frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 31 July 2024
DOI 10.3389/fnbot.2024.1423848

OPEN ACCESS

EDITED BY
Di Wu,
Southwest University, China

REVIEWED BY
Daxiong Ji,
Zhejiang University, China
Aiyang Han,
Nanjing University of Aeronautics and Astronautics, China

*CORRESPONDENCE
Xiujuan Du
✉ 124111397@qq.com

RECEIVED 26 April 2024
ACCEPTED 17 July 2024
PUBLISHED 31 July 2024

CITATION
Yan Q, Du X, Li C and Tian X (2024) CLIB: Contrastive learning of ignoring background for underwater fish image classification. *Front. Neurorobot.* 18:1423848. doi: 10.3389/fnbot.2024.1423848

CLIB: Contrastive learning of ignoring background for underwater fish image classification

Qiankun Yan[1,2], Xiujuan Du[1,2,3]*, Chong Li[1,2] and Xiaojing Tian[1,2]

[1]College of Computer, Qinghai Normal University, Xining, China, [2]Qinghai Provincial Key Laboratory of IoT, Xining, China, [3]The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining, China

Aiming at the problem that the existing methods are insufficient in dealing with the background noise anti-interference of underwater fish images, a contrastive learning method of ignoring background called CLIB for underwater fish image classification is proposed to improve the accuracy and robustness of underwater fish image classification. First, CLIB effectively separates the subject from the background in the image through the extraction module and applies it to contrastive learning by composing three complementary views with the original image. To further improve the adaptive ability of CLIB in complex underwater images, we propose a multi-view-based contrastive loss function, whose core idea is to enhance the similarity between the original image and the subject and maximize the difference between the subject and the background, making CLIB focus more on learning the core features of the subject during the training process, and effectively ignoring the interference of background noise. Experiments on the Fish4Knowledge, Fish-gres, WildFish-30, and QUTFish-89 public datasets show that our method performs well, with improvements of 1.43–6.75%, 8.16–8.95%, 13.1–14.82%, and 3.92–6.19%, respectively, compared with the baseline model, further validating the effectiveness of CLIB.

KEYWORDS

underwater fish image classification, contrastive learning, deep learning, self-supervised visual representation learning, background noise

1 Introduction

The ocean is one of the most important ecosystems on Earth and is an essential field for human survival and development. However, in recent years, the marine ecosystem has been continuously damaged (Georgian et al., 2022; Jiao et al., 2023). To protect the oceans, we need to understand the health of the oceans, and information such as the distribution of different species of fish and the number of fish in a particular watershed can well reflect the health of the ecological environment in that watershed (Trindade-Santos et al., 2022; Xuan et al., 2022; Yu et al., 2022). Therefore, the study of the species and number of fish through the collected images of underwater fishes is of great significance for further understanding the health of the oceans and protecting endangered species (Ovalle et al., 2022; Zhang D. et al., 2022).

However, underwater optical images are significantly different from land optical images. The lights with different wavelengths have different propagation characteristics in water, resulting in the collected underwater images being characterized by color distortion, visual

blurring, and low contrast (Chen et al., 2022; Wang et al., 2022; Lu et al., 2023), which is shown in Figure 1. Figure 1A shows the fish images taken in the terrestrial environment, and it can be seen that the texture of the images is clear and visible, Figure 1B shows the fish images taken in water environment, which are characterized by color distortion, visual blurring, etc. In addition, due to the high mobility of fish, the same species of fish may appear in different backgrounds while different species of fish may appear in the same background, which leads to the background not only having no positive effect but also interfering with the training of the model. The background noise brings about great challenges to the recognition of underwater fish images.

Traditional machine learning methods (Larsen et al., 2009) for underwater fish image classification usually use manually designed features to extract features, and it could be more scalable and generalizable in the face of large-scale datasets. The later emergence of supervised visual representation learning (Li et al., 2021) solved the drawbacks of manually designed features in traditional machine learning methods and attracted much response. However, supervised visual representation learning relies on manually labelled labels when training the model. When facing large-scale underwater fish image datasets, the labelling work on the labels consumes a lot of time and energy for oceanography experts. The emergence of self-supervised visual representation learning (Ericsson et al., 2022) in recent years has improved this problem by requiring only a small number of labels to fine-tune the model to achieve impressive results, significantly reducing the tediousness of labelling data. However, the current self-supervised visual representation learning methods are primarily designed for general-purpose models, which could not work well when facing underwater images of noisy noise. To address the aforementioned issues, this paper proposes contrastive learning of ignoring background for underwater fish image classification called CLIB, which is proposed to reduce the negative impact of background noise. The main contributions of this paper are as follows:

1. This paper reconstructs the view of contrastive learning based on the characteristics of underwater fish images. The subject and background of the image are extracted through an extraction module, and the original image, extracted subject, and background constitute three views for contrastive learning.
2. This paper proposes a multi-view-based contrastive loss function and defines a sample in the subject view as a positive sample of the corresponding image in the original view, while all other samples in the three views are negative samples of the original image.
3. This paper conducts a large number of comparative experiments from three perspectives: different resolutions, complex backgrounds, and few-sample to verify the superiority of the proposed CLIB method in underwater fish image classification.

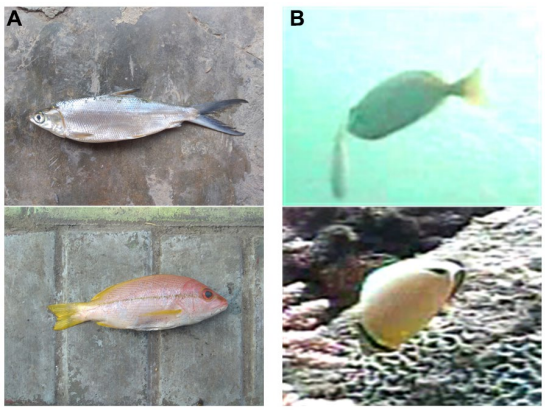

FIGURE 1
Images taken on land vs. images taken underwater. **(A)** Images taken on land. **(B)** Images taken underwater.

The rest of the paper is organized as follows. Section 2 reviews existing recognition methods of underwater fish and visual representation methods of self-supervising. Section 3 describes our proposed CLIB method. Section 4 presents the experimental results by comparing the CLIB with the mainstream methods. Section 5 concludes the paper.

2 Related work

2.1 Traditional machine learning methods

Scholars' research on fish image classification can be traced back to 1990 (Xu et al., 2020), and most of the early research combined traditional machine learning models with image processing techniques, which mainly focused on the design of feature extraction and improving the accuracy of classification by extracting more favorable information such as shape and texture. For example, Spampinato et al. achieved fish image classification by combining texture features and shape features (Spampinato et al., 2010). Texture features were extracted according to the statistical moments of the grayscale histogram, spatial Gabor filtering, and the properties of the co-occurrence matrix. Shape features were extracted by using curvature scale-spatial transformation and the histogram of boundary Fourier descriptors. Huang et al. achieved fish image classification by extracting 66 features from different parts of a fish composed of color, shape, and texture and reduced the feature dimensions by a forward sequential feature selection procedure (Huang et al., 2012). Fouad et al. described the local features extracted from a set of fish images to differentiate fish species through the algorithm supporting vector machine combined with an accelerated robust feature algorithm based on scale-invariant feature transformation (Fouad et al., 2013). Hu et al. extracted six sets of feature vectors, including the color features of the image, the color features of texture sub-images, the features of statistical texture, and the features of texture based on wavelets, and the feature vectors were fed into the supporting Vector Machine for classification (Hu et al., 2012). Khotimah et al. extracted eight texture and shape features from fish images using image processing methods and then used these features to create a classification model using a decision tree (Khotimah et al., 2015). Most early research on the classification of underwater fish image was based on traditional machine-learning models and image-processing techniques (Zhang et al., 2021; Zhang Z. et al., 2022). The main steps include (1) denoising and enhancing the underwater fish images using image

processing techniques, (2) extracting the pre-designed features artificially from the underwater fish images, and (3) training the traditional machine learning models based on the extracted features to classify the fish.

However, the classification effectiveness of traditional machine learning models is closely related to feature extraction methods designed manually, and most researchers rely on experience to design the features, which has the disadvantage of a certain degree of subjectivity and blindness. Although these methods achieve some classification results, most of the features designed by the researchers are designed for only a specific dataset. When faced with a new dataset or applied in practice, the classification results of the model usually have significant errors compared to the reality.

