Robotica (2024), **42**, pp. 1231–1245
doi:10.1017/S0263574724000195

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

IFE-net: improved feature enhancement network for weak feature target recognition in autonomous underwater vehicles

Lei Cai[1], Bingyuan Zhang[2], Yuejun Li[2] and Haojie Chai[1]

[1]School of Artificial Intelligence, Henan Institute of Science and Technology, Xinxiang, PR China and [2]School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, PR China
Corresponding author: Lei Cai; Email: cailei2014@126.com

Received: 17 July 2023; **Revised:** 11 January 2024; **Accepted:** 14 January 2024; **First published online:** 8 February 2024

Keywords: underwater target recognition; unbalanced category data; spatial and semantic feature enhancement; multi-scale feature comparison; ranking loss

Abstract

The recognizing underwater targets is a crucial component of autonomous underwater vehicle patrols and detection efforts. In the process of visual image recognition in real underwater environment, the spatial and semantic features of the target often appear to different degrees of loss, and the scarcity of specific types of underwater samples leads to unbalanced data on categories. This kind of problem makes the target features appear weak and seriously affects the accuracy of underwater target recognition. Traditional deep learning methods based on data and feature enhancement cannot achieve ideal recognition effect. Based on the above difficulties, this paper proposes an improved feature enhancement network for weak feature target recognition. Firstly, a multi-scale spatial and semantic feature enhancement module is constructed to extract the feature information of the extraction target accurately. Secondly, this paper solves the influence of target feature distortion on classification through multi-scale feature comparison of positive and negative samples. Finally, the Rank & Sort Loss function was used to train the depth target detection to solve the problem of recognition accuracy under highly unbalanced sample data. Experimental results show that the recognition accuracy of the proposed method is 2.28% and 3.84% higher than that of the existing algorithms in the recognition of underwater fuzzy and distorted target images, which demonstrates the effectiveness and superiority of the proposed method.

1. Introduction

Underwater target detection has a wide range of applications in marine environment monitoring and safety, and most existing target detection algorithms are implemented on Autonomous Underwater Vehicle (AUV). These AUV play a crucial role in enhancing marine monitoring and maintenance efforts [1]. Due to the influence of complex sea conditions and refraction of light transmission, the images obtained by AUV and other underwater vision equipment have weak features such as feature blur, loss and distortion. Most of the existing methods based on deep network solve the problem of weak feature target recognition by image deblurring, feature co-occurrence relationship for feature correction [2], distortion correction [3, 4] and other methods. However, this kind of algorithm can easily lead to the loss of original spatial and semantic information of small targets [5, 6], and increase the relevant modules and computing volume of the algorithm [7].

The acquisition of images of different targets in underwater environments by underwater vision devices, such as autonomous underwater vehicles, can be challenging, leading to an uneven distribution of labels in the training set. As a result, most targets lack training data and labels, which makes it difficult for the algorithm to accurately identify such targets with low signal-to-noise ratio. The existing

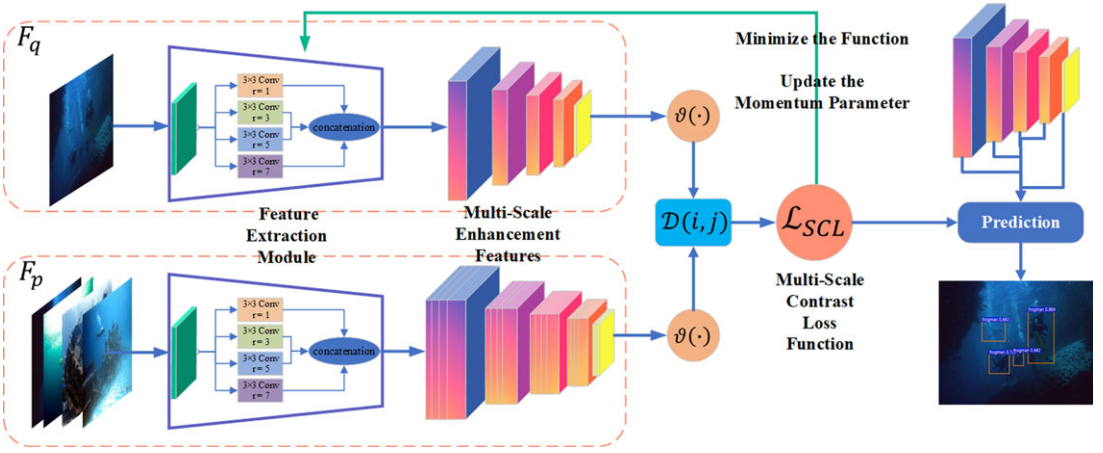

Figure 1. *General framework of improved feature enhancement network for weak feature target recognition.*

solution is to augment data, extract significant features from data images [6] and optimize data sample expansion and network framework. Although this kind of method can effectively improve the recognition accuracy, it will also cause the loss of basic correlation features of sample images in the recognition process. Therefore, in this case, the improvement effect of the algorithm on the recognition accuracy is limited. Graph convolutional networks and recurrent relationship trees play a key role in image semantic feature extraction [8, 9]. Unsupervised presentation learning can distinguish different targets by means of feature similarity measurement, which is helpful for downstream classification and recognition tasks and can effectively solve the above problem.

This paper proposes an improved feature enhancement network method for underwater target recognition with weak features. As shown in Fig. 1, the method proposed in this paper builds an identification network by combining spatial and semantic enhancement, unsupervised feature contrast extraction and Rank & Sort (RS) Loss, which is robust to unbalanced samples. The network achieves high recognition accuracy under the condition of uneven target samples and distorted images. The integration of image recognition algorithms with AUV can significantly enhance the inspection accuracy and efficiency of AUV.

The main contributions of this paper are as follows:

1. This paper introduces a multi-scale spatial and semantic feature enhancement module. This module enhances the extraction of multi-scale spatial features by utilizing feature mapping with self-learning parameters. It integrates semantic information from various sensory inputs to accurately extract target feature information.
2. A multi-scale feature comparison module is developed in this paper to address issues related to target feature blurring and distortion. It achieves this by conducting multi-scale feature comparisons between positive and negative samples. This network module is designed to ensure the robustness of the recognition network, especially in scenarios involving small sample sizes for underwater target recognition.
3. We propose a multi-scale target sorting detector based on RS Loss to train the deep target recognition network, particularly to address challenges associated with highly unbalanced sample data. We conducted comprehensive experiments, and the results demonstrate the effectiveness of our method in significantly improving detection accuracy, particularly for weakly characterized targets.

