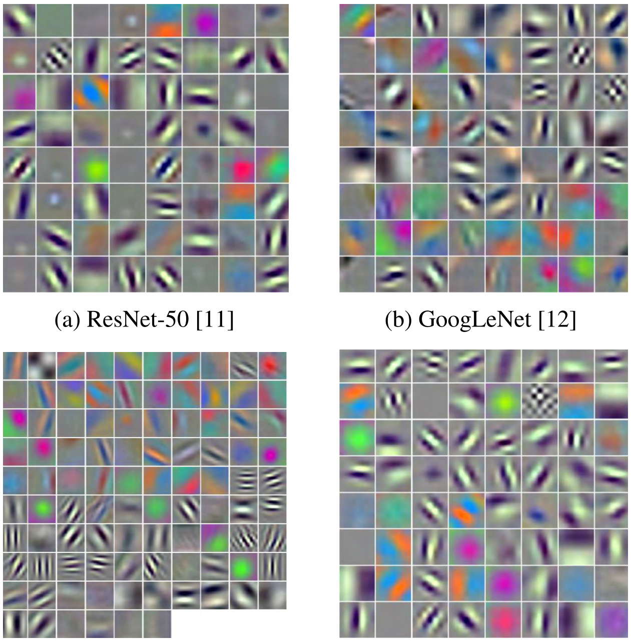
INDEX TERMS Gabor neural network detector, Gabor layer, side-scan sonar, mine-like objects.

I. INTRODUCTION

Over the past two decades, autonomous underwater vehicles (AUVs) have been increasingly used to survey the seabed. AUVs provide an effective platform for mounting high-resolution imaging sonars, e.g. side-scan or synthetic aperture sonars. Compared to radars and lidars, sonars are well-suited to the detection of small objects protruding from the seabed due to their abilities to visualize the dynamic underwater environments. Sound waves can propagate over a longer range than those of electromagnetic waves and light waves, due to their lower attenuation and dispersion in water. Compared to optical sensors, sonars are a more effective sensing modality for water-based activities in poor visibility, e.g. low-light or turbid conditions.

Automatic detection of mine-like objects (MLOs) in sonar imagery, which is a critical task for a mine clearance system, has attracted considerable research interest. As a cost-effective method in asymmetric warfare, underwater mines are commonly employed to block shipping lanes and restrict naval operations. Underwater mines can also cause long-lasting environmental damage due to the toxic explosive compounds. Despite its high demand in mine countermeasures, developing an automatic system for MLO detection is challenging for several reasons. First, a sufficient amount of labelled data is required to train a detection model. However, in practice, mine samples are extremely limited compared to other object detection tasks because of the costly and time-consuming data acquisition. Second, the acoustic features of echoes vary significantly depending on the range and aspect angle of sound pulses. As a result, an MLO (including its shadow) is often imaged with various shapes that cause difficulties for the detection process. Third, sonar imagery

The associate editor coordinating the review of this manuscript and approving it for publication was Mauro Tucci.



(a) ResNet-50 [11]

(b) GoogLeNet [12]

(c) AlexNet [13]

(d) DenseNet-201 [14]

FIGURE 1. Some convolutional kernels learned by the first layer of several well-known CNNs.

inherently includes the reverberation generated when transmitted acoustic beams strike the boundaries (i.e., water surface and seabed). The reverberation causes serious problems, especially in shallow water, since the clutter can dominate the background and completely cover the target objects.

Our Gabor-based approach is motivated by the biological and computational evidence of the Gabor filtering. It is widely accepted that the Gabor-like spatial functions are closely related to the mammalian vision systems, particularly in the perception of texture [1], [2]. Simple-cell receptive fields in the primary visual cortex of higher mammals are sensitive to orientations and spatial frequencies of the visual signal. Several neurophysiological studies showed that the simple cells found in the cat's striate cortex respond primarily to oriented edges and sinusoidal gratings, which can be approximated by the Gabor functions [3], [4]. Further studies conducted on macaques [5], [6] and humans [7], [8] also interpreted the computational models of the primary visual cortex as a bank of Gabor filters with selective orientation, spatial frequency, phase and bandwidth. Interestingly, such orientation-sensitive functions can be learned by many machine learning algorithms when applied to natural images. Several unsupervised methods, such as spike-and-slab sparse coding [9] and restricted Boltzmann machines [10], discover the features with Gabor-like weight patterns. In deep convolutional neural networks (CNNs) trained on large image datasets, many adaptive filters also converge to the Gabor functions, even from random initialization (see Fig. 1).

In this paper, we propose a Gabor-based neural network architecture for MLO detection in sonar imagery. Inspired by the YOLOv3 method [15], our approach adopts the detection framework with significant modifications in the network architecture. First, the Gabor filtering is embedded in the deep neural network for feature extraction and computational efficiency. As an effective way to control overfitting, the proposed Gabor layer has fewer trainable weights compared to the standard convolutional layer. The full hierarchical Gabor-based detector is trained in an end-to-end manner to discover the MLO features automatically. Second, our compact architecture is designed as a feature pyramid network (FPN) [16], where the low-resolution features are combined with the high-resolution features to compensate the information loss caused by the pooling effects. Compared to the original YOLOv3, the proposed Gabor detector enhances the semantic information of the feature pyramid at more scale levels to handle various MLO shapes (including shadows).

The main contributions of this paper can be highlighted as follows. First, we propose a new deep Gabor neural network (GNN) for MLO detection in sonar imagery. Second, we introduce the Gabor layer as a generic feature extractor for the design of compact neural architectures. Third, we conduct extensive experiments to evaluate the proposed method using a real sonar dataset provided by the Defence Science and Technology Group, Australia.

The remainder of the paper is organized as follows. Section II introduces the related work on the automatic detection of MLOs. Section III describes the proposed Gabor-based detection method. Section IV presents the experimental results and analysis, and finally, Section V gives the concluding remarks.

II. RELATED WORK

In this section, we first present a brief background on side-scan sonar imagery, and then provide a review of MLO detection methods.

A. SIDE-SCAN SONAR IMAGERY

A side-scan sonar provides high-resolution seabed morphology from both sides of an AUV, see Fig. 2. Typically, the sonar is mounted on a vehicle, which moves along a straight track at constant speed and altitude. Transducers on either side of the sonar periodically illuminate the seabed with fan-shaped beams of high-frequency acoustic signals perpendicular to the vehicle track. The backscattered intensities (as individual scan-lines) are then concatenated to form a two-sided sonar image. Note that such an image is represented in the *time* coordinate, instead of the Cartesian coordinate, where the echo amplitudes are displayed as image pixels. The vertical axis corresponds to the time when the acoustic pulse is emitted from the transducer, and the horizontal axis corresponds to the time of flight (i.e., slant range) in the across-track direction.

The seabed is commonly modeled as a Lambertian surface [17], which scatters incident energy uniformly in all directions. In other words, the echo amplitude depends only on the local angle of incidence δ formed by the incident pulse and the normal $\vec{n}$ to the surface. Let $p = (\vec{r}, \alpha)$ be a point

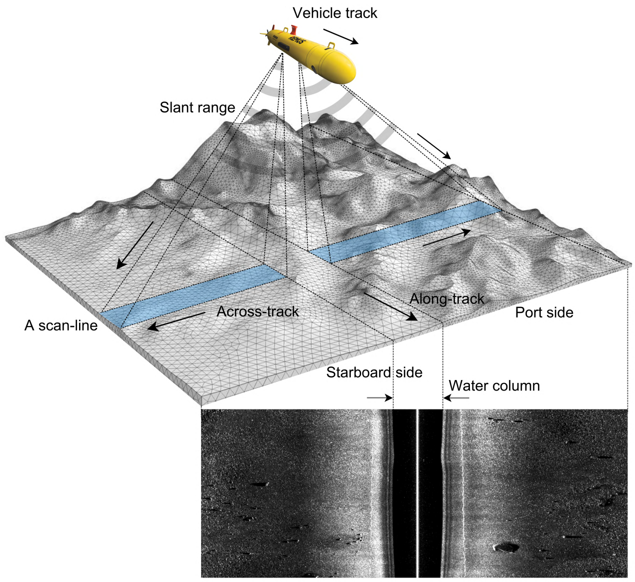


FIGURE 2. Principle of a side-scan sonar mounted on an autonomous underwater vehicle.

on the seabed ensonified by an anisotropic acoustic signal of intensity $\varphi(p)$. The backscattered intensity at p can be computed as

$$I(p) = \kappa\ \varphi(p)\ \mu(p)\ \frac{\vec{r}\cdot\vec{n}}{\|\ \vec{r}\ \|\ \|\ \vec{n}\ \|}, \tag{1}$$

where κ is a normalization constant, and $\mu(p)$ is the reflectivity coefficient of the seabed at p dependent on the sediment type. An example of sonar image formation is shown in Fig. 3.