2.2 Supervised visual representation learning

With the development of deep learning, deep learning-based methods have achieved good results in various fields in recent years. Different from traditional machine learning methods that rely on hand-designed feature extraction, deep learning methods are capable of automatic feature learning, which dramatically reduces the tediousness of design while improving performance. For example, Sun et al. solved the problem of limited discriminative information in low-resolution images by using deep learning and super-resolution methods to explicitly learn discriminative features in relatively low-resolution images (Sun et al., 2016). Deep et al. used convolutional neural networks to extract features and then used support vector machines and K-Nearest Neighbor to classify images (Deep and Dash, 2019). Based on the idea of contrastive learning, Zhang et al. encouraged the model to learn more discriminative features for different categories of images and similar features for images of the same category (Zhang et al., 2021). In addition, a regularization technique known as attentional suppression was used to prevent the model from paying much attention to the background. To reduce the effect of extreme noise in underwater images, Zhang et al. trained the model using adversarial perturbation images with the perturbation method, which helps to train a better recognition model from images containing extreme noise (Zhang D. et al., 2022). Li et al. used a method of multi-color space coding to fully integrate the feature advantages of different color spaces and then obtained the global and local deep features of the images in multiple dimensions through the multi-channel attention path aggregation strategy, and finally form a multi-channel attention network architecture through the embedding and stacking of multi-channel attention modules, which strengthens the perception of image features (Li et al., 2023).

Although supervised visual representation learning can extract data features, labels pre-labeled manually are still required in feature learning. Labeling data requires a lot of effort and time from the experts, and there may be some labeling errors, which can bring about great misleading in subsequent model learning. Self-supervised visual representation learning (Chen et al., 2020; Sang et al., 2022) solves this problem. With self-supervised representation learning, it is possible to learn the model without knowing the label information of the image. When ported to the downstream task, only a small amount of label information is needed to fine-tune the model to approximate or even exceed the effect of supervised learning.

2.3 Self-supervised visual representation learning

Self-supervised visual representation learning can provide powerful deep feature learning without the need for large amounts of labeled data and alleviate the annotation bottleneck to some extent (Ericsson et al., 2022). The most classical self-supervised visual representation learning is contrastive learning, which learns data representation by maximizing the similarity between positively correlated samples and minimizing the similarity between uncorrelated samples. With contrastive learning, a label is first derived from the unlabeled data by a pre-defined strategy, and then the model is trained using this label and the data. The key in contrastive learning is how to design the strategy for deriving a label, which is called a pretext task by scholars. The effectiveness of the pretext task determines the effectiveness of the model for downstream tasks. Consequently, the choice of the pretext task is vital for self-supervised visual representation learning. For example, He et al. thought that increasing the number of negative samples can increase the difficulty of comparison learning and enable the model to learn more detailed feature information (He et al., 2020). Therefore, they proposed the Momentum Contrast (MoCo) learning method, which achieved good results by adding a MEMORY BANK and updating the encoder parameters using momentum. Chen et al. explored the optimal combined method of data augmentation by eliminating memory banks and encouraging larger Batch sizes and longer training times (Chen et al., 2020). Chen et al. presented a self-supervised learning method without negative samples as well as without increasing the batch size (Chen and He, 2021). In the self-supervised learning method, the feature vectors obtained from one of the two-branch networks after passing through the encoder and the feature vectors from another one of the two-branch networks passing through the encoder and the multilayer perceptron are mutual positive samples. The network is trained by maximizing the similarity of the positive sample pairs, which achieved good results.

In addition to the above papers, some excellent methods of self-supervised visual representation learning are available. However, due to the particular characteristics of underwater images, applying general methods to underwater fish image classification does not achieve ideal results. In this paper, from the idea of focusing on the subject and ignoring the background, we innovatively design a contrastive learning of ignoring the background for underwater fish image classification, which is more suitable for underwater fish image classification.

3 The CLIB method

The training overview diagram of CLIB is shown in Figure 2. The subjects and backgrounds of the input images are first extracted by the extraction module. Then, the original images and the extracted subjects and backgrounds constitute three views, respectively, which are fed into the respective encoders after random data augmentation and then fed into the feature space through the projection head to compute the multi-view-based contrastive loss.

3.1 Symbol definition

This paper defines the main symbols as shown in Table 1.

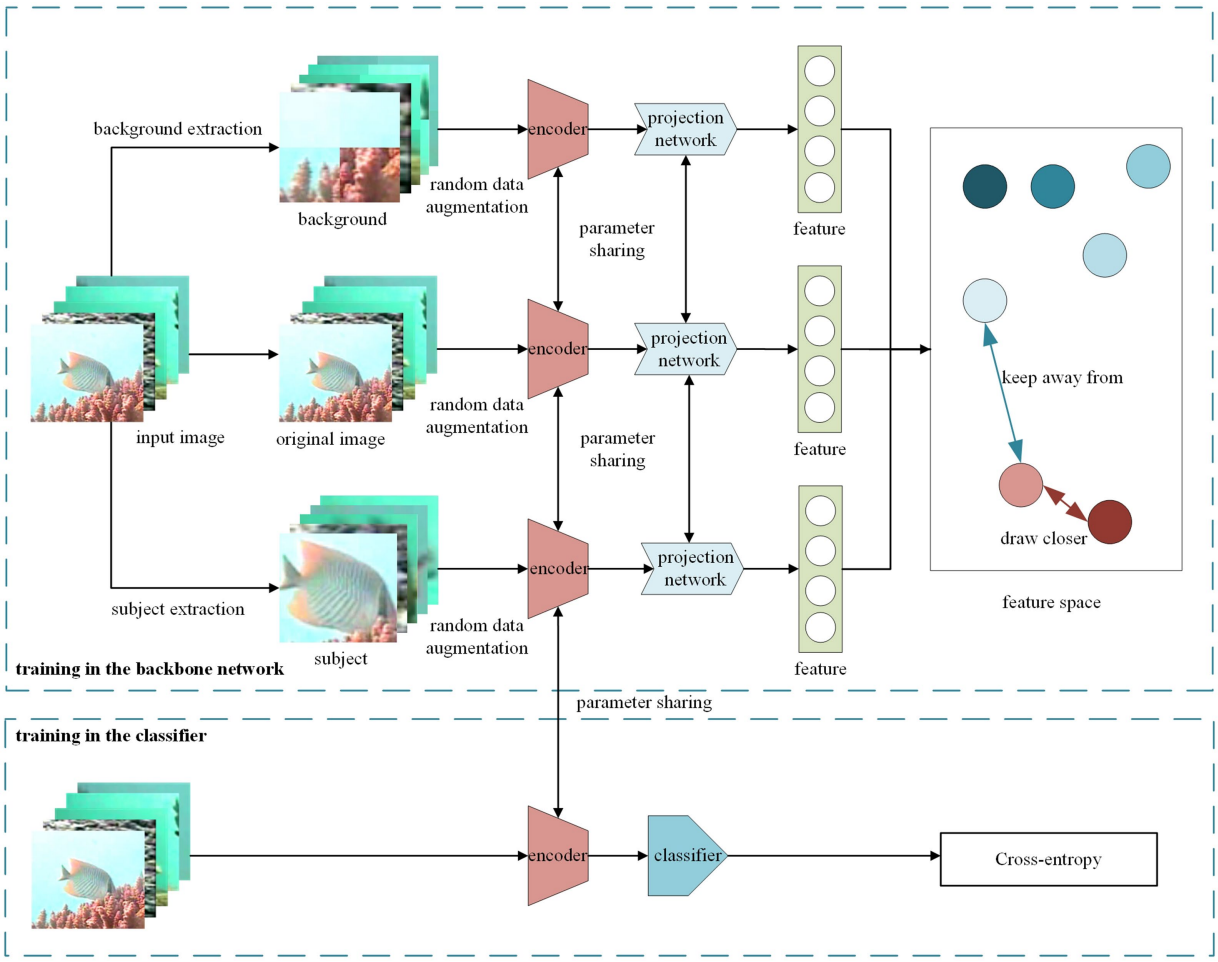

FIGURE 2
Overview of CLIB training. The upper half of the figure shows the self-supervised training of the backbone network with the CLIB method, and the lower half of the figure shows the classifier training in fine-tuning stage in which the parameters of the trained backbone network are freezed and a classifier is added for supervised training the whole network with a small amount of data.

TABLE 1 Main symbols.

Symbol	Meaning
$x \in X \in R^{(W*H*C)}$	Underwater fish image
W, H	Width and height of image
$t \in T$	Data augmentation
$f(\cdot)$	Backbone
$g(\cdot)$	Projection network
E	Subject and background extraction module

3.2 Subject and background extraction module

Before the original image is input into the model, the image is processed to the specified size, and then the subject and background of the image are extracted according to the extraction rules.