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

2. Related work

The integration of deep learning techniques into underwater vehicles for detection and identification represents a pivotal research domain in marine maintenance. Sun [1] introduced a kernelized correlation filter tracker and a novel fuzzy controller, which were trained using a deep learning model, resulting in favorable outcomes in visual tracking of underwater vehicles. Chu [10] utilized deep reinforcement learning based on a double-deep Q-network for autonomous underwater navigation, leading to effective path planning and obstacle avoidance. Tang [11] developed a deep learning-based underwater target detection model and a real-time underwater target detection method, effectively addressing the challenges associated with side-scan sonar recognition.

Refraction of light caused the loss of image features, making it more difficult for the algorithm to accurately identify the target. Rich semantic information is a prerequisite to achieve the classification task, while accurate spatial information is a prerequisite to achieve the localization task for target detection [7]. Cai [12] constructed a target recognition network by multi-intelligence collaboration, multi-view optical field reconstruction and migration reinforcement learning to enhance target feature data from both data sources and feature frameworks, with significant improvements in simplified computation and target recognition accuracy. Lin [13] employed a planar detection algorithm based on the semantic web to construct planar odometry for robots operating in structured environments. Inflationary convolution allows the extraction of sufficient semantic information by expanding the field of perception without changing the size of the feature map [14]. Cai [15] addressed the problem of distorted target recognition accuracy by supplementing missing salient features with spatial semantic information. Rabbi [16] used a super-resolution method for images to zoom in on objects, effectively solving the problem of feature loss and improving the detection performance of small targets in remote sensing images from the use of edge enhancement. The methods mentioned above excel at extracting detailed semantic information through deep features; however, they exhibit shortcomings in effectively utilizing spatial information and are less suitable for data characterized by imbalanced categories.

Multi-scale features contain richer spatial and semantic feature information and are beneficial for identifying weak targets at very low signal-to-noise ratios. Ma [17] fuses advanced feature detail enhancement and multi-scale features for contextual semantic edge detection. Ju [18] proposes the attention mechanism for multi-scale target detection by adaptively learning the importance and relevance ideas of features at different scales with the Lee [19] computed enhanced feature extraction networks by contrast learning. To cope with scale variations. Kuang [20] utilized semantic information extraction and environment matching to enhance the localization capabilities of mobile robots, enabling them to operate more efficiently in their environment. Cai [21] conducted fuzzy small target feature extraction by combining hybrid dilation convolution with multi-scale features. Liu [22] learns object features and contextual feature weights based on upsampling to fuse multi-layer feature maps to improve the detection performance of small object detection performance. Douadi [23] developed a method for rapidly and precisely constructing navigation maps for robots by modeling spatially stabilized keypoints. Reference [24] designed a shallow feature enhancement module to enhance the representation of weak feature objects with the help of rich contextual information. Since reusing feature information leads to unclear weights, Fang [25] proposed a densified, lightweight top-down network structure for effective integration of multi-scale features. Gupta [26] pointed out that the drawback of multi-scale feature methods is mainly that spatial details are ignored until the final fusion stage, so each channel in the aggregated features is weighted according to the adjacent layers to enhance the distinguishing power of the feature representation. In summary, reducing the loss of spatial information while ensuring the acquisition of strong semantic information is the key to improving classification and localization accuracy. While the methods described above make comprehensive use of spatial information, they may not be suitable for datasets with semantic information loss and category imbalance. Balancing the preservation of spatial information while ensuring the acquisition of robust semantic information is pivotal for enhancing classification and localization accuracy.

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

For the problem that accurate recognition cannot be achieved for specific categories with complex features and uneven sample size, Wang [27] used similarity constraints to capture the intrinsic connection between available information and feature weights, and fusion ranking loss to capture the dependency between labels. Zhi [28] proposed an end-to-end convolutional neural network based on multi-path structure. Gao [29] used a multi-category attention region module that maintains the diversity of feature data in the attention region. Jiang [30] extracts multi-scale features and learns multi-scale relationships between samples, which can alleviate the lack of performance of cross-entropy loss in the case of small samples. Khosla [31] proposes RS Loss based on the properties of ranking, which can train models in the presence of highly unbalanced data.

In summary, while these methods exhibit advantages in specific aspects of target detection and under particular scenarios, they fail to provide an effective solution to the challenges of spatial and semantic feature loss, especially when dealing with the scarcity of underwater class-specific samples [32]. To tackle these issues, this thesis introduces an improved feature enhancement network for enhancing the recognition accuracy of weak feature targets. Specifically, in situations where spatial and semantic features suffer varying degrees of loss, this thesis incorporates a spatial and semantic enhancement feature extraction module to enhance feature extraction. Moreover, to address the challenges arising from a shortage of samples in specific categories, often accompanied by feature blurring or distortion, this paper introduces a multi-scale feature comparison and selection module, aimed at precisely recognizing weak feature targets. By arranging the algorithm of this paper on the underwater vehicle, we can improve the inspection accuracy and efficiency of the underwater vehicle.

3. Proposed method

This section provides a comprehensive description of the overall architecture of the proposed improved feature enhancement network for recognizing weak feature targets. It comprises three primary components: a spatial and semantic enhancement feature extraction module, a multi-scale feature comparison selection module, and a multi-scale target ranking detector. The network architecture in this paper is configured with hyperparameters, such as convolutional kernel size and the number of feature channels, following the ResNet50 reference model.

3.1. Spatial and semantic enhanced feature extraction module

The weak light and complex background in underwater environment lead to the weak intensity of image objects acquired by Autonomous Underwater Vehicle (AUV), and the loss of spatial and semantic features to varying degrees, which seriously affects the accuracy of target recognition. The semantic information is sparse in the shallow feature map extracted by the traditional classification network, while the semantic information is rich in the deep feature map, but the spatial information decays seriously. To solve the above problems, In this paper, we introduce the spatial and semantic enhancement feature extraction module to improve the extraction of spatial and semantic features from the original classification network.

This article uses ResNet50 as the backbone network to extract features. The feature extraction network module has an overall composition of 6 layers of ResNet blocks. The input is a 256×256 pixel size image x. The features extracted by each ResNet block are denoted as $f_q(x)$. The multi-scale features of the image are extracted with different ResNet blocks $H = h_0, h_1, \ldots, h_i, i < 6$. The different scale feature maps obtained from each ResNet block are used as input to obtain a new multi-scale feature map enhanced with spatial and semantic information by the spatial and semantic feature enhancement module. This module is divided into two parts, spatial feature enhancement and dense upsampling. The overall structure of the spatial and semantic enhancement feature extraction module is shown in Fig. 2.