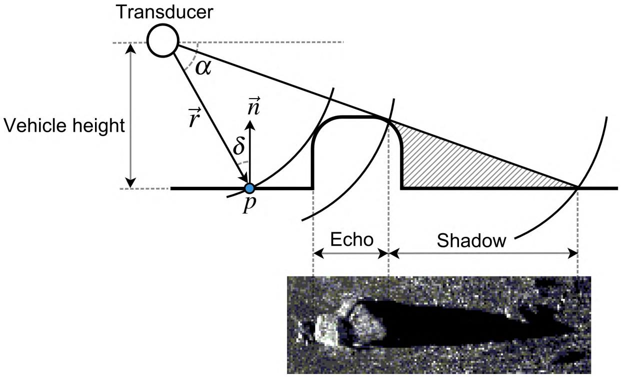


FIGURE 3. Sonar image formation from an object lying on the seabed.

B. TRADITIONAL MINE-LIKE OBJECT DETECTION METHODS

Over the past two decades, there have been several studies on automatic detection of MLOs using sonar imagery. This subsection presents a review of the traditional MLO detection methods.

Most existing MLO detection methods have employed feature-based algorithms to identify suspicious pixel regions. In [19], Sawas and Petillot applied the Haar-like features and a cascade of boosted classifiers, which were first introduced by Viola and Jones [31]. In [21], Barngrover *et al.* also utilized the Haar-like feature classifier to generate image patches (around regions of interest), which are then processed by subjects using the rapid serial visual presentation paradigm. Other feature-based methods used the geometric visual descriptors, such as scale-invariant feature transform (SIFT) [18], [32], [33] and local binary pattern (LBP) [20], [34]. In [18], Hollensen *et al.* adopted the dense SIFT feature extraction with various window sizes for computing orientation histograms. In [20], Barngrover *et al.* combined the LBP features and the AdaBoost algorithm to create an optimized cascade of features for classifying image windows. The existing feature-based methods have a limitation in that the feature extractors are manually designed to generate a feature vector from the input image window. However, finding an appropriate feature extractor to capture salient features of MLOs requires significant domain expertise.

In recent years, MLO detection methods have used deep neural networks to process sonar images in their raw form without manual feature engineering [22]–[24]. In [22], Gebhardt *et al.* proposed various CNNs, where a global average pooling (GAP) layer is employed before each fully-connected layer to produce a class activation map. In [24], Denos *et al.* introduced a four-step pipeline of MLO detection including synthetic data generation, one-class classification, background extraction, and binary classification. The second and fourth steps are performed using an auto-encoder and a pre-trained network VGG-19, respectively. In [23], McKay *et al.* utilized transfer learning with several pre-trained CNNs for mine feature extraction. The feature vectors are then used to train a support vector machine (SVM) on a small sonar dataset. The main limitation of the existing CNN-based methods is their computational cost. This is mainly due to the use of sliding windows for locating MLOs, where separate predictions are computed at every potential position. Furthermore, the existing methods do not handle MLOs with various shapes effectively, since the sliding windows (with a fixed aspect ratio) can lead to inaccurate bounding box detection.

C. GENERIC OBJECT DETECTION METHODS

MLO detection using sonar imagery can be considered as a subset of object detection. This subsection provides a brief survey of the generic object detectors in computer vision, which can be applied for the MLO detection.

With recent advances in deep learning, several techniques for generic object detection have been proposed, with state-of-the-art results. Such models can be categorized into two main types: i) two-stage detectors, and ii) one-stage detectors. Two-stage detectors, notably the R-CNN and its variations [25], [26], [30], perform object detection in two stages. In the first stage, a region proposal generation technique is used to remove most of the backgrounds. In the second stage, the remaining regions are categorized into different class

TABLE 1. Representative methods for MLO detection and generic object detection.

Application	Authors	Year	Technique
MLO detection	Hollensen *et al.* [18]	2011	SIFT features, SVMs
	Sawas and Petillot [19]	2012	Haar-like features, boosted classifiers
	Barngrover et al. [20]	2015	LBP features, boosted classifiers
	Barngrover *et al.* [21]	2016	Haar-like features, RSVP paradigm
	Gebhardt *et al.* [22]	2017	CNNs
	McKay *et al.* [23]	2017	Transfer learning, SVMs
	Denos *et al.* [24]	2017	Auto-encoder, CNN
Generic object detection	Girshick *et al.* [25]	2014	R-CNN
	Girshick [26]	2015	Fast R-CNN
	Liu *et al.* [27]	2016	SSD
	Redmond *et al.* [28]	2016	YOLOv1
	Redmond et al. [29]	2017	YOLOv2
	Ren *et al.* [30]	2017	Faster R-CNN
	Redmond *et al.* [15]	2018	YOLOv3

labels. In [25], Girshick *et al.* first introduced a method, called R-CNN (Regions with CNN features), where a selective search algorithm is employed to generate category-independent region proposals. Each candidate region is then classified using the AlexNet with the linear SVMs. In [26], Girshick proposed an improved version, called Fast R-CNN, where the feature maps are produced once from the entire image instead of region proposals. Based on the feature maps and the proposals suggested by the selective search, fixed-length feature vectors are then extracted for classification and regression using a region of interest (RoI) pooling layer. In [30], Ren *et al.* developed the Faster R-CNN with a separate fully-convolutional network, called Region Proposal Network (RPN), to predict candidate regions directly from the convolutional feature maps.

One-stage detectors, notably YOLO (You Only Look Once) [15], [28], [29] and SSD (Single Shot multi-box Detector) [27], predict bounding boxes directly from input images, without region proposal generation. In [28], Redmond *et al.* introduced the first version of YOLO, a real-time object detector. The main idea is to divide the image into grid cells, which are responsible for predicting the objects centered in these cells. For each grid cell, a CNN regressor is employed to predict several bounding boxes and the corresponding confidence scores. In [29] and [15], Redmond *et al.* adopted several powerful techniques to improve the detection performance of YOLO. In YOLOv2 [29], the fully-connected layers are removed from the base network Darknet-19, and multiple anchor boxes are utilized at each grid cell for predicting bounding boxes (similar to the Faster R-CNN). In YOLOv3 [15], the network Darknet-53 was proposed to make multiple predictions at different scales. In [27], Liu *et al.* proposed an object detector, called SSD, where six additional convolutional layers are appended to the base network VGG-16. Each additional layer produces feature maps at a scale for the detection prediction. SSD also adopts anchor boxes at multiple scales and aspect-ratios to predict objects on multiple feature maps. Essentially, SSD employs lower-resolution feature maps to detect large objects, and high-resolution feature maps to detect smaller objects. Table 1 presents a summary of representative methods for MLO detection and generic object detection.

III. PROPOSED DETECTION METHOD

This section presents the proposed detection method, including the deep Gabor neural network architecture (Section III-A), the proposed Gabor layer for feature extraction (Section III-B), the YOLOv3-based detection framework (Section III-C), the loss function for network training (Section III-D), and additional remarks on the conceptual contributions (Section III-E).

A. NETWORK ARCHITECTURE

The GNN detector utilizes a feature pyramid to make predictions at three different scales (see Fig. 4). The network comprises 17 Gabor layers with large kernel sizes in the early layers (i.e., 15 × 15 and 7 × 7 pixels) and smaller kernel sizes in the succeeding layers (i.e., 3 × 3 and 1 × 1 pixels). Each Gabor layer is followed by a batch-normalization layer and a LeakyReLU layer with the exception of the outputs. The network employs four max-pooling layers of size 2 × 2 pixels with stride of 2 for spatial dimensionality reduction.

Note that the high-resolution feature maps in the early Gabor layers are well-suited to locating small objects, but they contain semantically weak features. By contrast, the low-resolution feature maps in the succeeding Gabor layers contain semantically strong features, but the locations of MLOs are not precise due to the pooling effects. To overcome this problem, the proposed FPN architecture combines low-level features with high-level features using a bottom-up pathway, a top-down pathway, and two skip connections. This strategy not only enhances the semantic information

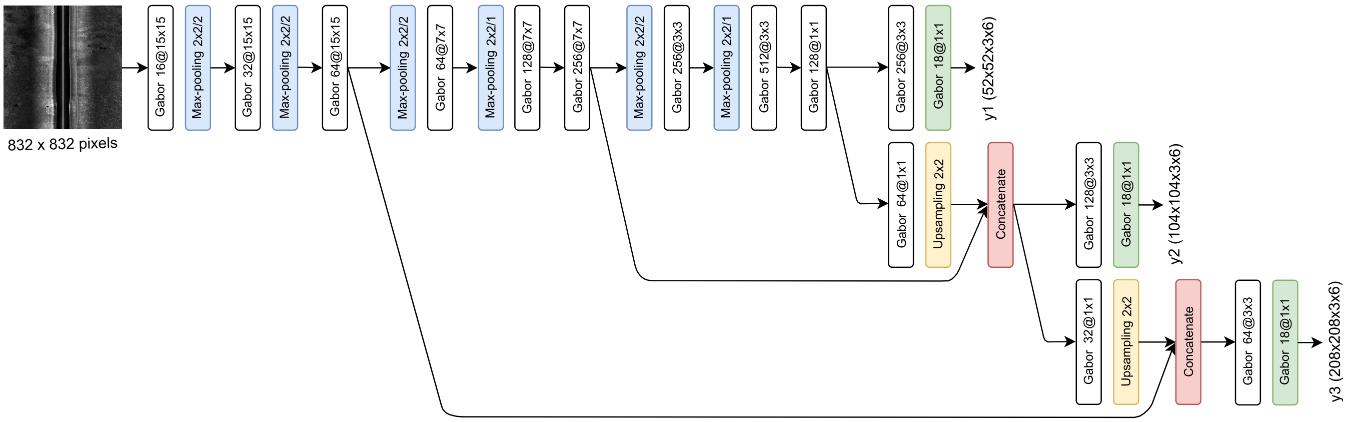


FIGURE 4. The proposed deep Gabor neural network for MLO detection in sonar images.

from both weak and strong features but also handles objects at multiple scales effectively.