The rule for extracting the subject is extracting the central region with the area size of the subject image is shown in,

$$S_{subject} = Area(\alpha, \beta) = (\alpha * H) * (\beta * W)$$

where $Area(\cdot,\cdot)$ denotes the solution function to the area, and α and β are the hyperparameter ratios of length and width, respectively, set $\alpha = \beta$.

The rule for extracting the background is extracting the four corners of the image. Afterwards, the four corners are pieced together to form a background image. The area size of the background image is shown in,

$$S_{background} = Area(\gamma, \delta) = (\gamma * H) * (\delta * W)$$

where γ and δ are the hyperparameter ratios of length and width, respectively, set $\gamma = \delta$.

After the sample x_i is input into the extraction module E, two samples are obtained at the output end (the subject sample and the background sample of the original image), which is given by,

$$Sub(x_i) = f_{crop}(x_i, S_{subject})$$

$$Bac(x_i) = f_{crop}(x_i, S_{background})$$

$$x_{i(subject)}, x_{i(background)} = Sub(x_i), Bac(x_i)$$

where f_{crop} is a crop function. $Sub(x_i)$ and $Bac(x_i)$ are the functions to exract the subject and background, respectively.

The three samples are further subjected to data augmentation, feature extraction, and projection mapping, and finally, three sets of feature vectors are acquired, which is given by,

$$z_i, z_{i(subject)}, z_{i(background)} = g\left(f\left(t\left(x_i, x_{i(subject)}, x_{i(background)}\right)\right)\right)$$

3.3 Build multi-views

The view of SimCLR (Chen et al., 2020) is constructed as follows. Firstly, randomly sample N sample images. After two random data augmentations, we obtain $2N$ expanded samples and two views. The two expanded samples originating from the same image are defined as mutual positive samples, and the remaining $2(N-1)$ samples are defined as the negative samples of the two positive samples.

The views of CLIB are constructed as follows. Firstly, randomly select N sample images (N original images) and input the selected N original images into the extraction module to obtain $2N$ samples, i.e., N subject samples and N background samples. After randomly augmenting the $3N$ samples, we obtain $3N$ expanded samples. The expanded sample of an original image and the expanded sample of the subject sample extracted from the original image are mutual positive samples, and the rest of the $3N-2$ expanded samples are defined as negative samples of the expanded sample of the original image or the subject sample. The acquired positive and negative samples are shown in Figure 3.

3.3.1 Multi-view-based contrastive loss function

SimCLR (Chen et al., 2020) follows the idea of contrastive learning. After constructing two views, each sample in the two views corresponds to one similar sample (positive sample) and $2(N-1)$ dissimilar samples (negative samples). $2N$ feature vectors are obtained after $2N$ samples are encoded through an encoder and projected through a projection network, and the similarity between two feature vectors is calculated according to the cosine similarity formula, which is given by,

$$sim(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\| \|z_j\|}$$

where z_i denotes the feature vector of sample i, z_j denotes the feature vector of sample j, T represents the transpose of the vector, and $\|z_i\|$ indicates the length of the vector z_i. For the positive samples pair (i,j), the definition of the loss function for SimCLR (Chen et al., 2020) is given by,

$$l_{i,j(SimCLR)} = -\log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(sim(z_i, z_k)/\tau)}$$

where $1_{[k \neq i]} \in \{0,1\}$ is an indicator function with value "1" when $k \uparrow i$ and value "0" when $k = i$, τ is a temperature parameter, and $\exp(\bullet)$ is the exponential function. The total loss is given by,

$$L_{SimCLR} = \frac{1}{2N} \sum_{k=1}^{N} \left[l_{(2k-1,2k)} + l_{(2k,2k-1)} \right]$$

where $(2k-1,2k)$ and $(2k,2k-1)$ represent the pairs of positive samples.

CLIB also follows the idea of contrastive learning, but unlike SimCLR (Chen et al., 2020), the number of negative samples in CLIB

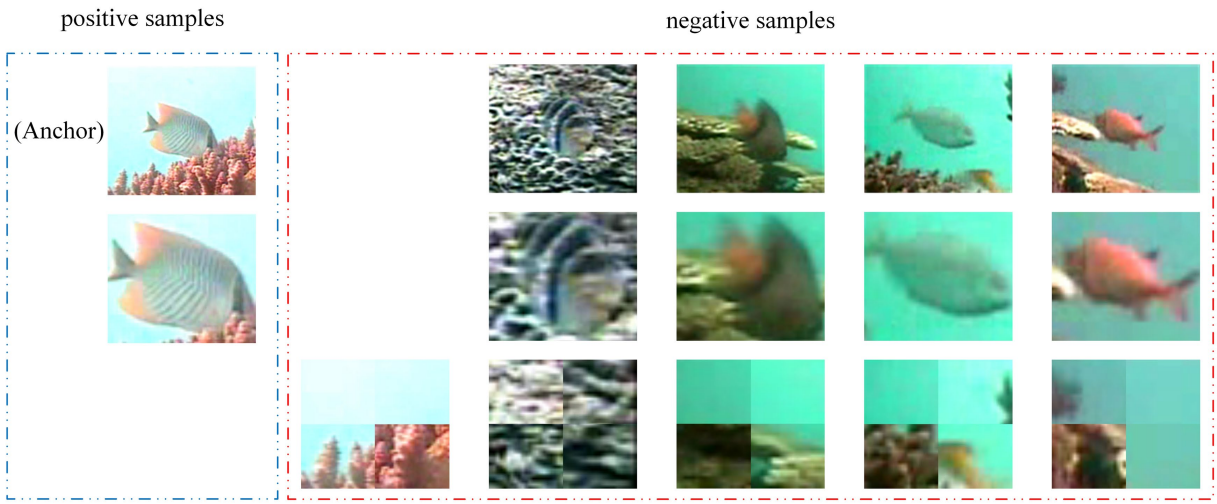

FIGURE 3
The diagram of positive and negative samples of CLIB. The first row of images are the original views, the second row of images are their respective subject views, and the third row of images are their respective background views. The upper left original image is regarded as an anchor, its subject view (the below image in blue line) is regarded as the positive sample of the anchor, and all other images are regarded as negative samples of the anchor (all images in red line).

increases significantly. This is because, besides the $2(N-1)$ negative samples in both the original image view and the subject view, the N samples in the background view are also defined as negative samples. After the three views are constructed, each sample in either the subject or the original image view has one similar sample (positive sample) and $3N-2$ dissimilar samples (negative samples). In CLIB, $3N$ feature vectors are obtained after $3N$ samples are encoded through an encoder and projected through a projection network. For the positive sample pair (i,j) and all the corresponding negative samples, the loss function of CLIB is given by,

$$l_{i,j(CLIB)} = -\log \frac{\exp\left(sim\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{3N} 1_{[k\leq 2N, k\neq i:1; k>2N:1]} \exp\left(sim\left(z_i, z_k\right)/\tau\right)}$$

where $1_{[k\leq 2N, k\neq i:1; k>2N:1]} \in \{0,1\}$ is the indicator function with value "1" when k " $2N$ and $k \uparrow i$ or $k > 2N$, in all other cases, the value of the function is 0. k " $2N$ means that k belongs to the original images view or the subject view. If $k > 2N$, the value of the indicator function is "1." τ is the temperature parameter, and $\exp(\bullet)$ is the exponential function. The total loss is given by,

$$L_{CLIB} = \frac{1}{2N} \sum_{k=1}^{N} \left[l_{(3k-1,3k-2)} + l_{(3k-2,3k-1)} \right]$$

where $(3k-1,3k-2)$ and $(3k-2,3k-1)$ represent the pairs of positive samples. It should be noted that the positive samples only appear in either the original view or the subject view, and all the samples in the background view are negative samples.

4 Experiments

In this section, the performance of CLIB is evaluated and compared with the benchmark model (SimCLR) as well as nine mainstream self-supervised visual representation learning methods through experiments in which the encoder is ResNet50 (He et al., 2016).

4.1 Datasets and experimental set-up

4.1.1 Datasets

The experiments are performed on four datasets. The types and quantities of the four datasets are shown in Table 2. The fish images in Fish4Knowledge (Boom B. et al., 2012; Boom B. J. et al., 2012), WildFish-30, and QUTFish-89 datasets are taken in water, while the fish images in the Fish-gres dataset (Prasetyo et al., 2020) are taken on land. The WildFish-30 (Zhuang et al., 2018) dataset is composed of the images in the 30 categories with the highest number of images. The QUTFish-89 dataset is composed of the images in the 89 few-sample categories from the QUTFish dataset (Anantharajah et al., 2014).