The spatial features are enhanced in a multi-branch parallel structure, and the context information of the target is extracted by enlarging the receptive field, which enriches the semantic information and

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

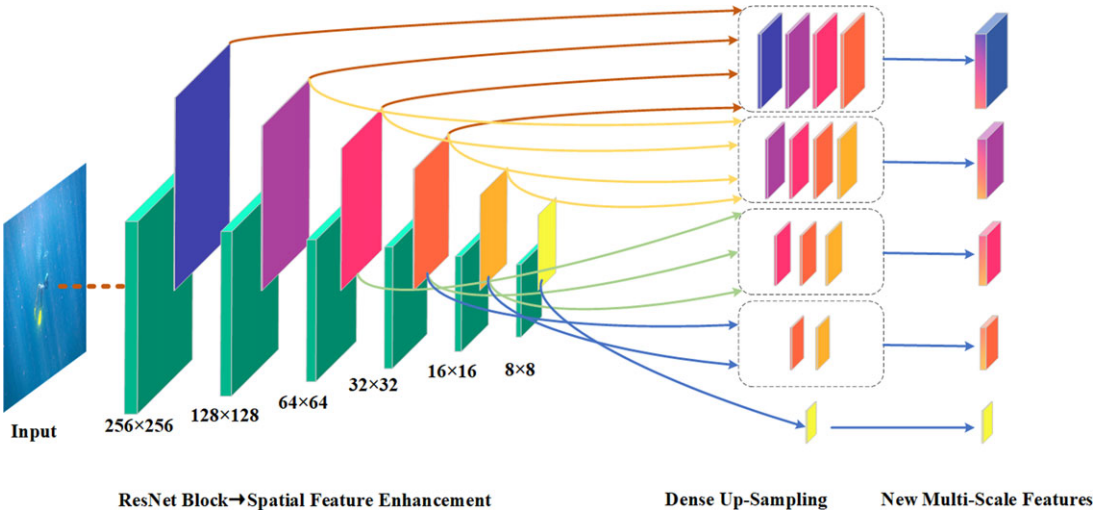

Figure 2. *Network framework diagram of spatial and semantic enhanced feature extraction module.*

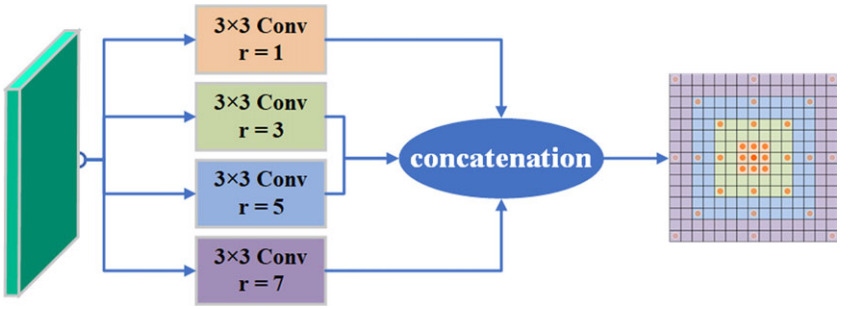

Figure 3. *Spatial feature enhancement of multi-branch parallel structures and their perceptual field representation.*

avoids the attenuation of spatial information as much as possible. This network can enhance the ability to extract spatial and semantic features. Multi-branch formal structure extracts the features of different receptive fields. As shown in Fig. 3, each branch uses expansion convolution with different expansion rates to carry out convolution operations on the same feature graph, so as to obtain isoscale feature graphs with different receptive fields. The features extracted from different receptive fields were restored to the original size and number of channels by cascade fusion.

Dense upsampling is then used to generate a new multi-scale feature map. Dense upsampling prevents information decay, and the feature map obtained by upsampling not only uses a top-down approach to transfer information layer by layer but also adds a form of direct transfer to complement the previous feature map. The multi-scale feature selection module of the reference [24] was used to transfer the semantic information of the 16×16 feature map to the 64×64 feature map as an example. It is structured through two parallel branches. In one, a top-down approach, the 16×16 feature map is first sampled to 32×32, fused with the original 32×32 feature map, with the size of 32×32 as the next sample, and then sampled to 64×64 and 128×128 in turn. In the other branch, it is directly sampled as 128×128, passed to the feature map of the corresponding size, which is used to guide the enhancement of shallow semantic features and avoid information loss.

Taking into account the impact of various texture details and objects on recognition results, the multi-scale model must determine the suitable size and number of feature maps for fusion. Each feature map of the multi-scale feature model is assigned a learnable parameter ω_x, and each feature map is adaptively

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

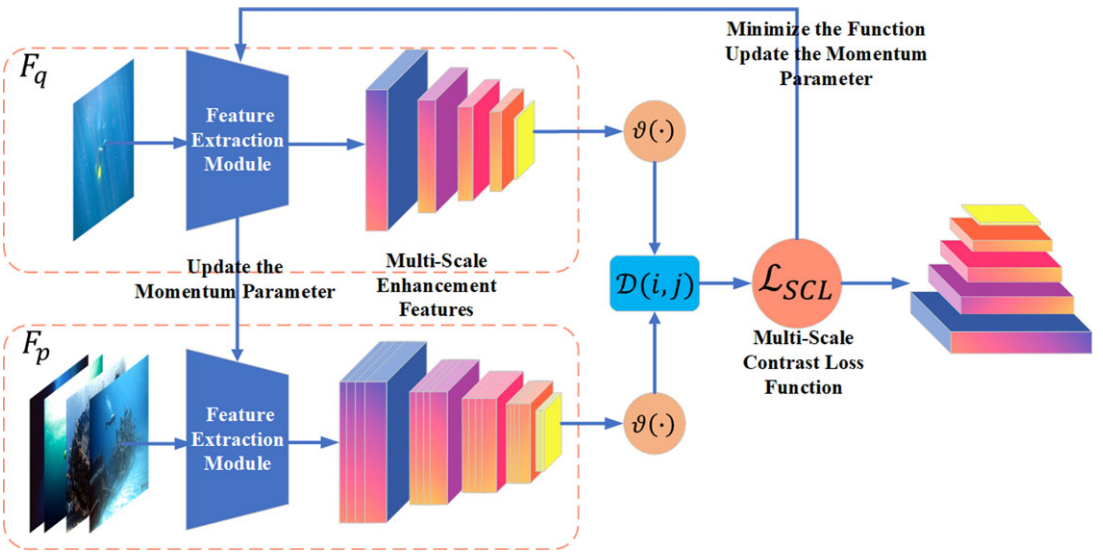

Figure 4. *Multi-scale feature comparison selection module structure.*

merged by training to find the optimal proportion of each feature map in the fusion. This design avoids the degradation of feature extraction performance caused by the decay of spatial information. The element summation method first adjusts feature maps of equal scale to the same number of dimensions and then sums the corresponding elements. When fusion is complete, it then reverts to its original dimensions. The multi-scale enhanced features of the image are represented as:

$$F = \{f_{e0}(x), f_{e1}(x), f_{e2}(x), f_{e3}(x), f_{e4}(x)\} \tag{1}$$

where x is the input test image. $f_{ei}(x)$ is the multi-scale enhancement feature output of the completed training enhancement feature extraction network.