The bottom-up pathway, which is the feed-forward computation of the backbone Gabor network, produces a feature hierarchy by reducing the spatial dimension gradually. Given an input sonar image of size 832 × 832 pixels, the first scale of 16 (i.e., 52 × 52 grid cells) is obtained at the top of the feature pyramid to predict large MLOs. The top-down pathway restores resolution from the semantically stronger (but spatially coarser) features by upsampling. The upsampled feature maps are then concatenated with those of identical spatial size from the bottom-up pathway via the skip connections. As a result, the second and the third scales of 8 and 4 (i.e., 104 × 104 and 208 × 208 grid cells) are produced to handle medium and small MLOs, respectively.

B. GABOR LAYER

A 2-D band-pass Gabor filter is an elliptical Gaussian envelope modulated by a complex sinusoidal wave of specific frequency and orientation. The harmonic component enables the filter to be sensitive to spatial frequencies, while the Gaussian component constrains the frequency sensitivity to localized regions of the input image. As an edge detector, Gabor filter responds strongly to patterns matching the orientation of sinusoidal strips, and suppresses those perpendicular to the orientation. This subsection introduces our Gabor-based feature extractor, called Gabor layer, which can be trained in an end-to-end manner.

Let σ_x and σ_y be the standard deviations of elliptical Gaussian envelope, which control the spatial scale of a Gabor filter. Let ϕ be the phase offset, which determines how much the sinusoidal component needs to be shifted with respect to the origin. A complex Gabor filter plane with real and imaginary components representing orthogonal directions is defined as

$$G(x, y) = \frac{1}{2\pi\ \gamma\ \sigma^2} \exp\{-\frac{\frac{\tilde{x}^2}{\gamma^2} + \tilde{y}^2}{2\sigma^2}\} \exp\{i\ 2\pi u_0\ (\tilde{x} + \phi)\}, \tag{2}$$

where $\sigma = \sigma_y$, and $\gamma = \sigma_x/\sigma_y$ is the spatial aspect ratio which reflects the ellipticity of the envelope. Here, $\tilde{x} = x\cos\theta + y\sin\theta$ and $\tilde{y} = -x\sin\theta + y\cos\theta$ denote the transformed coordinates, where θ specifies the orientation of the normal to the parallel stripes. In Eq. (2), $u_0 = \sqrt{u^2 + v^2}$ is the center frequency, where u and v are the spatial frequencies of the sinusoidal factors.

In practice, instead of specifying the value of σ directly, the receptive field is determined by the half-response spatial frequency bandwidth β, which is given by

$$\beta = \log_2 \frac{\frac{\sigma}{\lambda}\ \pi + \sqrt{\frac{\ln 2}{2}}}{\frac{\sigma}{\lambda}\ \pi - \sqrt{\frac{\ln 2}{2}}}. \tag{3}$$

Here, λ denotes the wavelength associated with the spatial frequency of the sinusoidal component. From Eq. (3), the standard deviation σ is related to the wavelength by

$$\frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\ \frac{2^\beta + 1}{2^\beta - 1}. \tag{4}$$

Note that the spatial frequency bandwidth determines the cutoff of the filter frequency response as frequency moves away from the center frequency u_0 (i.e., $1/\lambda$). The ratio σ/λ determines the number of parallel excitatory and inhibitory lobes observed in the receptive field. In summary, a single filter plane is controlled by five parameters $\lambda, \theta, \phi, \gamma$ and β, which are treated as the learnable parameters to be determined by the training algorithm.

In this paper, we adopt the terminology commonly used in deep learning literature when describing the network architecture [13], [28], [35]. Hereafter, a Gabor kernel is a 3D tensor that comprises several Gabor planes organized as a filter bank (see Fig. 5) so that the salient MLO features can be extracted at various orientations, scales and translations. In a deep hierarchical network, a Gabor layer employs several parameterized Gabor kernels as steerable feature extractors. These spatial kernels are then convolved with the input channels, yielding a Gabor space. We utilize the real impulse response of the complex-valued kernels for the convolutional

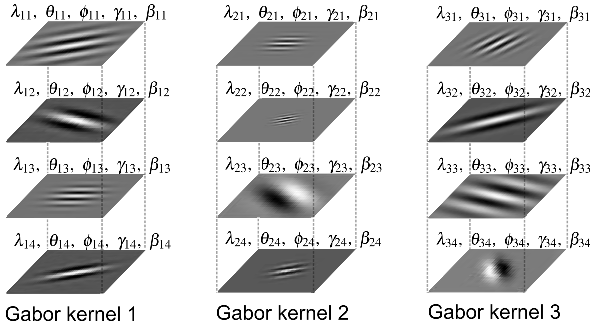


FIGURE 5. Visualization of three example Gabor kernels with four input channels used in a single Gabor layer. Here, the real components of the kernels are used.

computation since they resemble the receptive field found in the cat's striate cortex [36]. Mathematically, let O_i^l be the i-th feature map in the l-th Gabor layer, and $G_{i,j}^l$ be the i-th filter plane of the j-th Gabor kernel. The j-th output feature map can be computed as

$$O_j^{l+1} = f\left(\sum_{i=1}^{n} O_i^l * G_{i,j}^l\right), \tag{5}$$

where $*$ denotes the two-dimensional convolution operator, and f represents a non-linear activation function for the extraction of non-linear features.

C. DETECTION FRAMEWORK

Each grid cell in a certain scale level employs three anchor boxes (i.e., prior boxes) to predict bounding boxes. During the training phase, each object is assigned to a grid cell containing the object's center and an anchor box associated with the highest intersection over union (IoU). The network makes prediction as a logistic regression with six components: (i) four scores (x, y, w, h) reflecting the offset of predicted bounding box; (ii) an objectness score s representing the IoU between the predicted bounding box and the ground-truth; and (iii) a conditional class probability $p(\text{class} = \text{MLO}|\text{object})$. Here, the coordinates (x, y) are the object's center relative to the grid cell, and (w, h) are the width and height relative to the entire sonar image. Collectively, the prediction at each scale is encoded as a tensor of size $n \times n \times 3 \times 6$, where n is the number grid cells used in the scale level.

Note that our model predicts the relative offsets instead of the absolute coordinates. Inspired by the YOLOv3 detection technique [15], [29], we process the relative offsets to generate the absolute coordinates for the final output. Briefly, the predicted center coordinates (x, y) and the output objectness score s are squashed between 0 and 1 using a sigmoid function. Given the predicted sizes (w, h), the absolute outputs are obtained by computing the exponential then multiplying by the corresponding sizes of the anchor.

During the test phase, the predicted conditional class probabilities are multiplied by the corresponding objectness score to produce a class-specific score for each bounding box [29]. In other words, the class-specific score implicitly encodes: (i) the probability of an MLO occurring in the predicted box, and (ii) how well the box fits the object. Our method then removes detections with scores lower than a predefined confidence threshold, and sorts the remaining bounding boxes in the descending order of the class-specific score. An analysis of the confidence threshold selection is given in Section IV-D. Since multiple proposal boxes can be predicted for the same object, the non-maximum suppression (NMS) algorithm [28] with a pre-defined IoU threshold is adopted to remove duplicate detections.