4.1.2 Experimental set-up

The experiments in this paper are conducted under the same hardware and software environment. Specifically, the CPU used is Intel(R) Xeon(R) Platinum 8358P, while the GPU used is A40 (48GB). The size of the memory is 80GB. The Python version is 3.8, while the Pytorch version is 2.0. When training in the backbone network, all samples are put into the network for training, and the model with the lowest loss value is preserved. When training in the classifier, the dataset is divided into three subsets with a ratio of 1:1:8. One subset with 10% samples is used as the training set, another subset with 10% samples is used as the validation set, and the remaining subset with 80% samples is used as the test set. The model with the highest accuracy in the classifier training process on the validation set is reserved, and the final accuracy is obtained by testing the test set with the reserved model. The rest of the experimental setup is shown in Table 3.

TABLE 2 Type information of four datasets.

Dataset	Number of species	Number of images	Resized resolution
Fish4Knowledge	23	27,370	64 × 64
Fish-gres	8	3,248	224 × 224
WildFish-30	30	3,688	224 × 224
QUTFish-89	89	823	224 × 224

TABLE 3 Experimental set-up.

Environment and parameters	Set-up
Optimizer	SGD
Initial learning rate	0.01
Final learning rate	0.0001
Temperature	0.07
Training epochs of the backbone network	500
Training epochs of the classifier	100

4.2 Results of comparative experiments

To verify the effectiveness and superiority of the CLIB method, nine popular self-supervised methods of visual representation learning are selected for comparison, including SimCLR (a simple framework for contrastive learning of visual representations; Chen et al., 2020), MOCO (momentum contrast for unsupervised visual representation learning; He et al., 2020), SimSiam (exploring simple siamese representation learning; Chen and He, 2021), BYOL (bootstrap your own latent; Grill et al., 2020), TiCo (transformation invariance and covariance contrast for self-supervised visual representation learning; Zhu et al., 2022), NNCLR (nearest-neighbor contrastive learning of visual representations; Dwibedi et al., 2021), Dcl (decoupled contrastive learning; Yeh et al., 2022), Matrix-SSL (Matrix Information Theory for Self-Supervised Learning; Zhang et al., 2023), and Mixed Barlow Twins (Guarding Barlow Twins Against Overfitting with Mixed Samples; Bandara et al., 2023). To test the proposed CLIB method more objectively, to define four metrics, Accuracy, Precision, Recall, and F1 value, as the metrics to evaluate the classification performance of the methods.

4.2.1 Comparative experiments on underwater images with different resolutions

To further explore the actual effect of CLIB, we conduct experiments on Fish4Knowledge and WildFish-30 datasets with very different resolutions, and the results are shown in Table 4. Meanwhile, the validation accuracy is shown in Figures 4, 5.

The experimental results show that the CLIB method performs excellently on the lower-resolution Fish4Knowledge dataset and the higher-resolution WildFish-30 dataset. It is worth noting that the effect of CLIB on the higher-resolution WildFish-30 dataset is more prominent. This is because the original image resolution is high, CLIB extracts more pixel points of the subject and background image parts, which provides the model with richer information about the negative samples compared to other methods and makes the model pay attention to the subject part of the image and ignore the background part. In conclusion, experimental results and analysis on the Fish4Knowledge and WildFish-30 datasets show that CLIB performs excellently in underwater fish image classification of different resolutions with background noise.

4.2.2 Comparative experiments in complex backgrounds

To reflect the main idea that CLIB focuses more attention on the subject of the image and ignores the background, we purposely conduct experiments on the dataset of Fish-gres, in which the fish images are taken from land, and the backgrounds of images of fishes

TABLE 4 Experimental results on the Fish4Knowledge and WildFish-30 datasets.

Method	Fish4Knowledge				WildFish-30			
	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Acc (%)	Pre (%)	Rec (%)	F1 (%)
SimCLR (Chen et al., 2020)	94.58	86.83	69.19	74.10	46.56	49.88	46.17	46.82
Moco (He et al., 2020)	93.12	78.78	66.02	69.06	44.55	46.50	44.48	44.33
SimSiam (Chen and He, 2021)	88.28	41.32	34.57	35.33	21.99	24.35	21.94	21.10
BYOL (Grill et al., 2020)	91.01	60.97	46.82	48.54	39.55	41.65	39.32	39.52
TiCo (Zhu et al., 2022)	86.36	58.63	48.30	49.20	32.48	34.42	32.21	31.16
NNCLR (Dwibedi et al., 2021)	94.07	69.17	66.47	65.93	41.60	44.75	41.45	41.53
Dcl (Yeh et al., 2022)	94.83	85.43	71.19	74.60	45.02	46.43	44.78	44.60
Matrix-SSL (Zhang et al., 2023)	95.93	70.80	56.56	59.70	40.96	42.41	40.78	41.02
Mixed Barlow Twins (Bandara et al., 2023)	93.21	53.25	47.28	47.86	33.48	36.24	33.18	32.51
CLIB	96.01	91.48	75.94	80.16	60.94	62.98	60.99	61.44

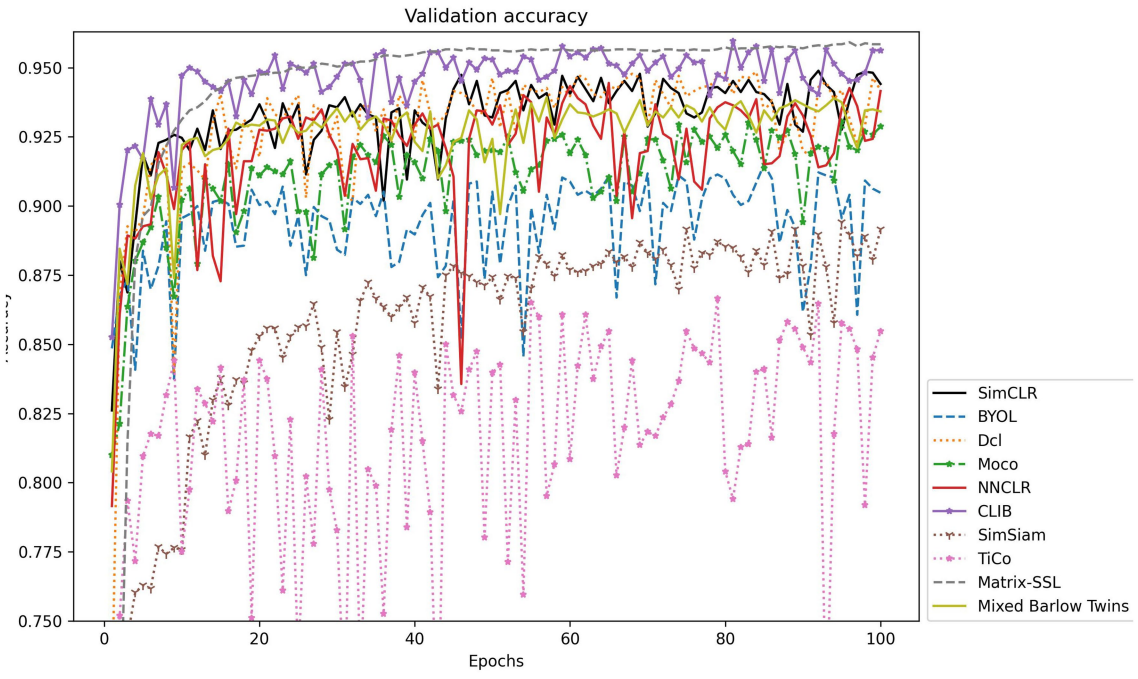

FIGURE 4
Accuracy validation on the Fish4Knowledge dataset.