3.2. Multi-scale feature comparison selection module

The shortage of samples in specific categories is compounded by issues like feature blurring or distortion. This challenging scenario results in the low-accuracy recognition of weak feature targets. To address this, the multi-scale feature contrast selection module characterizes feature contrast metrics through unsupervised learning. This, in turn, furnishes multi-scale prominent features and reference sample data to assist the target detector in accurately recognizing weak feature targets.

Data expansion by image distortion, blurring, cropping, flipping, etc. In this paper, the positive samples are the original example samples and the samples obtained by data expansion, while the negative examples are all other categories of strength samples in the sampled training batch and the expanded samples. The goal of feature comparison is to learn the similarity representation of samples. It should minimize or keep constant the feature distance for different pairs of positive examples while maximizing the feature distance between pairs of negative examples. The objective of the method is to select sample features from a series of labeled sample images $X = x_1, x_2, \ldots, x_n$ that are close to the instance image features, thus complementing the semantic information of the instance features.

The entire multi-scale feature comparison selection module contains two encoder branches F_q and F_p, as shown in Fig. 4, with both branches applying the same spatial and semantic enhancements to the feature extraction network. The feature mapping module $\vartheta(\bullet)$, which connects one after each feature-enhanced scale feature map, passes the feature vector $f_{ei}(x)$ through a fully connected layer $\vartheta(\bullet)$. Here, the feature mapping $\vartheta(\bullet)$ maps the multi-scale enhanced feature map of the image to the feature space vector $z = z_1, z_2, z_3, z_4, z_5$. The similarity measure of the feature space is then performed by the cosine

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

similarity function $D(i, j)$ of the samples. Where x_i and x_j are the example image representations and the image representations in the training samples of the comparison batch, respectively.

In this paper, the multi-scale feature contrast loss used SCLLoss [33] function to train the two-branch space with semantically enhanced feature extraction model $F_\tau(\bullet)$ and feature mapping $\vartheta(\bullet)$. The contrast loss function is defined as follows:

$$\mathcal{L}_{MECL} = -\frac{\beta}{N}\sum_{i=1}^{N}\sum_{c=1}^{c} y_{i,c}\cdot \log \hat{y}_{i,c} - \sum_{i=1}^{N}\frac{\lambda}{N_{y_i}-1}\sum_{j=1}^{N}\log\frac{exp(\vartheta(x_i)\cdot\vartheta(x_j)/\tau)}{\sum_{k=1}^{N} exp(\vartheta(x_i)\cdot\vartheta(x_j)/\tau)} \tag{2}$$

where N is the image batch size of the F_p structure input, the input is the training image of $x_i, y_{i i=1,\ldots,N}$ of the training images. x_i is the image sample and x_j is the label corresponding to the image sample. N_{y_i} is the total number of instances in the batch with the same label as y_i. x_i denotes the positive sample images in the image training batch. x_j is the negative sample images in the batch. $y_{i,c}$ denotes the input instance images of label data, and $\hat{y}_{i,c}$ denotes the probability that the model output instance is of category C. β and λ are a scalar-weighted hyperparameter,$\beta = 1 - \lambda$. *tau* is an adjustable scalar temperature parameter that controls feature class differentiation.

The construction of salient features through image feature space comparison not only makes the number of negative samples larger but also improves the training effect. However, this approach presents challenges in terms of iteratively updating the encoder F_p. To address this, we utilize the momentum updating approach introduced in literature [15]. This approach dynamically updates encoder F_p using encoder F_q. The parameter of encoder F_q is denoted as ϑ_q, and the updating method is stochastic gradient descent. The parameter of encoder F_p is denoted as ϑ_p, and the update method is $z\vartheta_p + (1 - \vartheta_q)$, with momentum coefficients $z \in [0, 1)$. ϑ_p updates in accordance with the changes in ϑ_q. When the training of the multi-scale feature contrast selection module is completed, the multi-scale space and semantic features are extracted from the multi-scale enhanced features F'_e obtained by branching F_q of the network, and the multiple similar feature labels $y_{i,c}$ obtained by branching F_p. F'_e and $y_{i,c}$ were used to guide through the subsequent target recognition tasks.

3.3. Multi-scale target sequencing detector

To address the challenge of imbalanced samples in specific underwater categories, this paper presents an improved feature enhancement network for weak feature target recognition. This approach is based on RS Loss and constructs a multi-scale target sorting detector to perform recognition tasks across various scales. The recognition network ultimately achieves the most accurate results by continuously updating the feature comparison selection and prediction rankings.

Firstly, 5 scales of spatial and semantic enhancement features were input to the target detector to generate a series of Anchor Boxes and prediction category information. Then, an update sorting classification with continuous Intersection over Union (IoU) prediction task is performed. RS loss $\mathcal{L}_{RS}$ [31] was used for the classification task, and the continuous prediction category information data consisted of two parts, the prediction categories generated by the instance image after the target detector and the multiple similar feature labels $y_{i,c}$ category data obtained in the multi-scale feature comparison selection module. The $\mathbf{IoU}(\hat{b}_i, b_i)$ between the predicted position box $\hat{b}_i$ and the true position box of the data (b_i) is used as the continuous label. In the continuous IoUs prediction task, K-means is first used to generate Anchor Boxes with larger IoU values than ground truth, and the clustering centers of Anchor Boxes are overlapped with the centroids of ground truth to select the target candidate bounding boxes. The minimum outer rectangle of the two prediction boxes is then calculated by Distance-IoU, which is used to characterize the distance between the two target boxes. To compute the loss $\mathcal{L}_r$ of the multi-scale target sorting detector, we define x_i as the real label of the instance image and x_j as the label of multiple similar feature samples. The RS Loss, denoted as $\mathcal{L}_{RS}$, is the difference between the current sorting error term $\ell_{RS}(x_i)$ and the desired sorting error term $\ell^*_{RS}(x_i)$. It also includes the overall average error term for updating sorting, given by $(\ell_S(x_j) - \ell^*_S(x_j))p_S(x_i \mid x_j)$, where $\ell_S(x_j) - \ell^*_S(x_j \geq 0)$ represents the positive