D. LOSS FUNCTION

During training, we minimize the YOLOv3-based loss function which is defined as

$$\begin{aligned}
\mathscr{L} = & \sum_{i=1}^{n\times n}\sum_{j=1}^{3} \mathbb{1}_{ij}^{\text{MLO}}[l(x_i, \hat{x}_i) + l(y_i, \hat{y}_i))] \\
& + \sum_{i=1}^{n\times n}\sum_{j=1}^{3} \mathbb{1}_{ij}^{\text{MLO}}[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \\
& + \sum_{i=1}^{n\times n}\sum_{j=1}^{3} \mathbb{1}_{ij}^{\text{MLO}} l(s_i, \hat{s}_i) \\
& + 0.5 \sum_{i=1}^{n\times n}\sum_{j=1}^{3} \mathbb{1}_{ij}^{\text{noMLO}} l(s_i, \hat{s}_i) \\
& + \sum_{i=1}^{n\times n} \mathbb{1}_{i}^{\text{MLO}} l(p_i, \hat{p}_i).
\end{aligned} \tag{6}$$

Equation 6 can be explained as follows:

- The loss function $\mathscr{L}$ consists of three components: (i) localization loss, (ii) confidence loss, and (iii) classification loss.
- The first and second terms denote the localization loss, which measures the errors in the offsets of the predicted bounding box. To consider the regression errors with respect to the bounding box sizes, we apply the square root operator, which reduces the significance of high regression errors for large boxes.
- The third and fourth terms denote the confidence loss, which measures the errors in the objectness score of the bounding box in both cases, with and without an MLO detected in the box.
- The fifth term denotes the classification loss measuring the difference between the actual and predicted class probabilities if an MLO is present in the grid cell.
- $\mathbb{1}_{ij}^{\text{MLO}} = 1$ if the j-th bounding box in the i-th grid cell is responsible for detecting an MLO, otherwise 0, and $\mathbb{1}_{ij}^{\text{noMLO}}$ is the complement of $\mathbb{1}_{ij}^{\text{MLO}}$. The function l is a binary cross-entropy loss given by

$$l(a, \hat{a}) = -a \log \hat{a} - (1 - a) \log(1 - \hat{a}). \tag{7}$$

E. REMARKS AND DISCUSSION

Before presenting the experimental results and analysis, we provide brief remarks on the proposed Gabor layer and GNN detector to highlight the contributions.

It is worth noting that the number of trainable parameters of a single Gabor kernel is independent of the kernel size. In designing deep networks, the receptive field (the kernel size) needs to cover the entire relevant image region. A sufficiently large receptive field is required to capture the local context around every single pixel when making the prediction. Existing attempts to extend the receptive field have used large convolutional kernels in the early layers [13], or stacking several layers with small kernels [11], [37], [38]. However, increasing the receptive field size leads to a rapid growth in the number of trainable parameters and computational cost. Given a standard convolutional layer, let k be the number of kernels of size $m \times n$ pixels, and c be the number of input feature maps. The number of trainable weights in this convolutional layer is $(m \times n \times c + 1) \times k$. By contrast, the proposed Gabor-based approach represents each filter plane with only five parameters, regardless of the kernel size. Thus, the number of trainable weights is reduced to $(5 \times c + 1) \times k$. As a generic feature extractor, the Gabor kernel enables us to design compact networks with fewer free parameters compared to the convolutional counterparts.

The GNN detector has several conceptual merits compared to the relevant approaches of MLO detection. In terms of network architecture, the proposed method extracts MLO features at multiple scales, while maintaining a compact architecture with fewer trainable parameters. Compared to the tiny YOLOv3 method which decomposes the input image at two scales of 32 and 16, our network performs the detection at three scales of 16, 8, and 4. In other words, the proposed detector employs smaller grid cells at various sizes to handle MLOs effectively. Compared to the full YOLOv3 with the feature extractor Darknet-53 [15], the proposed GNN achieves roughly 30 times reduction in the total number of trainable weights. A small network size enables the entire GNN model to be deployed on various survey platforms (e.g., AUVs) as an efficient on-chip architecture.

In terms of detection framework, our approach processes the entire input sonar image with a single feed-forward propagation through the Gabor network, instead of using the sliding window and region proposal techniques. This improves the detection speed and the contextual information of the extracted features. The proposed one-stage method performs MLO detection as a regression problem, where bounding box offsets and class probability are obtained directly from image pixels. In other words, this enables us to maintain a simple detection pipeline without the softmax and classification layers.

In terms of feature extraction, the Gabor filtering enhances not only the scale and orientation decomposition of images but also the invariant properties of the extracted features [39]. Compared to the standard convolutional kernels with randomly-initialized weights, the Gabor kernels follow patterns that are steerable to specific frequencies. A bank of several Gabor filters can effectively extract the directional texture features (e.g., shadows and strong edges) representing structural properties of MLOs.

TABLE 2. Summary of sonar data acquisition and experimental setup.

Description	Specification
Sonar equipment	MST side-scan sonar
Transmitted frequency	900 kHz
Maximum resolution	0.2 m
Operation range (port and starboard sides)	30 m
AUV	REMUS 100 AUV
Image size	1000 × 1024
Number of MLOs	216
Number of sonar images	190
Number of CV folds	5
Training/test images in each CV fold	153/37
Augmented training images in each CV fold	1683

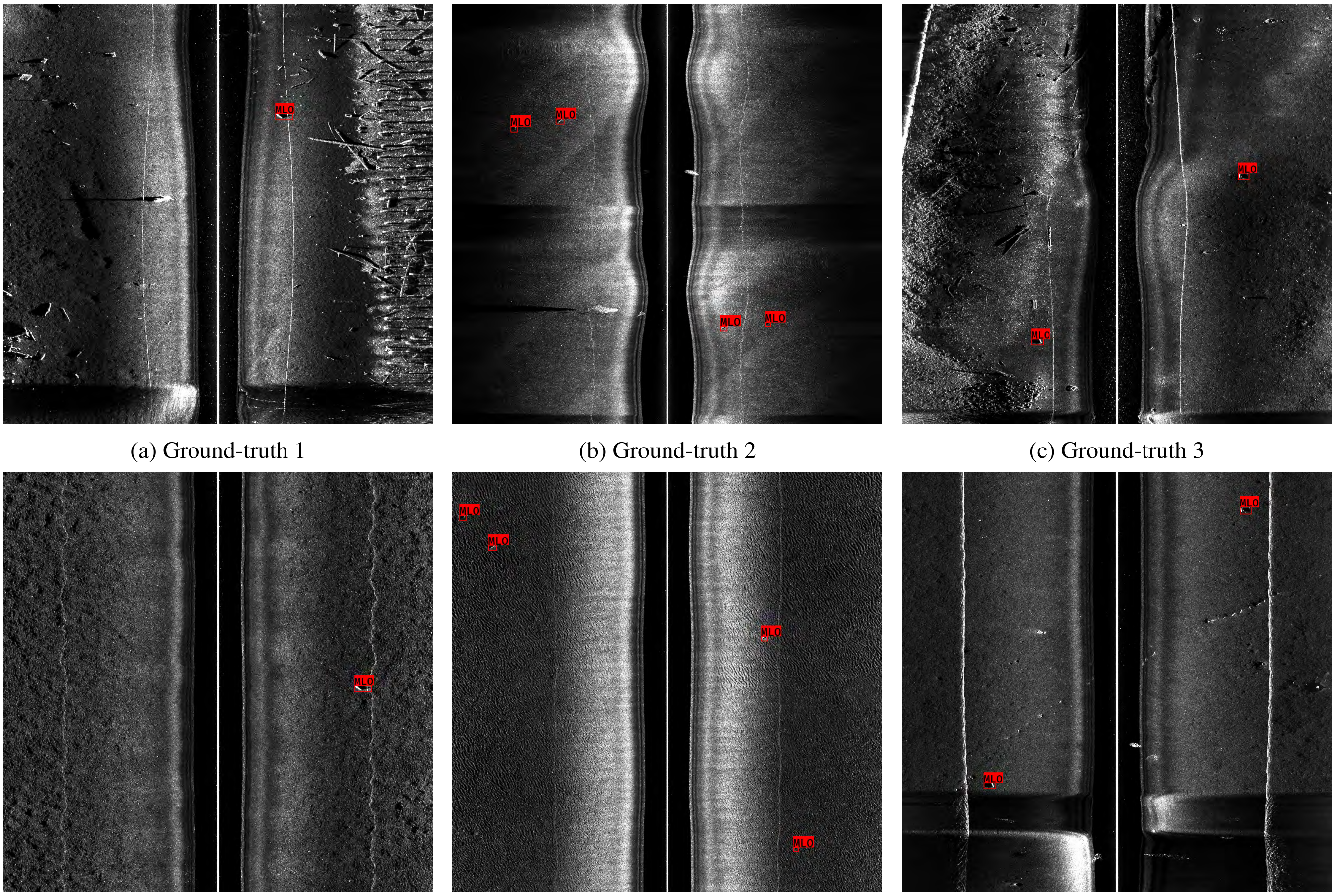
IV. RESULTS AND ANALYSIS

In this section, we first describe the data acquisition (Section IV-A) and the detection evaluation metrics (Section IV-B), then investigate the anchor box selection (Section IV-C) and confidence threshold selection (Section IV-D). Finally, we compare the proposed method with six state-of-the-art generic object detectors in computer vision (Section IV-E) and four relevant representative MLO detection methods (Section IV-F).