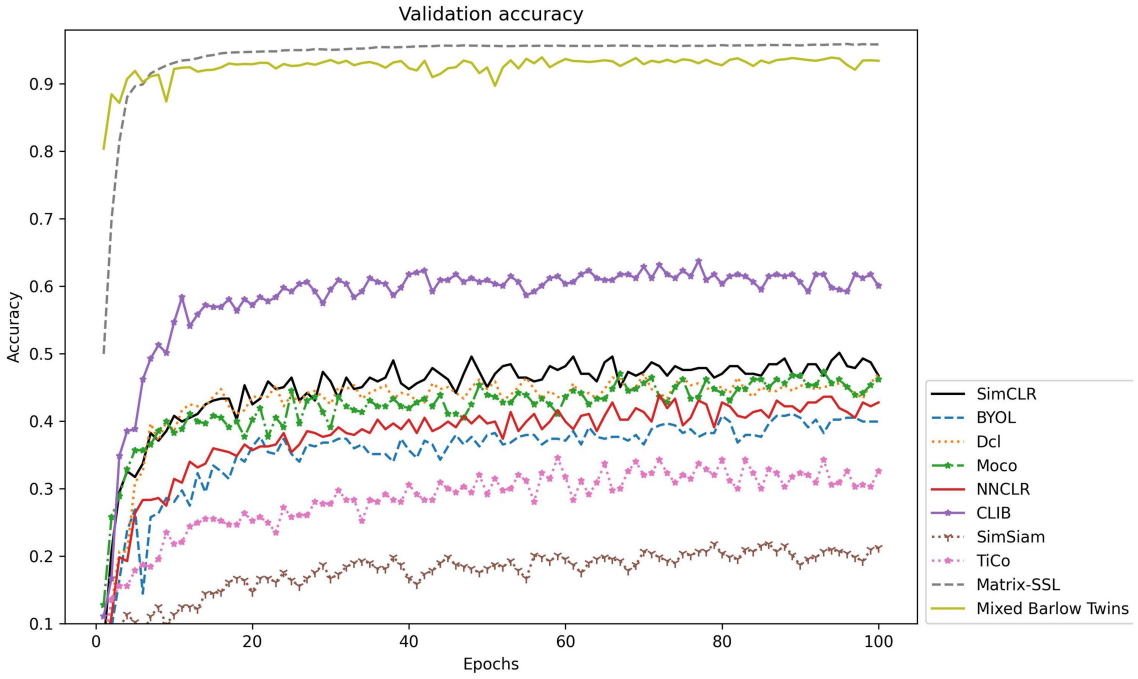

FIGURE 5
Accuracy validation on the WildFish-30 dataset.

belonging to the same species are different greatly. The experimental results on the Fish-gres datasets are shown in Table 5, and the validation accuracy is shown in Figure 6.

For datasets like Fish-gres with large background differences between similar classes, it should be more difficult for the ordinary contrastive learning methods to achieve feature learning by zooming in the feature mapping of positive sample pairs and zooming out the feature mapping of negative samples. In this paper, we propose the CLIB method. The main idea of CLIB is to pay more attention to the subject of the image and ignore the background of the image. Theoretically, CLIB should perform much better on the Fish-gres dataset. The experimental results show that CLIB achieves accuracy of 87.59%, precision of 88.38%, recall of 86.35%, and F1 value of 87.17% on the dataset of Fish-gres, which outperforms SimCLR (8.72%), Moco (8.61%), SimSiam (20.94%), BYOL (19.56%), TiCo (19.63%), NNCLR (7.42%), Dcl (9.84%), Matrix-SSL (13.49%), and Mixed Barlow Twins (28.43%) in terms of accuracy. Furthermore, CLIB outperforms the nine methods in terms of precision, recall, and F1 value, which verifies the validity of the CLIB's idea of ignoring the background and focusing on the subject.

TABLE 5 Experimental results on the Fish-gres dataset.

Method	Fish-gres			
	Acc (%)	Pre (%)	Rec (%)	F1 (%)
SimCLR (Chen et al., 2020)	78.87	80.22	77.45	78.22
Moco (He et al., 2020)	78.98	80.35	77.88	78.84
SimSiam (Chen and He, 2021)	66.65	66.61	64.73	65.38
BYOL (Grill et al., 2020)	68.03	70.14	64.65	65.98
TiCo (Zhu et al., 2022)	67.96	69.01	66.89	66.44
NNCLR (Dwibedi et al., 2021)	80.17	81.43	80.10	80.32
Dcl (Yeh et al., 2022)	77.75	80.82	76.28	77.96
Matrix-SSL (Zhang et al., 2023)	74.10	74.11	71.04	71.94
Mixed Barlow Twins (Bandara et al., 2023)	59.16	62.16	55.30	55.90
CLIB	87.59	88.38	86.35	87.17

4.2.3 Comparative experiments with few-sample datasets

When taking underwater fish images, it is often difficult to capture enough fish images due to the sparse number of fish in some species, resulting in some categories of the dataset presenting a low sample size. To better adapt to this situation, we further evaluate CLIB's ability to learn with few samples and its generalization by constructing a few-sample dataset. Eighty-nine categories with few samples are extracted from the QUTFish dataset to form the QUTFish-89 dataset, most of which have less than 10 images in the category. Not only that, only 34 images were used to fine-tune the model for the classifier, and far fewer than the categorized categories 89. The experimental results on the QUTFish-89 dataset are shown in Table 6, and the validation accuracy is shown in Figure 7.

Under such demanding conditions, the CLIB method still obtains good results compared to the other nine methods, with significant advantages in all four metrics. This is because in the training phase of the backbone network, benefiting from the idea of focusing on the subject of the image and ignoring the background of the image, CLIB is less affected by the background in the process of learning, and the model can more accurately capture the differences between different categories of images and the sameness between the same categories.

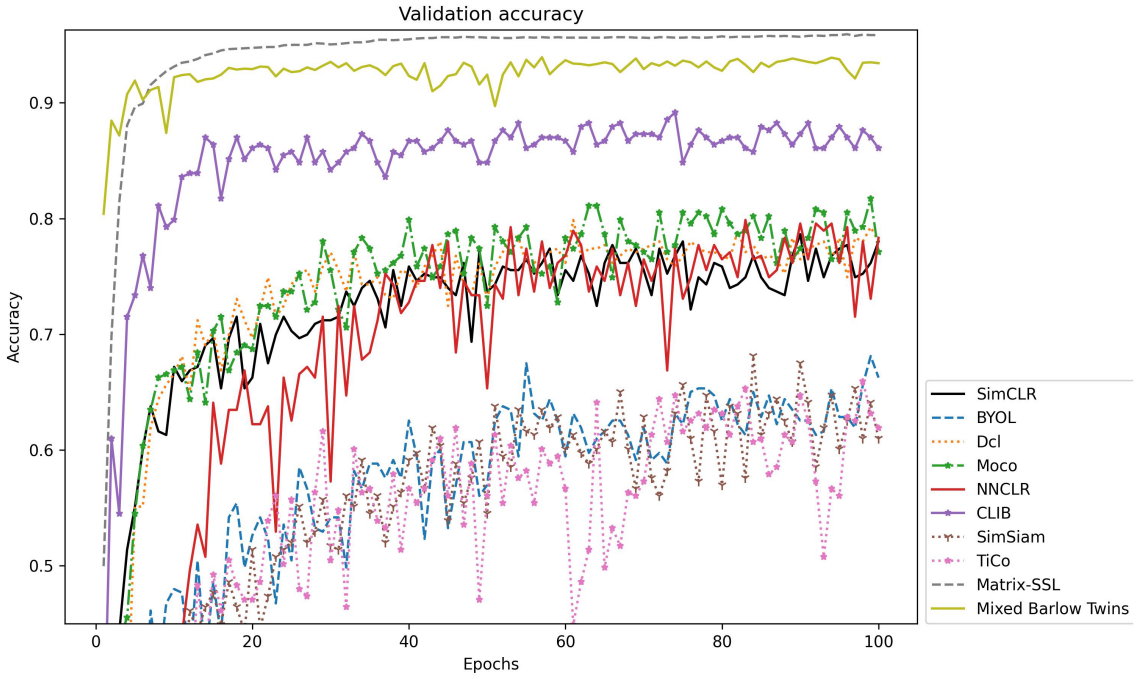

FIGURE 6
Accuracy validation on the Fish-gres dataset.

TABLE 6 Experimental results on the QUTFish-89 dataset.

Method	QUTFish-89			
	Acc (%)	Pre (%)	Rec (%)	F1 (%)
SimCLR (Chen et al., 2020)	10.71	6.98	9.16	6.39
Moco (He et al., 2020)	4.36	3.60	3.90	2.68
SimSiam (Chen and He, 2021)	5.29	4.21	4.62	3.33
BYOL (Grill et al., 2020)	6.87	4.79	5.74	4.04
TiCo (Zhu et al., 2022)	6.21	4.64	5.49	3.54
NNCLR (Dwibedi et al., 2021)	7.53	6.12	6.73	4.63
Dcl (Yeh et al., 2022)	8.99	5.94	7.88	5.22
Matrix-SSL (Zhang et al., 2023)	11.11	6.69	9.65	6.53
Mixed Barlow Twins (Bandara et al., 2023)	6.21	6.88	5.60	3.51
CLIB	16.13	13.17	13.89	10.31

As a result, CLIB can also distinguish different categories of fish images well when faced with this sparse number of samples. At the same time, the other nine methods make it more difficult for the model to learn the homogeneity between the same categories when facing image data with sparse samples because of the scarcity of data in the same category. In addition, the backgrounds of fish images between different categories are highly similar, while fish images of the same category can have large differences, and the sparse data make it more difficult for these methods to learn the differences between images of different categories and the homogeneity between the same categories.