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

case signal enhancement term for $x_i \in P$, and $(\ell_S(x_j) - \ell_S^*(x_j))p_S(x_i \mid x_j) \leq 0$ represents the signal degradation term for $x_j \in P$. Here, $p_S(x_i \mid x_j)$ is the category probability mass function based on the probability mass function (pmf) of $\ell_{RS}(x_i)$. The RS Loss $\mathcal{L}_{RS}$ is integrated into the multi-scale sorted target detector loss $\mathcal{L}_r$, which is calculated as follows:

$$\mathcal{L}_{RS} = \frac{1}{|\mathcal{P}|} \sum_{x_j \in \mathcal{P}} (\ell_{RS}(x_i) - \ell_{RS}^*(x_i) + (\ell_S(x_j) - \ell_S^*(x_j))p_S(x_i \mid x_j)) \tag{3}$$

$$\mathcal{L}_r = \mathcal{L}_{RS} + \frac{|\partial \mathcal{L}_{RS}|}{|\partial \hat{s}|} \Big/ \frac{|\partial \mathcal{L}_{RS}|}{|\partial \hat{b}|} \mathcal{L}_{box} \tag{4}$$

The multi-scale target sorting detector loss, denoted as $\mathcal{L}_r$, combines RS Loss $\mathcal{L}_{RS}$ and IoU Loss. The iterative parameter for IoU Loss, represented as $\frac{|\partial \mathcal{L}_{RS}|}{|\partial \hat{s}|} / \frac{|\partial \mathcal{L}_{RS}|}{|\partial \hat{b}|}$, is computed using a magnitude-based tuning algorithm based on the L1 paradigm. Here, $\hat{b}$ and $\hat{s}$ denote the predicted boundaries, encompassing both box regression and classification header outputs. $\mathcal{L}_{box}$ corresponds to the IoU Loss.

4. Experimental

The experiments were conducted on a small server with 64G of memory and dual GPUs configured with RTX 3090Ti. Model algorithm simulations were performed on a Pytorch platform configured under Ubuntu 18.04.

4.1. Experimental dataset

The data for training and testing of the method in this paper are extracted from various sources, including CADDY, Underwater Image Enhancement Benchmark, Underwater Target dataset (UTD) and self-collected images. Among them, the UTD is a self-constructed public underwater dataset by the author's team. This dataset, which is publicly available on GitHub at https://github.com/Robotics-Institute-HIST/Dataset.git, consists of 2616 images across various object categories, including AUV, submarine, Frogman and torpedo. These images encompass various environments, including turbid water, uneven light, overlapping targets and others, resulting in a dataset rich in diversity.

Our team also collected self-acquired images using AUV in different underwater scenarios with varying target categories, attitudes, and lighting environments. We obtained 200 effective images that were supplemented to the training process to enhance the diversity of the dataset and improve the model's generalization ability. Specifically, in the multi-scale feature comparison module, our self-acquired dataset facilitates positive sample selection and accelerates the learning of target feature similarity. To ensure the applicability of the model, it was deployed on our team's AUV to perform real-time recognition of self-mined images. This approach verified the accuracy and generalization ability of the model's recognition effect. The target recognition algorithm is effectively integrated with the autonomous underwater vehicle to enhance the inspection accuracy and efficiency of the underwater vehicle. Using a combination of three publicly available datasets and self-collected images in equal ratio, this paper trained and tested the spatial and semantic enhancement feature extraction model and target recognition network using 1200 labeled images. The datasets were divided into training and testing sets in a 7:3 ratio.

4.2. Implementation details

For model training, an Adaptive Learning Rate (AdamW) optimizer was used to train the recognition method model. On the server side, the entire training process was iterated 50,000 times with the initial learning rate set to 0.01, and the weights decayed to 0.0001. The default numbers of suggested frames suggested features and iterations were 1000, 1000 and 60, respectively.

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

Figure 5. *Obvious underwater recognition of specific categories of targets.*

4.3. Experimental results

Experiments were conducted to validate the recognition of underwater targets by an improved feature enhancement network method for weak feature target recognition. As for different types of underwater weak feature target images, four sets of simulation experiments are designed in this section to verify the effectiveness of the proposed algorithm and compare it with MR-CNN [22], SLMS-SSD [24], DNTDF [25] and ALMNet [26]. Experiments in this paper have been conducted to recognize several specific classes of target images, and four of them, Torpedo, Submarine, Frogman and AUV, were selected for analysis of recognition accuracy in this paper. The algorithms were evaluated in terms of category confidence, average recognition accuracy (mAP), and the number of frames per second (FPS) that the method network can process.

The first experiments were conducted with well-featured underwater category-specific target recognition, and the visualization results are shown in Fig. 5. In terms of position frame and category confidence,

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

Table I. *Obvious data on the results of underwater category-specific target identification.*

Method	Torpedo	Submarine	Frogman	AUV	mAP	FPS
MR-CNN	0.8016	0.9549	0.8682	0.9326	0.8893	18
DNTDF	0.8242	0.9704	0.8597	0.9383	0.8982	16
ALMNet	0.8858	0.9623	0.8762	0.9574	0.9204	23
SLMS-SSD	0.8672	0.9862	0.8688	0.9568	0.9198	42
OURS	0.9062	0.9835	0.8943	0.9532	0.9343	29

Table II. *Results data for recognition of underwater blurred images.*

Method	Torpedo	Submarine	Frogman	AUV	mAP	FPS
MR-CNN	0.7253	0.8052	0.7632	0.8289	0.7807	14
DNTDF	0.7349	0.8259	0.7548	0.8461	0.7904	16
ALMNet	0.7651	0.8143	0.7739	0.8573	0.8027	20
SLMS-SSD	0.7632	0.8758	0.7713	0.8616	0.8180	36
OURS	0.7843	0.8716	0.8132	0.8942	0.8408	26

Table III. *The recognition time and accuracy of distorted underwater image.*

Method	Torpedo	Submarine	Frogman	AUV	mAP	FPS
MR-CNN	0.6432	0.7572	0.6879	0.7668	0.7138	9
DNTDF	0.6876	0.7685	0.6982	0.7559	0.7201	11.5
ALMNet	0.6493	0.7592	0.6851	0.8026	0.7241	19
SLMS-SSD	0.6658	0.7664	0.7072	0.8159	0.7388	32
OURS	0.7182	0.7943	0.7408	0.8556	0.7772	27

the comparison algorithm generally suffers from small target misses with weak features when it comes to frogman target image recognition. In contrast, the algorithm in this paper is able to recognize more accurately and with higher accuracy than the comparison algorithm. The underwater clear image test set recognition result data is shown in Table I, with the first four columns showing the accuracy in each category (same as Tables II and III). The bold black font in the table indicates the outstanding metrics of each algorithm. The data show that this paper method has the highest recognition accuracy in both torpedo and frogman target categories recognition with 0.9062 and 0.8943 respectively. Among the recognition results of the AUV test sample images, SLMS-SSD method has the highest accuracy of 0.9862. Our paper method is only second to SLMS-SSD method accuracy of 0.0027. In the AUV test sample image recognition results, ALMNet has the best recognition result with an accuracy value of 0.9574. Its value is only higher than that of the present method at 0.0042. and the present method has the highest mAP value compared to the comparison algorithms with a value of 0.9343. In terms of algorithm speed comparison, SLMS-SSD exhibits the fastest recognition speed, boasting impressive FPS data of up to 42. While the method proposed in this paper cannot rival SLMS-SSD in speed, it nevertheless demonstrates a speed advantage when compared to other methods. Moreover, the method presented in this paper achieves the highest mean Average Precision (mAP) value, highlighting its effectiveness and accuracy.