A. SONAR DATA ACQUISITION AND ANNOTATION

The sonar data were provided by the Defence Science and Technology (DST) Group in a naval mine-shape recovery operation in Australia [40]. A Marine Sonic Technology (MST) side-scan sonar with dual frequencies was employed for data acquisition. This sonar equipment has: (i) a 900 kHz channel with a resolution of 0.2 m and a practical maximum range of 30 to 40 m; and (ii) a 1800 kHz channel with a resolution of 0.05 to 0.1 m and a maximum range of 10 to 15 m. In the surveys conducted by the DST Group, the first channel of 900 kHz was used, and the maximum range of sonar operation for both port and starboard sides was set to 30 m. The REMUS 100 AUV by Kongsberg Maritime was utilized as an unmanned platform for rapidly detecting MLOs on the seabed. The REMUS 100 AUV is a compact, lightweight vehicle designed for operation in coastal environments. It has a maximum depth of 100 m, and an endurance of up to 12 hours at the standard cruising speed of 1.5 m/s (i.e., 3 knots) dependent on the sensor configuration. The MLOs in the acquired sonar images were annotated by the DST experts. There are 216 MLOs in 190 sonar images of size 1000 × 1024 pixels.

The original images were resized to 832 × 832 pixels to satisfy the designed input shape (i.e., multiple of 32)



(a) Ground-truth 1

(b) Ground-truth 2

(c) Ground-truth 3

(d) Synthesized image from (a)

(e) Synthesized image from (b)

(f) Synthesized image from (c)

FIGURE 6. Data augmentation for training the MLO detectors. Top row: original sonar images with the MLO ground-truth. Bottom row: synthesized sonar images with the MLO ground-truth. See electronic color images.

before being partitioned randomly into five cross-validation folds. Thus, each case of cross-validation contains 153 sonar images for training and 37 images for testing. For each fold, we applied data augmentation to the training set to synthesize additional training images as follows. The annotated MLOs were extracted from the original images and then overlaid on seabed backgrounds (without MLOs) at random locations. The overlaying was performed such that the shadow direction of the MLO matched to the shadow direction in the background image (i.e., across-track direction). Finally, each augmented case of cross-validation contains 1683 images for training and 37 images for testing. A summary of sonar data acquisition and experimental setup is shown in Table 2. Figure 6 presents three examples of original sonar images with MLOs and the corresponding synthesized images for data augmentation in our dataset.

B. DETECTION EVALUATION METRICS

To measure the detection performance, we adopted the evaluation metric of the PASCAL Visual Object Classes (VOC) Challenge [41], which has been widely accepted as the benchmark for detection tasks. The principal quantitative metric is the average precision (AP) using all-point interpolation, which can be closely estimated as the area under the precision-recall curve (AUC). Note that, to compute the precisions and recalls, the detections are converted to classifications based on a pre-defined threshold of IoU. The predicted bounding boxes having IoU scores (with the ground-truths) above the threshold are considered as true positives, and those with IoU scores below the threshold are considered to be false positives. If multiple bounding boxes detect the same MLO, the box with the highest IoU is counted as a correct detection, and the remaining boxes are interpreted as false detections.

Let $r_i \in [0, 1]$ be the i-th recall value, and $\rho(r_i)$ be the measured precision at r_i. A version of the precision-recall curve with precision monotonically decreasing is obtained by setting $\rho(r_i)$ to the maximum precision for any recall $\tilde{r} \geq r_i$. The AP (i.e., AUC) interpolated over n unique recall values can be computed as

$$AP = \sum_{i=2}^{n-1} (r_i - r_{i-1})\, \rho_{int}(r_i), \tag{8}$$

where $\rho_{int}(r_i) = \max_{\tilde{r} \geq r_i} \rho(\tilde{r})$.

C. ANCHOR BOX SELECTION

Anchor boxes (i.e., prior boxes) affect significantly the efficiency and accuracy of an object detector. Such pre-defined boxes are commonly used to capture the aspect ratio of specific object classes and handle multiple objects associated with the same grid cell. Inspired by YOLOv2 [29], our approach present the anchor boxes by running k-means clustering on the training MLO bounding boxes. Instead of using Euclidean distance as in the standard k-means algorithm, we use the IoU distance metric in clustering, which aims to avoid the errors caused by the scale of boxes. The IoU metric is computed as

$$d(box, centroid) = 1 - \text{IoU}(box, centroid). \qquad (9)$$

To investigate the effects of the number of anchor boxes used for each grid cell, we varied its value from 1 to 15 with a step of 1. Figure 7 shows the average IoU as a function of the number of anchors. In practice, the average IoU should be greater than 0.5, so that anchor boxes overlap well with bounding boxes in the training data. Increasing the number of anchors improves the average IoU measure, but using more anchor boxes may cause overfitting and increase the computational cost [29]. Note that the number of anchors used in our case must be a multiple of 3, since the proposed Gabor detector produces three output scales. Among the evaluated values, we selected nine candidate anchor boxes with an average IoU of 0.813 for all subsequent experiments.

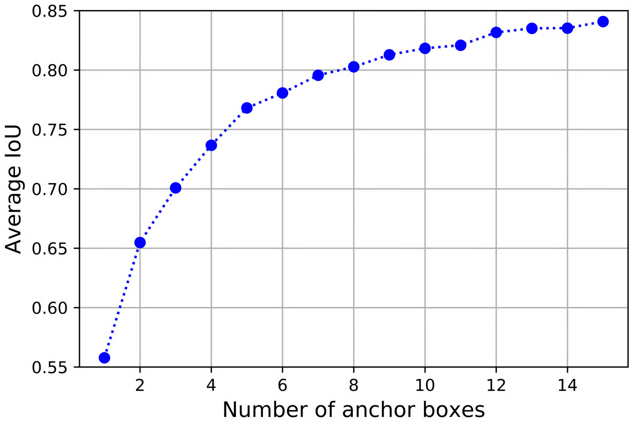


FIGURE 7. **Relationship between the number of anchors and the average IoU.**

D. CONFIDENCE THRESHOLD SELECTION

During the test phase, the proposed method employs a pre-defined confidence threshold to discard weak detections. The higher is the threshold value, the more candidate bounding boxes are removed from the final detections. To investigate the effects of the confidence threshold on the detection performance, we varied its value from 0.05 to 0.85 with a step of 0.05. The AP was measured at IoU = 0.5 as in the PASCAL VOC metric. Figure 8 shows the AP as a function of the confidence threshold. The experimental validation indicates that the suitable range for the threshold is [0.05, 0.15], where the AP measure remains stable. Based on these results, we employ the threshold value of 0.15 for the subsequent experiments.

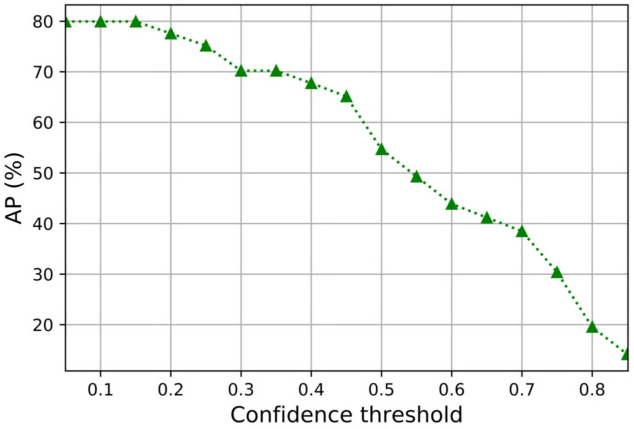


FIGURE 8. **Relationship between the confidence threshold and the detection accuracy.**

E. COMPARISON WITH THE STATE-OF-THE-ART OBJECT DETECTORS

The proposed Gabor detector is compared to six state-of-the-art generic object detectors: 1) R-CNN [25], 2) Fast R-CNN [26], 3) Faster R-CNN [30], 4) SSD300 [27], 5) tiny YOLOv3, and 6) full YOLOv3 [15]. All experiments were conducted on a computer with Intel Xeon Gold 5115 2.40 GHz processor and NVIDIA TITAN Xp GP102 graphics card.

- For the R-CNN detector and its variants (i.e., Fast R-CNN, Faster R-CNN), the ResNet-50 [11] was employed as a backbone network for feature extraction. A new classification layer, a regression layer, and a ROI max-pooling layer (applied to the Fast R-CNN and Faster R-CNN) were then added to the backbone to support object detection. To generate the region proposals for the R-CNN and the Fast R-CNN, we employed the Edge Boxes algorithm [42], which has been shown to be more computationally efficient than the Selective Search algorithm. The maximum number of strongest region proposals used for generating training samples was set to 2,000. The negative and positive ranges, which are used to determine the negative and positive training samples if the region proposals overlap with the ground-truths, were set to [0, 0.3] and [0.3, 1], respectively.
- For the SSD300 detector, we utilized the standard input shape of 300 × 300 pixels. The confidence threshold for removing the weak detections was set to 0.4.
- For the tiny and full YOLOv3 detectors, we employed the pre-trained tiny weights and Darknet-53 weights [15], respectively. The confidence threshold and the IoU threshold of the NMS algorithm [28] were set to 0.3 and 0.15, respectively.