4.3 Results of ablation experiments

To further verify the superiority of CLIB, three groups of ablation experiments are designed:

4.3.1 Baseline

The experiments of the first group are conducted with the baseline model SimCLR, which consists of two views of the two original images derived from the same image with two different data enhancements.

4.3.2 CLIB-ablation

The experiments of the second group, the background view is added in the SimCLR model as expanded negative samples of two enhanced image views, and forms the third view.

4.3.3 CLIB

The experiments of the third group are conducted with the proposed CLIB method, with three views in total: original view, body view, and background view.

The results of the ablation experiments are shown in Tables 7, 8. The experiment results of the second group are either better or worse than those of the first group. This is because in the second set of methods, on the one hand, the mappings in the feature space between one of the enhanced original views and the background view is zoomed out in the process of training, which results in the model ignoring the background. On the other hand, the mappings in the feature space between one enhanced original image and another enhanced original image is zoomed in during the training process. However, due to the two images in the two original views being with background, the mapping between the two backgrounds in the two

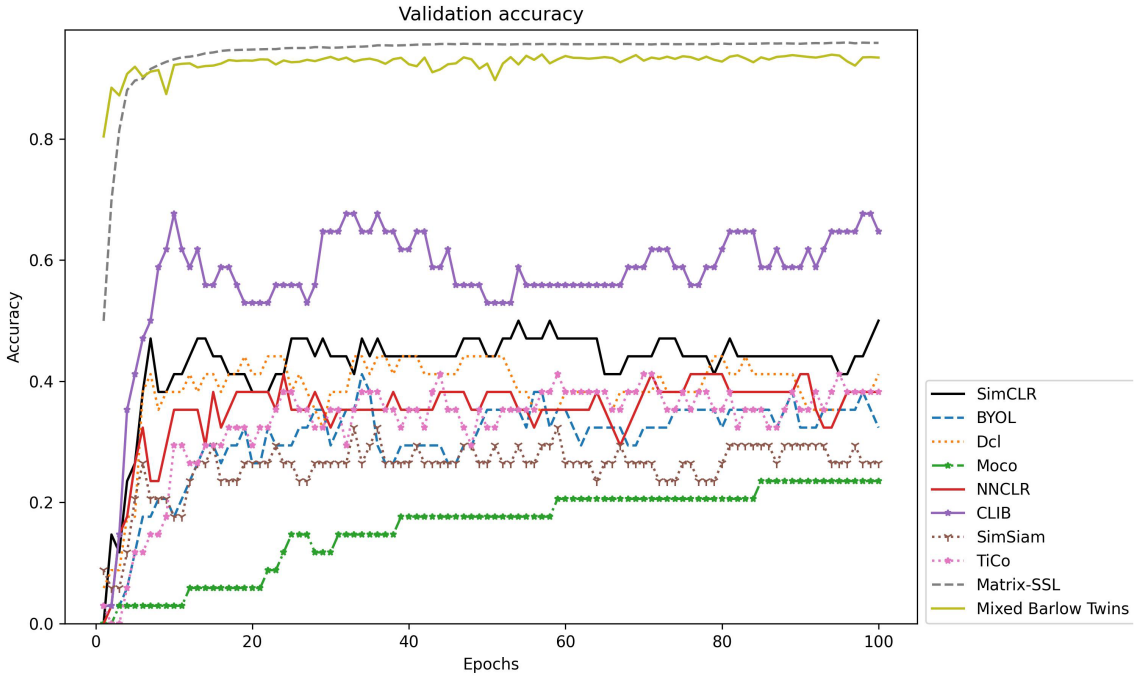

FIGURE 7
Accuracy validation on the QUTFish-89 dataset.

TABLE 7 Experimental results on Fish4Knowledge and Fish-gres datasets.

Method	Fish4Knowledge				Fish-gres			
	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Baseline	94.58	86.83	69.19	74.10	78.87	80.22	77.45	78.22
CLIB-ablation	94.73	81.22	70.57	73.72	79.40	80.03	78.68	79.07
CLIB	96.01	91.48	75.94	80.16	87.59	88.38	86.35	87.17

TABLE 8 Experimental results on WildFish-30 and QUTFish-89 datasets.

Method	WildFish-30				QUTFish-89			
	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Baseline	46.56	49.88	46.17	46.82	10.71	6.98	9.16	6.39
CLIB-ablation	48.44	49.50	48.08	48.09	12.43	8.93	10.70	7.56
CLIB	60.94	62.98	60.99	61.44	16.13	13.17	13.89	10.31

original views is also zoomed in, which results in the model failing to ignore the background. Therefore, it can be concluded that simply adding background views to SimCLR to expand the negative samples does not improve the performance of the model.

The ablation experiments of the third group are conducted using the CLIB method proposed in this paper. Consequently, compared to the experiments in the first or second group, the performance of the CLIB method is significantly improved, which is verified by the experimental results in Tables 7, 8. This is because the CLIB method does not suffer from the contradiction in the second set of experiments. Thus, the ablation experiments verify the validity of the idea of the CLIB by focusing on the subject and ignoring the background.

4.4 Visualization

To visualize the idea of CLIB of focusing on the subject and ignoring the background, experiments on visualizing network attention using class activation maps (Grad-Cam; Selvaraju et al., 2017) are conducted. The results are shown in Figure 8. The redder the area, the more critical it is for decision or classification, and the SimCLR model is used as the baseline in the visualization experiment. Most general methods tend to regard the background in the image as part of the fish for classification. Figure 8 shows the subject area is concerned with both the benchmark model and CLIB. However, with the benchmark model, the recognition results are easily affected by backgrounds. For example, in the first row, although the benchmark

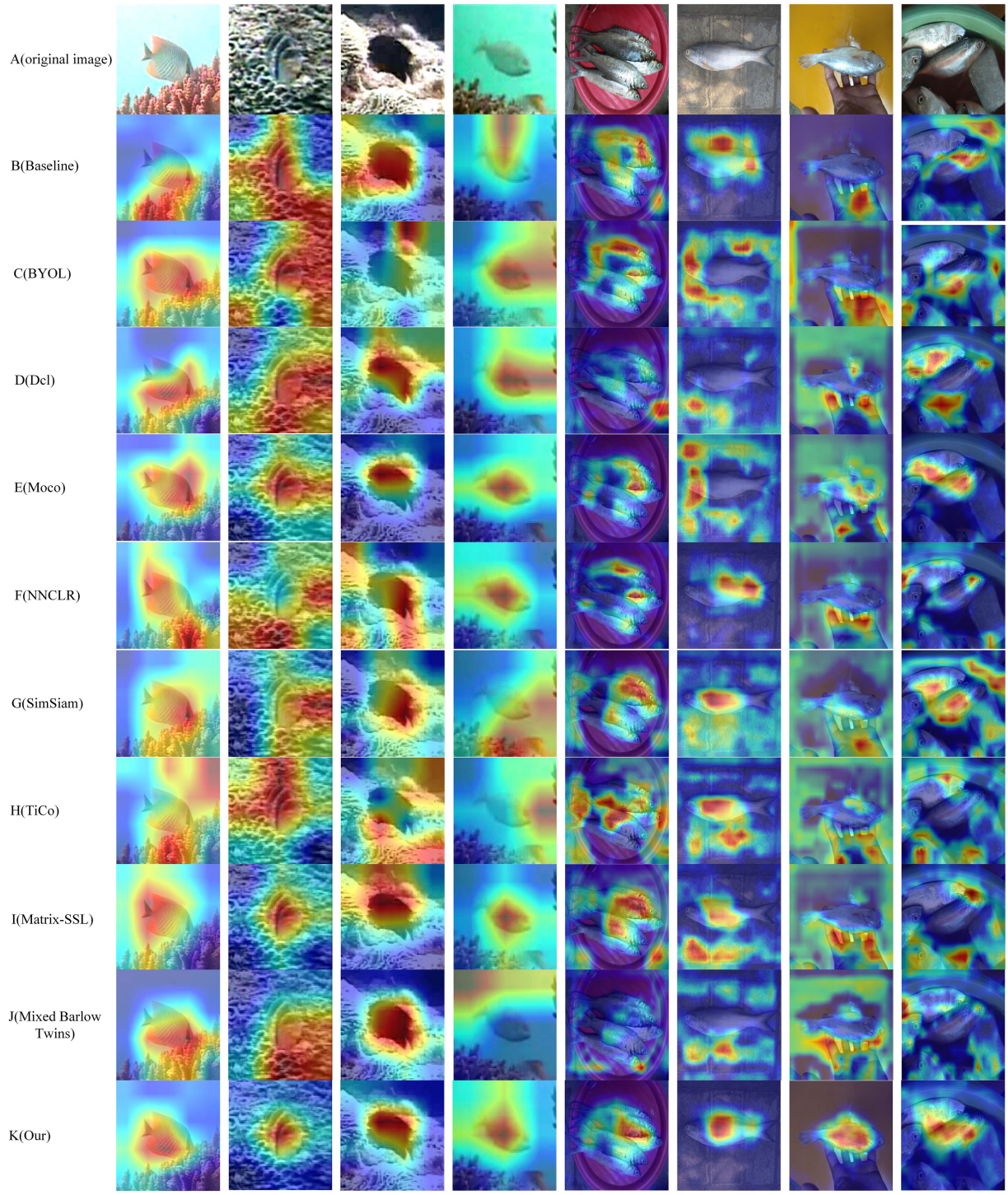

FIGURE 8
Network attention visualization. The redder the area, the more critical it is for decision or classification.