The next step was to conduct experiments on target recognition of underwater blurred images. The recognition accuracy and visualization results of each method for blurred images are shown in Fig. 6 and Table II. Due to the interference of image blurring on the features, the recognition accuracy of each method is significantly reduced. The recognition data show that the mAP of the method in this paper is 0.8408, which is the highest when comparing methods, and is 0.0202 higher than the mAP value

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

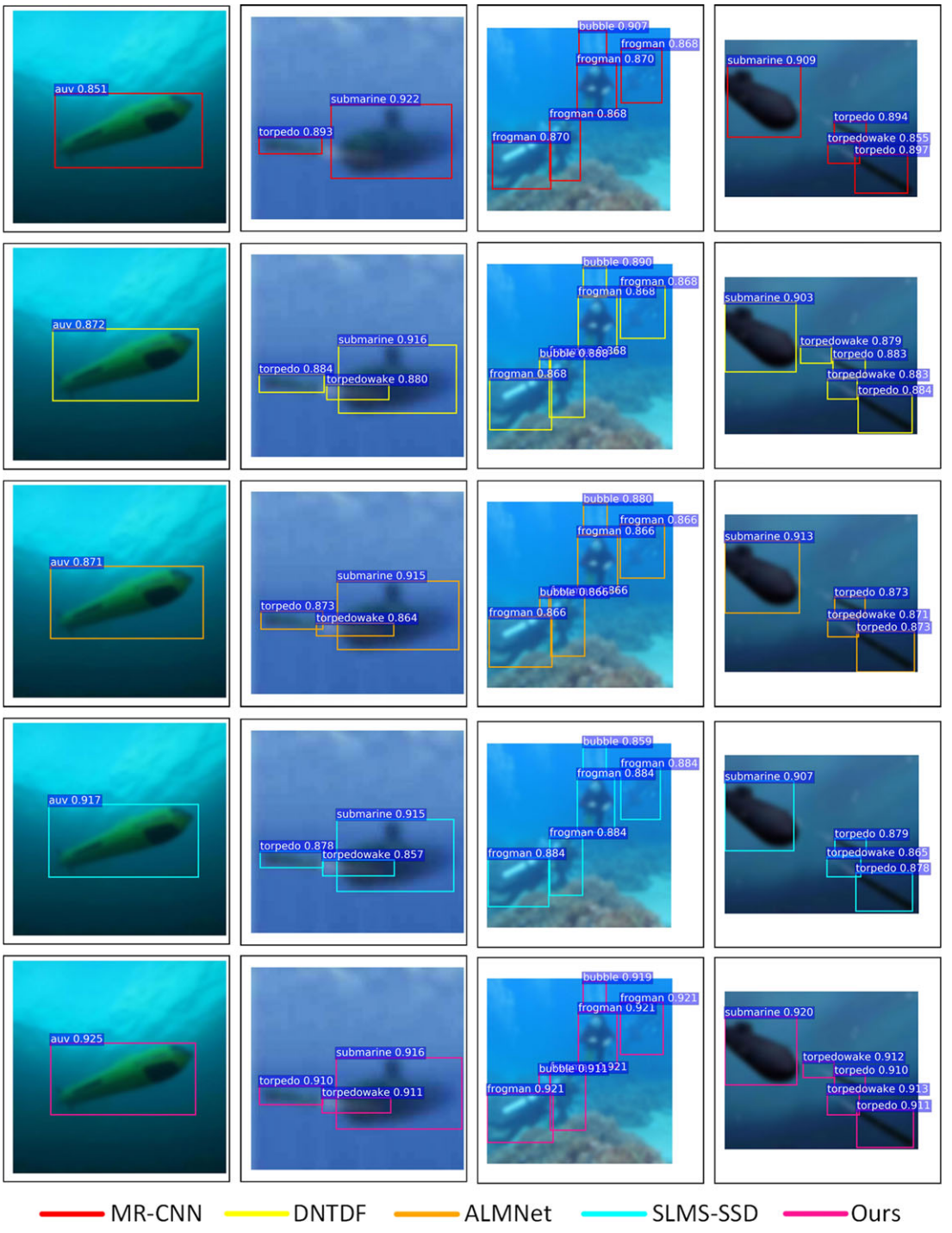

Figure 6. *The target recognition results of underwater blurred images.*

of the second-place SLMS-SSD method. mAP of the SLMS-SSD method is 0.8758, but only 0.0042 higher than the algorithm of this paper. Among the remaining category recognition results shown, the algorithm of this paper is relatively optimal. The recognition speed of this method is slightly inferior to that of the SLMS-SSD method, but it is 2.28% ahead in recognition accuracy.