Table 3 presents the detection performance of the evaluated methods. In terms of accuracy, it is clear that the proposed

TABLE 3. Detection performance of the proposed GNN and other object detectors.

Method	AP ± SD (%)	Frames/second	# Total trainable weights	Model size (KB)	Model size relative to the GNN
GNN	**79.93 ± 7.66**	3.01	**2,084,880**	**8,305**	1 ×
R-CNN [25]	37.62 ± 11.85	0.30	25,502,912	100,632	12.11×
Fast R-CNN [26]	18.26 ± 3.90	1.29	25,436,865	100,371	12.01×
Faster R-CNN [30]	9.41 ± 3.16	2.89	25,703,510	101,423	12.21×
Tiny YOLOv3 [15]	70.54 ± 8.91	**28.41**	8,669,876	33,990	4.10×
Full YOLOv3 [15]	72.76 ± 9.53	15.07	61,523,734	241,082	29.01×
SSD300 [27]	27.08 ± 6.90	7.17	23,371,782	91,427	11.01×

AP was measured at IoU = 0.5 (as in PASCAL VOC metric). SD stands for the standard deviation. Model size was calculated for the HDF5 file storage.

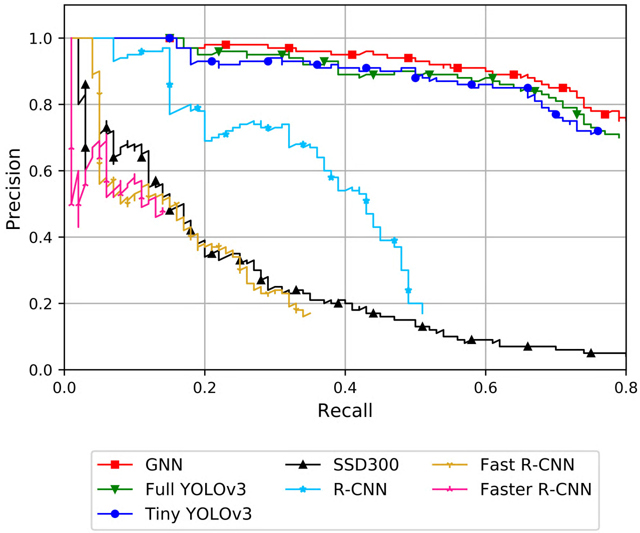


FIGURE 9. Precision-recall curves of the GNN and other object detectors over the five cross-validation folds.

GNN detector outperforms the existing object detectors. Among the evaluated methods, the proposed method achieves the highest AP of 79.93%, while the AP yielded by the existing methods varies from 9.41% to 72.76%. Compared to the full YOLOv3 and tiny YOLOv3, the best and second-best existing detectors, the GNN detector produces an improvement of 7.17% and 9.39%, respectively. In terms of model size, the proposed compact GNN achieves a significant reduction compared to other methods. The model size of the GNN detector is 4.1 times smaller than that of the tiny YOLOv3 detector.

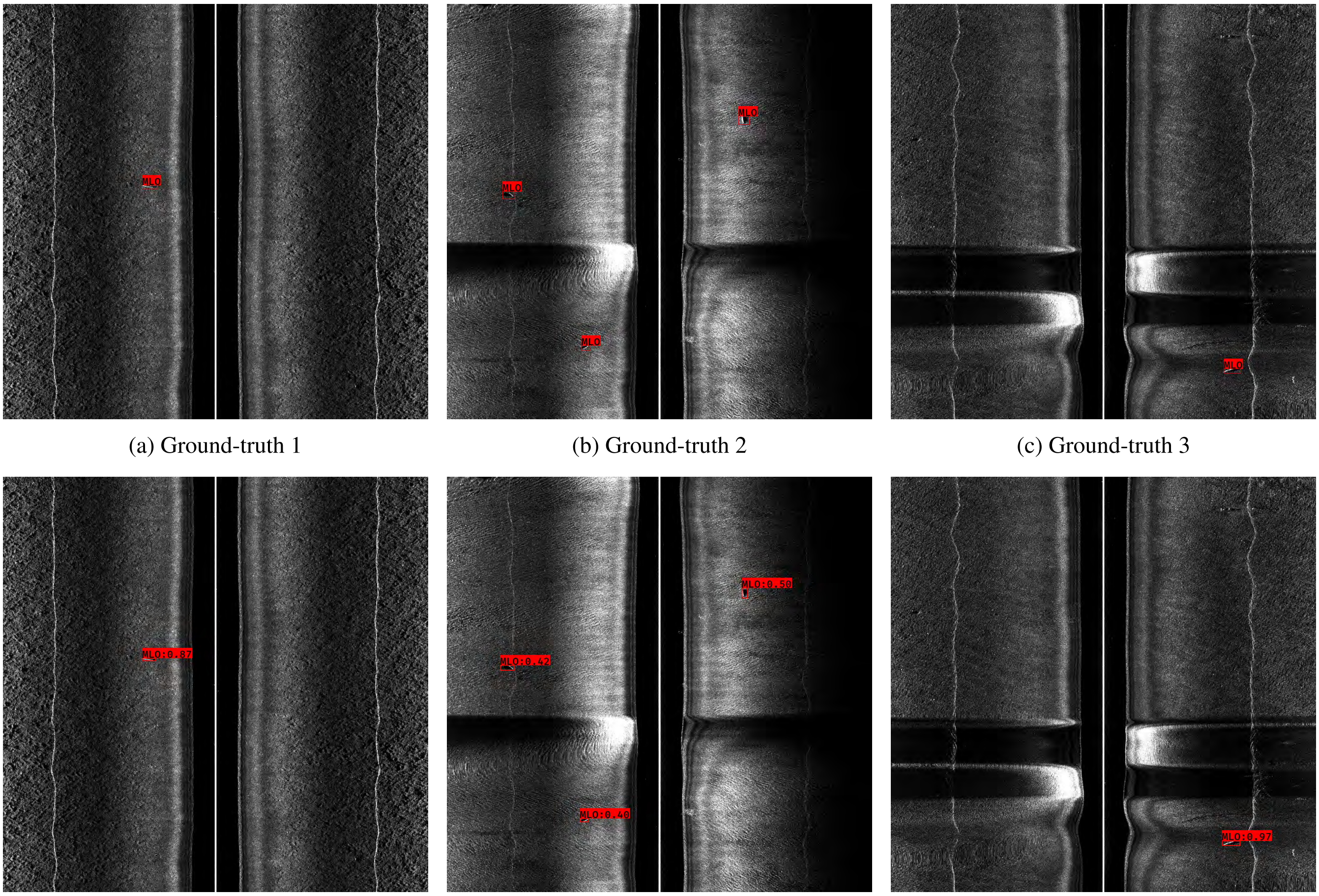
In terms of detection speed, Table 3 shows that the proposed method is faster than the two-stage detectors (R-CNN, Fast R-CNN, and Faster R-CNN), and slower than the existing one-stage detectors (YOLOv3 and SSD300). It can operate at a speed of 3.01 frames/s, which is 10 times faster than the R-CNN, and 5 times slower than the full YOLOv3. Note that this paper focuses on improving the detection accuracy due to the user demand of a reliable MLO detection algorithm. Although the current detection speed is acceptable to the users, it would be useful to improve the inference time by investigating more compact networks and optimizing the Python implementation of the Gabor layer. Both directions are feasible, and we leave their detailed explorations for future studies.

Figure 9 presents the precision-recall curves over the five cross-validation folds for further insights into the detection capability of the evaluated object detectors. Clearly, the precision-recall curve produced by the proposed GNN is better than the others because it produces a higher precision at each level of recall. The detection performance of the GNN is also more stable than those of the existing methods. Several outputs of the GNN detector are presented in Fig. 10. The experimental results show that the proposed method can detect MLOs with various shapes, in different seabed terrains.

On our sonar image dataset, YOLOv3 is found to have better detection accuracy than Faster R-CNN. On benchmark datasets such as MS COCO, Faster R-CNN is shown to have similar detection accuracy as YOLOv3 [15], [26], [30]. A possible explanation for the different findings is the small number of sonar images available for training. Our sonar dataset contains 190 sonar images (before data augmentation) with 216 MLOs, as it costs several thousand dollar to deploy an underwater mine, record sonar images, and retrieve the mine. In comparison, the MS COCO dataset for object detection task contains more than 200,000 images with over 500,000 object instances categorized into 80 classes [43]. Furthermore, Faster R-CNN is a two-stage detector that uses an additional fully-convolutional network (i.e. the RPN) for predicting candidate regions, whereas YOLOv3 is a one-stage detector. It is possible that Faster R-CNN needs more training images to reach a similar detection performance as YOLOv3.