model focuses on the fish body of the image, it also focuses on the grass under the fish. In the second row, the subject of the image is similar to the background, which makes it challenging to locate the fish even with the human eye. Since the baseline model cannot accurately differentiate between the subject and the background, the baseline model incorrectly regards the grass that is highly similar to the fish as the subject of the image. However, this is not the case for CLIB. The CLIB model accurately draws the outline of the fish. Similarly, from the visualization results on the Fish-gres dataset in the seventh row, the baseline model focuses mainly on the wrist and ignores the fish in the hand, while CLIB can accurately focus on the fish in the image. In summary, the superiority of the CLIB is further verified by visualizing experiments on network attention using class activation graphs.

4.5 Analysis of parameter sensitivity

In this subsection, the performance of CLIB under different settings of parameters is shown in Figure 9. It is worth stating that the experiments are conducted on the dataset of Fish4Knowledge with resnet50 as the backbone network, and experiments can be performed under the same settings of parameters on other datasets and backbone networks as well.

The first parameter is the ratio of α to γ of half of the side length of the background image to the side length of the original image in the extraction module. Figure 9 shows the performance of CLIB changes with the ratio, and the most effective ratio is 0.25. This is because when the ratio is 0.25, the size of the mosaic of backgrounds at the four corners of the original image is equal to that of the original image. Thus, more pixel points can be used in contrastive learning, and the result is better.

The second parameter is the temperature in contrastive learning, which is used to control the model's ability to differentiate between positive and negative samples. The performance index values of CLIB in the training phase of the backbone network under different temperature parameters are shown in Figures 9, 10. It is seen that the higher the temperature parameter is, the weaker ability to differentiate the positive and negative samples of the model is, and the higher loss value in the training process of the model is. However, if the temperature is set very small, the model is difficult to converge or has poor generalization ability.The most effective temperature coefficient of CLIB is 0.07.

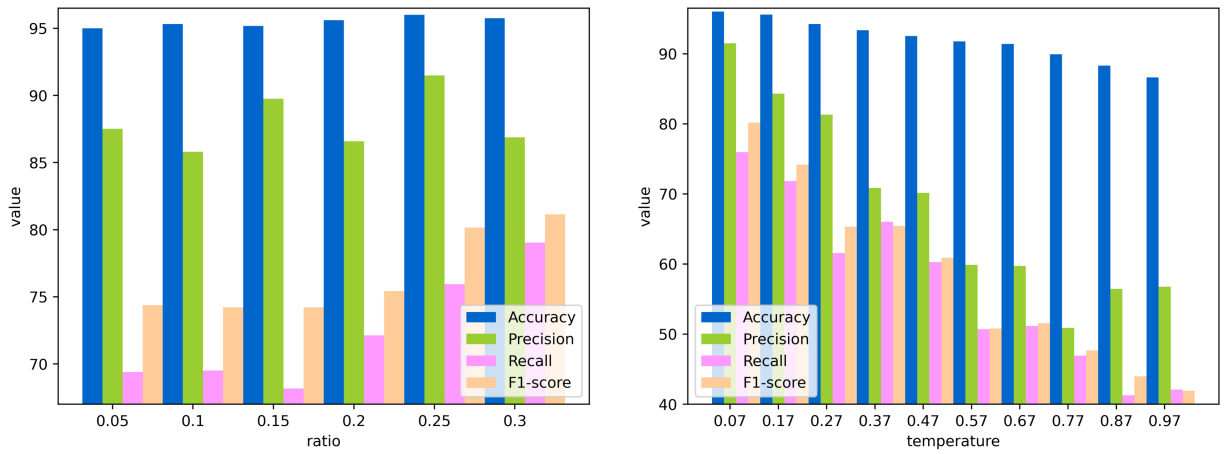

FIGURE 9
Performance of CLIB under different ratios and temperatures. **(A)** Performance under different ratios. **(B)** Performance under different temperatures.

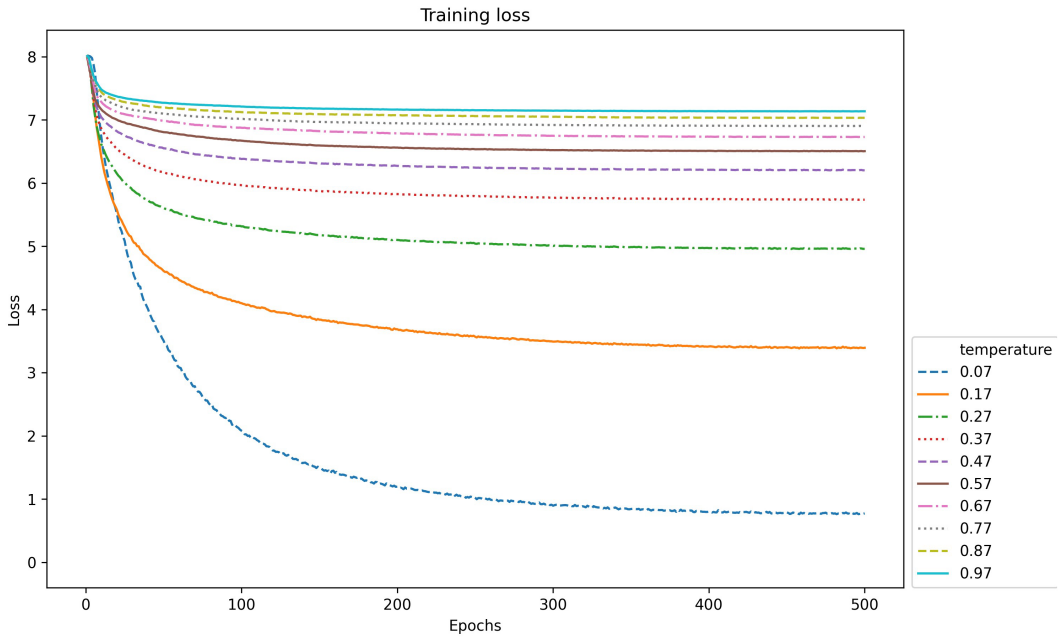

FIGURE 10
The value of loss function in backbone network training of CLIB under different temperatures and epochs.

5 Conclusion

In this paper, to solve the problem of background noise having a tremendous negative impact on underwater fish image classification, we propose the contrastive learning method of ignoring background for underwater fish image classification called CLIB. CLIB redefines views and loss function in contrastive learning based on the characteristics of underwater fish images. We demonstrate the effectiveness of CLIB in underwater fish image classification, especially when facing different resolutions, complex backgrounds, and few-sample. When the subject is located in the center of the image, the proposed CLIB method achieves the best classification effect. However, if the fish is located in a corner of the image, the CLIB treats the fish as the background, and the classification effect is decreased, which is a problem we need to solve in our subsequent work.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

QY: Conceptualization, Methodology, Writing – original draft. XD: Funding acquisition, Methodology, Writing – original draft. CL: Software, Writing – review & editing. XT: Conceptualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the National Natural Science Foundation of China (No. 61962052) and in part by the Natural Science Foundation of Qinghai Province (2024-ZJ-929).

Conflict of interest

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

Publisher's note

References

Anantharajah, K., Ge, ZY., McCool, C., Denman, S., Fookes, C., Corke, P., et al. (2014) "Local inter-session variability modelling for object classification," in *Proceedings of the IEEE winter conference on applications of computer vision*. pp. 309–316.

Bandara, W G C., De, Melo C M., and Patel, V M. (2023). Guarding Barlow twins against overfitting with mixed samples. arXiv [preprint]. arXiv:2312.02151. doi: 10.48550/arXiv.2312.02151

Boom, B., Huang, P., Beyan, C., Spampinato, C., Palazzo, S., He, J., et al. (2012). "Long-term underwater camera surveillance for monitoring and analysis of fish populations," in *Proceedings of the 21st international conference on pattern recognition*. pp. 1–4.