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

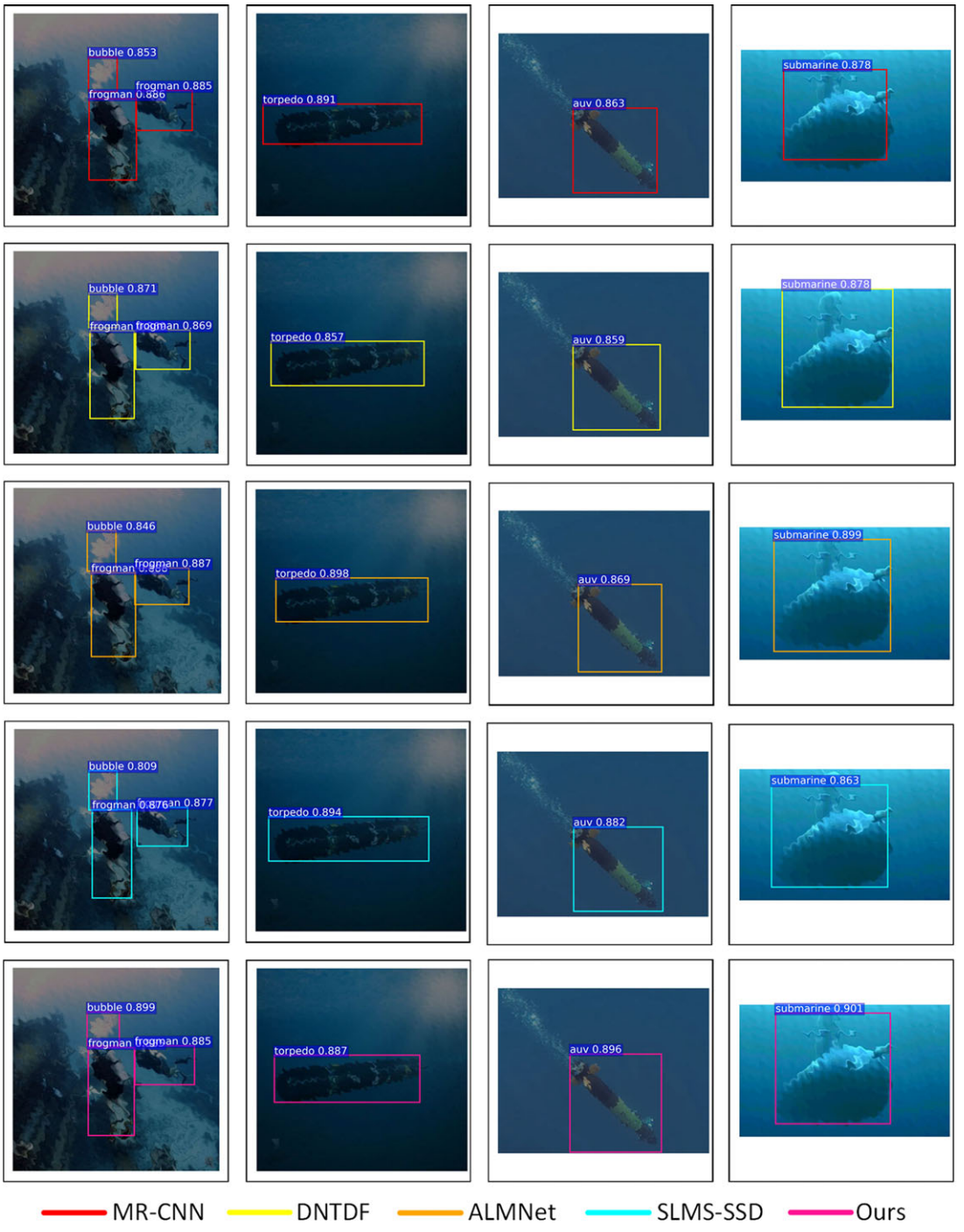

Figure 7. *The recognition results of underwater distorted image.*

In addition, experiments were carried out for the recognition of underwater distorted images. The recognition accuracy and visualization results for each method in the experiments are shown in Fig. 7 and Table III. The image distortion and blurred feature data were affected differently and the relative recognition had a different impact on the results. Different methods have different optimization strategies

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

Table IV. *Ablation experiment identification result data.*

Method	Torpedo	Submarine	Frogman	AUV	mAP	FPS
Backbone	0.7113	0.7953	0.7672	0.8501	0.7809	34
OURS-1	0.7438	0.8184	0.7805	0.8650	0.8019	31
OURS-2	0.7353	0.8027	0.7879	0.8674	0.7983	31
OURS	0.7597	0.8428	0.7971	0.8772	0.8192	29

according to their own recognition contexts and therefore have different performance in the recognition results. As can be seen from the data of recognition results in the table, the mAP of this paper's method is 0.7772 in recognizing underwater distorted targets. It still has the highest mAP value among the compared algorithms and is 3.84% higher than the SLMS-SSD method which is in the second place in terms of accuracy values. Our method has only the highest accuracy value in the comparison of recognition accuracy values for each category of targets. In addition, in terms of the recognition speed of each method for underwater distorted target images, our method is second only to the SLMS-SSD method with the highest FPS value.

In the three different experiments mentioned above, this paper's method increases the computing time and makes the number of negative samples larger to improve the training effect by constructing salient features through image feature space comparison in the multi-scale feature comparison selection module. With a slight decrease in FPS compared to the SLMS-SSD method, the proposed method significantly enhances the recognition accuracy of images.

Finally, we conducted an ablation experiment to determine whether the spatial and semantic feature enhancement module and the multi-scale enhancement feature comparison selection module, contribute to improved feature extraction performance. ResNet50 was chosen as the backbone feature extraction network, and the results were validated on the dataset. The experimental findings are presented in Table IV. We used ResNet50 as the backbone for feature extraction. OURS-1 represents the recognition results with the addition of the spatial and semantic feature enhancement module, while OURS-2 represents the results with the inclusion of the multi-scale enhancement feature comparison selection module. OURS is the recognition result obtained by fusing both modules. As shown in Table IV, adding the spatial and semantic feature enhancement module to the backbone network led to a 0.0210 increase in mAP, demonstrating that the multi-scale feature fusion and inflationary convolution operation can enhance spatial and semantic feature extraction capabilities. Incorporating the enhanced feature comparison selection module into the multi-scale backbone network resulted in a 0.0174 increase in mAP, confirming that the feature comparison selection module can enhance accuracy in recognizing distorted targets within imbalanced categories. Combining these two modules in experiments led to the highest mAP, showcasing the significant mutual enhancement of the designed modules, collectively improving the detection performance of weak feature target recognition.

5. Conclusion

Aiming at the difficulties such as serious interference in underwater environment and the difficulty of data acquisition by equipment and the scarcity of specific types of samples, this paper proposes an improved feature enhancement network method for weak feature target recognition. In this method, a multi-scale spatial and semantic feature enhancement module is constructed to extract the feature information of the extraction target accurately. Secondly, the influence of target feature distortion on classification is solved by the multi-scale feature comparison of positive and negative samples. Finally, RS Loss based on ranking is integrated to train the depth target recognition network and solve the problem of recognition accuracy under highly unbalanced sample data. This approach enhances the inspection accuracy and efficiency of underwater vehicles. The accuracy of the proposed method is 2.28% and 3.84% higher than that of the existing algorithms.

https://doi.org/10.1017/S0263574724000195 Published online by Cambridge University Press

Financial support. The work was financially supported by the Key Research and Development Program of Henan Province (231111220700) and the Major Science and Technology Program of Henan Province (221100110500).

Competing interests. The authors declare that there are no competing interests regarding the publication of this article.

Ethical standards. None.

Author contributions. Lei Cai conceived and designed the study. Bingyuan Zhang, Yuejun Li and Haojie Chai conducted data gathering. Bingyuan Zhang wrote the article.