F. COMPARISON WITH THE RELEVANT MLO DETECTION METHODS

The proposed GNN detector is compared to four representative existing methods that were specifically designed for MLO detection: (1) Haar-like cascade detector [19], (2) LBP cascade detector [20], (3) the pre-trained VGG-19 with an SVM classifier [23], and (4) CNNs with GAP layer [22].



(a) Ground-truth 1 (b) Ground-truth 2 (c) Ground-truth 3

(d) GNN output of (a) (e) GNN output of (b) (f) GNN output of (c)

FIGURE 10. Representative visual results produced by the proposed GNN detector. Top row: test sonar images with the MLO ground-truth. Bottom row: detection results by the GNN. See electronic color images.

- For Method (1) and (2), we found that the number of cascade stages giving the best performance is from 5 to 7, which agrees with [19]. Note that the more cascade stages we use, the more image data are required to train the detector. For the subsequent experiments, we employed the value of 5 which is well-suited to our available sonar data. A scaling factor of 1.1, which determines the amount of scaling applied to the input image after each increment, was employed to enable multi-scale detection.
- For Method (3) and (4), we implemented the network architecture as suggested in [22], [23]. A sliding window of fixed size 101 × 101 pixels and a sliding step of 20 pixels was utilized to locate the MLOs. For Method (4), the network consists of 9 convolutional layers and a GAP layer added after the last convolutional layer. The input image size of 832 × 832 pixels for these methods was the same as those of the GNN detector.

Note that the cascade detectors do not produce the confidence scores, which are employed to sort the detections before calculating the precisions and recalls. The CNN-based methods merely classify the sliding window without returning the offsets of bounding boxes. Hence, instead of using the AP metric to evaluate the detection performance, we recorded three performance measures: 1) the number of correct detections (i.e., true positives), 2) the number of incorrect detections (i.e., false positives), and 3) the number of ground-truths not detected (i.e., false negatives). A predicted sliding window containing an MLO is considered as a correct detection. When multiple windows cover the same MLO, the first predicted window is counted as a correct detection, and the remaining windows are interpreted as incorrect detections. The scores were accumulated over the five cross-validation folds.

Table 4 shows the performance of four existing MLO detection methods. Clearly, the proposed GNN detector outperforms the existing methods in terms of both the correct detection rate and the frame rate. The GNN detector achieves a detection rate of 80.5% (i.e., 174/216), which is 3.8 times higher than that of the VGG-19 method. The results also indicate that the GNN detector is more reliable than the existing methods: it produced the smallest number of incorrect detections (46) over the five test folds. Compared to the cascade detectors with a frame rate of roughly 0.05 frames/s,

TABLE 4. Detection performance of the proposed GNN and relevant MLO detection methods.

Method	# Correct detections	# Incorrect detections	# Ground-truths not detected	Frames/second
GNN detector	**174**	**46**	**42**	**3.018**
Haar-like cascade detector [19]	18	319	198	0.053
LBP cascade detector [20]	23	267	193	0.051
VGG-19 + SVM [23]	46	107	170	0.004
CNN + GAP [22]	9	134	207	0.007

the proposed method is 57 times faster. The CNN-based methods using sliding window are the slowest with the frame rates between 0.004 to 0.007 frames/s.

V. CONCLUSION

In this paper, a novel Gabor-based deep neural network architecture is proposed for automatic detection of MLOs in sonar imagery. The steerable Gabor filtering modules are embedded within the cascaded layers to enhance the scale and orientation decomposition of images. The proposed GNN is designed as a FPN-like architecture with a small number of trainable weights, which can be trained in an end-to-end manner to extract the MLO features automatically. The experimental results on a real sonar dataset, provided by the DST Group, Australia, indicates that the proposed GNN is an effective MLO detection method for AUVs in terms of the accuracy and the model size. Compared to the state-of-the-art object detectors in computer vision, the proposed GNN demonstrates a significant improvement in the AP metric and at least 4 times reduction in the model size. Compared to the relevant MLO detection methods, our approach not only achieves a higher detection rate but also improves the detection speed significantly.

APPENDIX

DERIVATION OF GABOR ERROR GRADIENT

This section presents the derivation of Gabor error gradient, which is used for end-to-end training of the proposed network.

1) $o^l_j(x, y)$ is the output of neuron (x, y) in the j-th feature map of the l-th Gabor layer:

$$o^l_j(x, y) = f(s^l_j(x, y)), \tag{10}$$

where f denotes an activation function.

2) $s^l_j(x, y)$ is the weighted sum input to neuron (x, y) in the j-th feature map of the l-th Gabor layer produced by convolutional computation:

$$s^l_j(x, y) = \sum_{i=1}^{n} \sum_{x'} \sum_{y'} g^l_{i,j}(x', y')\, o^{l-1}_i(x', y'). \tag{11}$$

3) $g^l_{i,j}(x, y)$ is a real impulse response of the i-th filter plane in the j-th Gabor kernel. The value of $g^l_{i,j}(x, y)$ yielded from the trainable Gabor weights is defined by Eq. (2).

4) Using the chain rule of differentiation, we can express the partial derivative of the total error with respect to (w.r.t.) the k-th weight for the i-th filter plane in the j-th Gabor kernel (i.e., $\lambda^l_{i,j}$, $\theta^l_{i,j}$, $\phi^l_{i,j}$, $\gamma^l_{i,j}$ and $\beta^l_{i,j}$) as

$$\frac{\partial E}{\partial w^l_{i,j}(k)} = \frac{\partial E}{\partial o^l_j(x, y)} \frac{\partial o^l_j(x, y)}{\partial s^l_j(x, y)} \frac{\partial s^l_j(x, y)}{\partial g^l_{i,j}(x, y)} \frac{\partial g^l_{i,j}(x, y)}{\partial w^l_{i,j}(k)}. \tag{12}$$

Assuming the rectified linear unit (ReLU) is used as the activation function, we can rewrite Eq. (12) as

$$\frac{\partial E}{\partial w^l_{i,j}(k)} = \frac{\partial E}{\partial o^l_j(x, y)} \frac{\partial s^l_j(x, y)}{\partial g^l_{i,j}(x', y')} \frac{\partial g^l_{i,j}(x', y')}{\partial w^l_{i,j}(k)}. \tag{13}$$

Substituting the derivative obtained from (11) into (13) gives

$$\frac{\partial E}{\partial w^l_{i,j}(k)} = \frac{\partial E}{\partial o^l_j(x, y)}\, o^{l-1}_i(x', y') \frac{\partial g^l_{i,j}(x', y')}{\partial w^l_{i,j}(k)}. \tag{14}$$

Here, the partial derivatives of the Gabor function with respect to Gabor parameters γ and ϕ can be computed directly as

$$\begin{aligned}\frac{\partial g(x, y)}{\partial \gamma} &= \frac{1}{2\pi\sigma^2\gamma^2}\left(\frac{\tilde{x}^2}{\gamma^2\sigma^2} - 1\right) \exp\{-\frac{\frac{\tilde{x}^2}{\gamma^2} + \tilde{y}^2}{2\sigma^2}\} \\ &\quad \times \cos\{\frac{2\pi}{\lambda}(\tilde{x} + \phi)\},\end{aligned} \tag{15}$$

$$\frac{\partial g(x, y)}{\partial \phi} = -\frac{1}{\lambda\gamma\sigma^2} \exp\{-\frac{\frac{\tilde{x}^2}{\gamma^2} + \tilde{y}^2}{2\sigma^2}\} \sin\{\frac{2\pi}{\lambda}(\tilde{x} + \phi)\}. \tag{16}$$

For parameter θ, the partial derivative can be obtained using the chain rule as follows:

$$\begin{aligned}&\frac{\partial g(x, y)}{\partial \theta} \\ &= \frac{\partial g(x, y)}{\partial \tilde{x}} \frac{\partial \tilde{x}}{\partial \theta} = -\frac{\tilde{y}}{2\pi\gamma\sigma^2} \exp\{-\frac{\frac{\tilde{x}^2}{\gamma^2} + \tilde{y}^2}{2\sigma^2}\} \\ &\quad \times \left[\frac{2\pi}{\lambda} \sin\{\frac{2\pi}{\lambda}(\tilde{x} + \phi) + \frac{\tilde{x}}{\sigma^2}\left(\frac{1}{\gamma^2} - 1\right) \cos\{\frac{2\pi}{\lambda}(\tilde{x} + \phi)\right],\end{aligned} \tag{17}$$

Similar to (17), the partial derivative of the Gabor function with respect to parameter λ is given by

$$\begin{aligned}\frac{\partial g(x,y)}{\partial\lambda} &= \frac{\partial g(x,y)}{\partial\sigma}\,\frac{\partial\sigma}{\partial\lambda}\\ &= -\frac{1}{\gamma\lambda\sigma^2}\,\exp\{-\frac{\frac{\tilde{x}^2}{\gamma^2}+\tilde{y}^2}{2\sigma^2}\}\\ &\quad\times[\frac{\tilde{x}}{\lambda}\sin\{\frac{2\pi}{\lambda}(\tilde{x}+\phi)\}+\frac{1}{2\pi}(\frac{\frac{\tilde{x}^2}{\gamma^2}+\tilde{y}^2}{\sigma^2}-2)\cos\{\frac{2\pi}{\lambda}(\tilde{x}+\phi)\}].\end{aligned} \tag{18}$$