Boom, BJ., Huang, PX., He, J, and Fisher, RB. (2012). "Supporting ground-truth annotation of image datasets using clustering," in *Proceedings of the 21st international conference on pattern recognition*. pp. 1542–1545.

Chen, X., and He, K. (2021). "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 15750–15758.

Chen, T., Kornblith, S., Norouzi, M., Hinton, and Geoffrey. (2020). "A simple framework for contrastive learning of visual representations," in *Proceedings of the international conference on machine learning*. pp. 1597–1607.

Chen, J., Wu, H. T., Lu, L., Luo, X., and Hu, J. (2022). Single underwater image haze removal with a learning-based approach to blurriness estimation. *J. Vis. Commun. Image Represent.* 89:103656. doi: 10.1016/j.jvcir.2022.103656

Deep, BV., Dash, R. (2019). "Underwater fish species recognition using deep learning techniques," in *Proceedings of the 6th international conference on signal processing and integrated networks*. pp. 665–669.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. (2021) "With a little help from my friends: nearest-neighbor contrastive learning of visual representations," in *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9588–9597.

Ericsson, L., Gouk, H., Loy, C. C., Loy, C. C., and Hospedales, T. M. (2022). Self-supervised representation learning: introduction, advances, and challenges. *IEEE Signal Process. Mag.* 39, 42–62. doi: 10.1109/MSP.2021.3134634

Fouad, MMM., Zawbaa, HM., El-Bendary, N, and Hassanien, AE. (2013). "Automatic nile tilapia fish classification approach using machine learning techniques," in *Proceedings of the 13th international conference on hybrid intelligent systems*. pp. 173–178.

Georgian, S., Hameed, S., Morgan, L., Amon, D. J., Sumaila, U. R., Johns, D., et al. (2022). Scientists' warning of an imperiled ocean. *Biol. Conserv.* 272:109595. doi: 10.1016/j.biocon.2022.109595

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Proces. Syst.* 33, 21271–21284. doi: 10.48550/arXiv.2006.07733

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., et al. (2020). "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9729–9738.

He, K., Zhang, X., Ren, S., Sun, and Jian. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778.

Hu, J., Li, D., Duan, Q., Han, Y., Chen, G., and Si, X. (2012). Fish species classification by color, texture and multi-class support vector machine using computer vision. *Comput. Electron. Agric.* 88, 133–140. doi: 10.1016/j.compag.2012.07.008

Huang, PX., Boom, BJ., Fisher, RB. (2012). "Underwater live fish recognition using a balance-guaranteed optimized tree," in *Proceedings of the computer vision–ACCV 2012: 11th Asian conference on computer vision*. pp. 422–433.

Jiao, N., Luo, T., Liu, J., Chen, Q., Zhu, C., et al. (2023). Ocean negative carbon emissions in the context of earth system science. *Bull. Chin. Acad. Sci.* 38, 1294–1305. doi: 10.16418/j.issn.1000-3045.20230726004

Khotimah, WN., Arifin, AZ., Yuniarti, A., Wijaya, AY., Navastara, DA., Kalbuadi, MA., et al. (2015). "Tuna fish classification using decision tree algorithm and image processing method," in *Proceedings of the 2015 international conference on computer, control, informatics and its applications*. pp. 126–131.

Larsen, R., Olafsdottir, H., and Ersbøll, BK. (2009). "Shape and texture based classification of fish species," in *Proceedings of the image analysis: 16th Scandinavian conference*. pp. 745–749.

Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J., et al. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Networks Learn. Syst.* 33, 6999–7019. doi: 10.1109/TNNLS.2021.3084827

Li, G., Wang, F., Zhou, L., Jin, S., Xie, X., and Ding, C. (2023). MCANet: multi-channel attention network with multi-color space encoder for underwater image classification. *Comput. Electr. Eng.* 108:108724. doi: 10.1016/j.compeleceng.2023.108724

Lu, S., Guan, F., Zhang, H., and Lai, H. (2023). Underwater image enhancement method based on denoising diffusion probabilistic model. *J. Vis. Commun. Image Represent.* 96:103926. doi: 10.1016/j.jvcir.2023.103926

Ovalle, J. C., Vilas, C., and Antelo, L. T. (2022). On the use of deep learning for fish species recognition and quantification on board fishing vessels. *Mar. Policy* 139:105015. doi: 10.1016/j.marpol.2022.105015

Prasetyo, E., Suciati, N., and Fatichah, C. (2020). Fish-gres dataset for fish species classification. Mendeley Data. 10.17632/76cr3wfhff.1.

Sang, Q., Shu, Z., Liu, L., Hu, C., and Wu, Q. (2022). Image quality assessment based on self-supervised learning and knowledge distillation. *J. Vis. Commun. Image Represent.* 90:103708. doi: 10.1016/j.jvcir.2022.103708

Selvaraju, RR., Cogswell, M., Das, A, Vedantam, R, Parikh, D, Batra, D, et al. (2017) "Grad-cam: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*. pp. 618–626.

Spampinato, C., Giordano, D., Di, Salvo R, Chen-Burger, YHJ., Fisher, RB., Nadarajan, G., et al. (2010). "Automatic fish classification for underwater species behavior understanding," in *Proceedings of the first ACM international workshop on analysis and retrieval of tracked events and motion in imagery streams*. pp. 45–50.

Sun, X., Shi, J., Dong, J., and Wang, X. (2016). "Fish recognition from low-resolution underwater images," in *Proceedings of the 9th international congress on image and signal processing, BioMedical engineering and informatics*. pp. 471–476.

Trindade-Santos, I., Moyes, F., and Magurran, A. E. (2022). Global patterns in functional rarity of marine fish. *Nat. Commun.* 13:877. doi: 10.1038/s41467-022-28488-1

Wang, L., Xu, L., Tian, W., Zhang, Y., Feng, H., and Chen, Z. (2022). Underwater image super-resolution and enhancement via progressive frequency-interleaved network. *J. Vis. Commun. Image Represent.* 86:103545. doi: 10.1016/j.jvcir.2022.103545

Xu, X., Li, W., and Duan, Q. (2020). Transfer learning and SE-ResNet152 networks-based for small-scale unbalanced fish species identification. *Comput. Electron. Agric.* 180:105878. doi: 10.1016/j.compag.2020.105878

Xuan, Z. Y., Jiang, T., Liu, H. B., Chen, X. B., Hu, Y. H., Yang, J., et al. (2022). Advances in the application of otolith microchemistry analysis in fish population ecology. *Progress Fish. Sci.* 43, 01–14. doi: 10.19663/j.issn2095-9869.20210528002

Yeh, CH., Hong, CY., Hsu, YC, Liu, TL, Chen, Y, LeCun, Y, et al. (2022) "Decoupled contrastive learning," in *Proceedings of the European conference on computer vision*. pp. 668–684.

Yu, F., Liu, F., Xia, Z., Xu, C., Wang, J., Tang, R., et al. (2022). Integration of ABC curve, three dimensions of alpha diversity indices, and spatial patterns of fish assemblages into the health assessment of the Chishui River basin, China. *Environ. Sci. Pollution Res.* 29, 75057–75071. doi: 10.1007/s11356-022-20648-6

Zhang, Z., du, X., Jin, L., Han, D., Li, C., and Liu, X. (2021). Discriminative feature learning for underwater fish recognition. *J. Elect. Imaging* 30:023020. doi: 10.1117/1.JEI.30.2.023020

Zhang, Z., du, X., Jin, L., Wang, S., Wang, L., and Liu, X. (2022). Large-scale underwater fish recognition via deep adversarial learning. *Knowl. Inf. Syst.* 64, 353–379. doi: 10.1007/s10115-021-01643-8

Zhang, D., O'Conner, N. E., Simpson, A. J., Cao, C., Little, S., Wu, B., et al. (2022). Coastal fisheries resource monitoring through a deep learning-based underwater video analysis. *Estuar. Coast. Shelf Sci.* 269:107815. doi: 10.1016/j.ecss.2022.107815

Zhang, Y., Tan, Z., Yang, J., Huang, W., and Yuan, Y. (2023). Matrix information theory for self-supervised learning. arXiv [preprint]. arXiv:2305.17326.

Zhu, J., Moraes, RM., Karakulak, S., Sobol, V., Canziani, A., LeCun, Y., et al. (2022). Tico: transformation invariance and covariance contrast for self-supervised visual representation learning. arXiv [preprint]. arXiv: 2206.10698. doi: 10.48550/arXiv.2206.10698

Zhuang, P., Wang, Y., and Qiao, Y. (2018) "Wildfish: a large benchmark for fish recognition in the wild," in *Proceedings of the 26th ACM international conference on multimedia*. pp. 1301–1309.