References

[1] C. Sun, Z. Wan, H. Huang, G. Zhang, X. Bao, J. Li, M. Sheng and X. Yang, "Intelligent target visual tracking and control strategy for open frame underwater vehicles," *Robotica* **39**(10), 1791–1805 (2021).
[2] Z.-M. Chen, Q. Cui, X.-S. Wei, X. Jin and Y. Guo, "Disentangling, embedding and ranking label cues for multi-label image recognition," *IEEE Trans Multimedia* **23**(4), 1827–1840 (2021).
[3] L. Cai, Y. Li, C. Chen and H. Chai, "Dynamic multiscale feature fusion method for underwater target recognition," *J Sens* **2022**(13), 25–35 (2022).
[4] T.-S. Pan, H.-C. Huang, J.-C. Lee and C.-H. Chen, "Multi-scale ResNet for real-time underwater object detection," *Signal Image Video Process* **15**(5), 941–949 (2021).
[5] Q. D. Le, T. T. C. Vu and T. Q. Vo, "Application of 3D face recognition in the access control system," *Robotica* **40**(7), 2449–2467 (2022).
[6] J. Cheng, Y. Sun and M. Q.-H. Meng, "Robust semantic mapping in challenging environments," *Robotica* **38**(2), 256–270 (2020).
[7] B. Sun, B. Li, S. Cai, Y. Yuanand C. Zhang, "Fsce: Few-Shot Object Detection via Contrastive Proposal Encoding," **In:** *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA (IEEE, 2021) pp. 7352–7362.
[8] F. Qin, S. Qiu, S. Gao and J. Bai, "3D CAD model retrieval based on sketch and unsupervised variational autoencoder," *Adv Eng Inform* **51**, 101427 (2022).
[9] J. Hou, C. Luo, F. Qin, Y. Shao and X. Chen, "FuS-GCN: Efficient B-rep based graph convolutional networks for 3D-CAD model classification and retrieval," *Adv Eng Inform* **56**, 102008 (2023).
[10] Z. Chu, F. Wang, T. Lei and C. Luo, "Path planning based on deep reinforcement learning for autonomous underwater vehicles under ocean current disturbance," *IEEE Trans Intel Veh* **8**(1), 108–120 (2023).
[11] Y. Tang, L. Wang, S. Jin, J. Zhao, C. Huang and Y. Yu, "AUV-based side-scan sonar real-time method for underwater-target detection," *J Mar Sci Eng* **11**(4), 690 (2023).
[12] L. Cai, P. Luo, G. Zhou, T. Xu and Z. Chen, "Multiperspective light field reconstruction method via transfer reinforcement learning," *Comput Intel Neurosc* **2020**(2), 1–14 (2020).
[13] T. Lin, F. Pan and X. Wang, "Sparse point-plane odometry in structured environments," *Robotica* **40**(7), 2381–2394 (2022).
[14] T. Wang, C. Ma, H. Su and W. Wang, "SSFENet: Spatial and Semantic Feature Enhancement Network for Object Detection," **In:** *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada (IEEE, 2021) pp. 1500–1504.
[15] L. Cai, C. Chen and H. Chai, "Underwater distortion target recognition network (UDTRNet) via enhanced image features," *Comput Intell Neurosci* **2021**(3), 1–10 (2021).
[16] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network," *Remote Sens* **12**(9), 1432 (2020).
[17] W. Ma, C. Gong, S. Xu and X. Zhang, "Multi-scale spatial context-based semantic edge detection," *Inf Fusion* **64**(5), 238–251 (2020).
[18] M. Ju, J. Luo, Z. Wang and H. Luo, "Adaptive feature fusion with attention mechanism for multi-scale target detection," *Neural Comput Appl* **33**(7), 2769–2781 (2021).
[19] T. Lee and S. Yoo, "Augmenting few-shot learning with supervised contrastive learning," *IEEE Access* **9**(2), 61466–61474 (2021).
[20] H. Kuang, Y. Li, Y. Zhang, Y. Wan and G. Ge, "Research on rapid location method of mobile robot based on semantic grid map in large scene similar environment," *Robotica* **40**(11), 4011–4030 (2022).
[21] L. Cai, X. Qin and T. Xu, "EHDC: Enhanced dilated convolution framework for underwater blurred target recognition," *Robotica* **41**(3), 900–911 (2023).
[22] Z. Liu, J. Du, F. Tian and J. Wen, "MR-CNN: A multi-scale region-based convolutional neural network for small traffic sign recognition," *IEEE Access* **7**(1), 57120–57128 (2019).
[23] L. Douadi, Y. Dupuis and P. Vasseur, "Stable keypoints selection for 2D LiDAR based place recognition with map data reduction," *Robotica* **40**(11), 3786–3810 (2022).
[24] K. Wang, Y. Wang, S. Zhang, Y. Tian and D. Li, "SLMS-SSD: Improving the balance of semantic and spatial information in object detection," *Expert Syst Appl* **206**(6), 117682 (2022).

[25] C. Fang, H. Tian, D. Zhang, Q. Zhang, J. Han and J. Han, "Densely nested top-down flows for salient object detection," *Sci China Inf Sci* **65**(8), 182103 (2022).
[26] A. Gupta, A. Seal, P. Khanna, E. Herrera-Viedma and O. Krejcar, ALMNet: Adjacent layer driven multiscale features for salient object detection," *IEEE Trans Instrum Meas* **70**(3), 1–14 (2021).
[27] S. Wang, S. Chen, T. Chen and X. Shi, "Learning with privileged information for multi-label classification," *Pattern Recognit* **81**(5), 60–70 (2018).
[28] C. Zhi, "Mmnet: A Multi-Method Network for Multi-Label Classification," **In:** *2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA)*, Zhangjiajie, China (IEEE, 2020) pp. 441–445.
[29] B.-B. Gao and H.-Y. Zhou, "Learning to discover multi-class attentional regions for multi-label image recognition," *IEEE Trans Image Process* **30**(12), 5920–5932 (2021).
[30] W. Jiang, K. Huang, J. Geng and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Trans Circuits Syst Video Technol* **31**(3), 1091–1102 (2021).
[31] K. Oksuz, B. C. Cam, E. Akbas and S. Kalkan, "Rank & Sort Loss for Object Detection and Instance Segmentation," **In:** *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada (IEEE, 2021) pp. 2989–2998.
[32] Y. Zhang, L. Wang and Y. Dai, "PLOT: A 3D point cloud object detection network for autonomous driving," *Robotica* **41**(5), 1483–1499 (2023).
[33] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu and D. Krishnan, "Supervised contrastive learning," *Adv Neural Inf Process Syst* **33**(2), 18661–18673 (2020).

Cite this article: L. Cai, B. Zhang, Y. Li and H. Chai (2024). "IFE-net: improved feature enhancement network for weak feature target recognition in autonomous underwater vehicles", Robotica **42**, 1231–1245. https://doi.org/10.1017/S0263574724000195