Using (4) and applying the chain rule, we obtain the partial derivative of the Gabor function w.r.t. parameter β:

$$\begin{aligned}\frac{\partial g(x,y)}{\partial\beta} &= \frac{\partial g(x,y)}{\partial\sigma}\,\frac{\partial\sigma}{\partial\beta} = \frac{\lambda}{\gamma\pi^2\sigma^3}\sqrt{\frac{\ln^3 2}{2}}\,\frac{2^\beta}{2^\beta-1}\\ &\quad\times(\frac{\frac{\tilde{x}^2}{\gamma^2}+\tilde{y}^2}{\sigma^2}-2)\exp\{-\frac{\frac{\tilde{x}^2}{\gamma^2}+\tilde{y}^2}{2\sigma^2}\}\cos\{\frac{2\pi}{\lambda}(\tilde{x}+\phi)\}.\end{aligned} \tag{19}$$

REFERENCES

[1] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[2] D. A. Mely and T. Serre, *Towards a Theory of Computation in the Visual Cortex*. Singapore: Springer, 2017, pp. 59–84.

[3] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1233–1258, Dec. 1987.

[4] R. P. Scobey and A. J. Gabor, "Orientation discrimination sensitivity of single units in cat primary visual cortex," *Exp. Brain Res.*, vol. 77, no. 2, pp. 398–406, Sep. 1989.

[5] D. B. Hamilton, D. G. Albrecht, and W. S. Geisler, "Visual cortical receptive fields in monkey and cat: Spatial and temporal phase transfer function," *Vis. Res.*, vol. 29, no. 10, pp. 1285–1308, Jan. 1989.

[6] R. A. Frazor, D. G. Albrecht, W. S. Geisler, and A. M. Crane, "Visual cortex neurons of monkeys and cats: Temporal dynamics of the spatial frequency response function," *J. Neurophysiol.*, vol. 91, no. 6, pp. 2607–2627, Jun. 2004.

[7] G. M. Boynton and E. M. Finney, "Orientation-specific adaptation in human visual cortex," *J. Neurosci.*, vol. 23, no. 25, pp. 8781–8787, Sep. 2003.

[8] F. Fang, S. O. Murray, D. Kersten, and S. He, "Orientation-tuned fMRI adaptation in human visual cortex," *J. Neurophysiol.*, vol. 94, no. 6, pp. 4188–4195, Dec. 2005.

[9] I. J. Goodfellow, A. Courville, and Y. Bengio, "Scaling up Spike-and-Slab models for unsupervised feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1902–1914, Aug. 2013.

[10] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[17] E. Coiras, Y. Petillot, and D. M. Lane, "Multiresolution 3-D reconstruction from side-scan sonar images," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 382–390, Feb. 2007.

[18] P. Hollesen, W. A. Connors, and T. Trappenberg, "Comparison of learned versus engineered features for classification of mine like objects from raw sonar images," in *Proc. Can. Conf. Artif. Intell.*, May 2011, pp. 174–185.

[19] J. Sawas and Y. Petillot, "Cascade of boosted classifiers for automatic target recognition in synthetic aperture sonar imagery," in *Proc. Meetings Acoust. ECUA*, vol. 17, no. 1, 2012, Art. no. 070074.

[20] C. Barngrover, R. Kastner, and S. Belongie, "Semisynthetic versus real-world sonar training data for the classification of mine-like objects," *IEEE J. Ocean. Eng.*, vol. 40, no. 1, pp. 48–56, Jan. 2015.

[21] C. Barngrover, A. Althoff, P. DeGuzman, and R. Kastner, "A brain–computer interface (BCI) for the detection of mine-like objects in sidescan sonar imagery," *IEEE J. Ocean. Eng.*, vol. 41, no. 1, pp. 123–138, Jan. 2016.

[22] D. Gebhardt, K. Parikh, I. Dzieciuch, M. Walton, and N. A. V. Hoang, "Hunting for naval mines with deep neural networks," in *Proc. IEEE OCEANS Conf.*, Sep. 2017, pp. 1–5.

[23] J. McKay, I. Gerg, V. Monga, and R. G. Raj, "What's mine is yours: Pretrained CNNs for limited training sonar ATR," in *Proc. IEEE OCEANS Conf.*, Sep. 2017, pp. 1–7.

[24] K. Denos, M. Ravaut, A. Fagette, and H.-S. Lim, "Deep learning applied to underwater mine warfare," in *Proc. Oceans Aberdeen*, Jun. 2017, pp. 1–7.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[26] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[29] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[31] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, Dec. 2001, pp. 511–518.

[32] Z. Zhu, X. Xu, L. Yang, H. Yan, S. Peng, and J. Xu, "A model-based sonar image ATR method based on SIFT features," in *Proc. Oceans TAIPEI*, Apr. 2014, pp. 1–4.

[33] P. Tueller, R. Kastner, and R. Diamant, "A comparison of feature detectors for underwater sonar imagery," in *Proc. Oceans MTS/IEEE Charleston*, Oct. 2018, pp. 1–6.

[34] J. Dale, A. Galusha, J. Keller, and A. Zare, "Evaluation of image features for discriminating targets from false positives in synthetic aperture sonar imagery," *Proc. SPIE*, vol. 11012, May 2019, Art. no. 110120A.

[35] X. Li, R. Jimmy, L. Ce, and J. Jiaya, "Deep convolutional neural network for image deconvolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1790–1798.

[36] J. P. Jones and L. A. Palmer, "The two-dimensional spatial structure of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1187–1211, Dec. 1987.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 630–645.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–6.

[39] J.-K. Kamarainen, V. Kyrki, and H. Kalviainen, "Invariance properties of Gabor filter-based features-overview and applications," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1088–1099, May 2006.

[40] P. B. Chapple, "Unsupervised detection of mine-like objects in seabed imagery from autonomous underwater vehicles," in *Proc. Oceans*, Oct. 2009, pp. 1–6.

[41] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[42] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.

HOANG THANH LE received the B.Eng. degree in computer science from Nha Trang University, Vietnam, in 2008, and the M.Sc. degree in computer science from the University of Queensland, Australia, in 2012. He is currently pursuing the Ph.D. degree in computer engineering with the University of Wollongong, Australia. His research interests include image processing, pattern recognition, machine learning, and computer vision.

SON LAM PHUNG (Senior Member, IEEE) received the B.Eng. degree (Hons.) in computer engineering and the Ph.D. degree in computer engineering from Edith Cowan University, Australia, in 1999 and 2003, respectively. He is currently an Associate Professor with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong. His research interests include image and signal processing, neural networks, pattern recognition, and machine learning. He was awarded the University and Faculty Medals, in 2000.

PHILIP B. CHAPPLE received the B.Sc. (Hons.) and Ph.D. degrees in experimental physics (multiphoton laser spectroscopy) from the Australian National University and the Master of Mathematical Sciences degree in signal and information processing. He is currently the Group Leader Littoral Systems Autonomy, Maritime Division of Defence Science and Technology (DST). His group conducts research in maritime autonomous systems to support naval capability development in mine hunting, hydrography, environmental assessment and blue-water naval operations. He has worked for 20 years in the areas of maritime survey, maritime rapid environmental assessment, mine countermeasures, sonar image processing (automatic target recognition), and maritime autonomous systems.

ABDESSELAM BOUZERDOUM (Senior Member, IEEE) graduated with M.S.E.E. and Ph.D. degrees in electrical engineering from the University of Washington, Seattle, USA. He is currently serving as the Head of Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Qatar, and a Senior Professor of Computer Engineering from the University of Wollongong. His research interests include radar imaging and signal processing, image processing, vision, machine learning, and pattern recognition.

CHRISTIAN H. RITZ (Senior Member, IEEE) received the B.Eng. degree in electrical engineering, and the B.Math. and Ph.D. degrees from the University of Wollongong, in 1998 and 2003, respectively. He is currently a Professor with the School of Electrical, Computer, and Telecommunications Engineering. His current research interests include signal and information processing for spatial audio, microphone arrays, sound classification, and sonar image classification.

LE CHUNG TRAN (Senior Member, IEEE) received the B.E. degree (Hons.) from the University of Transport and Communication, Vietnam, in 1997, the M.E. degree from the University of Science and Technology, Vietnam, in 2000, and the Ph.D. degree from the University of Wollongong, Australia, in 2006, all in telecommunications engineering. He is currently a Senior Lecturer with the School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong. His research interests include 5G, MIMO, space-time-frequency processing, WBANs, the IoT, biomedical engineering, ultra-wideband, millimetre wave, cooperative and cognitive communications, software defined radio, network coding, and digital signal processing for communications.