frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 28 May 2024
DOI 10.3389/fmars.2024.1373755

OPEN ACCESS

EDITED BY
Haiyong Zheng,
Ocean University of China, China

REVIEWED BY
Zhuhua Hu,
Hainan University, China
Deepayan Bhowmik,
Newcastle University, United Kingdom

*CORRESPONDENCE
Ye Li
yli@geo.ecnu.edu.cn
Min Liu
mliu@geo.ecnu.edu.cn

RECEIVED 20 January 2024
ACCEPTED 02 May 2024
PUBLISHED 28 May 2024

CITATION
Ma D, Wei J, Zhu L, Zhao F, Wu H, Chen X, Li Y and Liu M (2024) Semi-supervised learning advances species recognition for aquatic biodiversity monitoring. *Front. Mar. Sci.* 11:1373755. doi: 10.3389/fmars.2024.1373755

Semi-supervised learning advances species recognition for aquatic biodiversity monitoring

Dongliang Ma[1], Jine Wei[2], Likai Zhu[1], Fang Zhao[1], Hao Wu[3], Xi Chen[1], Ye Li[1*] and Min Liu[1*]

[1]Key Laboratory of Geographic Information Science, Ministry of Education, School of Geographic Sciences, East China Normal University, Shanghai, China, [2]State Key Laboratory of Estuarine and Coastal Research, East China Normal Normal University, Shanghai, China, [3]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China

Aquatic biodiversity monitoring relies on species recognition from images. While deep learning (DL) streamlines the recognition process, the performance of these method is closely linked to the large-scale labeled datasets, necessitating manual processing with expert knowledge and consume substantial time, labor, and financial resources. Semi-supervised learning (SSL) offers a promising avenue to improve the performance of DL models by utilizing the extensive unlabeled samples. However, the complex collection environments and the long-tailed class imbalance of aquatic species make SSL difficult to implement effectively. To address these challenges in aquatic species recognition within the SSL scheme, we propose a Wavelet Fusion Network and the Consistency Equilibrium Loss function. The former mitigates the influence of data collection environment by fusing image information at different frequencies decomposed through wavelet transform. The latter improves the SSL scheme by refining the consistency loss function and adaptively adjusting the margin for each class. Extensive experiments are conducted on the large-scale FishNet dataset. As expected, our method improves the existing SSL scheme by up to 9.34% in overall classification accuracy. With the accumulation of image data, the improved SSL method with limited labeled data, shows the potential to advance species recognition for aquatic biodiversity monitoring and conservation.

KEYWORDS

deep learning, semi-supervised learning, aquatic species recognition, wavelet transform, consistency loss

1 Introduction

Aquatic biodiversity plays a crucial role in maintaining the structural integrity, stability, and overall health of ecosystems (Sala et al., 2021). However, anthropogenic pressures from human activities have progressively intensified in recent decades, posing gradual challenges to the preservation of aquatic biodiversity (Visbeck, 2018; Irfan and Alatawi, 2019).

A critical step in conserving aquatic biodiversity is monitoring the information regarding the abundance and distribution of aquatic animals, which relies heavily on extensive collections of underwater images and videos. Deep learning (DL) techniques have recently demonstrated significant progress in several computer vision tasks (LeCun et al., 2015), and offer a promising solution to automatic and effective species recognition from images (Rubbens et al., 2023). Due to the profound influence of dataset size and diversity on the accuracy of DL methods, many previous efforts have focused on building extensive and publicly available labeled image datasets specifically for aquatic species recognition (Zhuang et al., 2020; Katija et al., 2022; Khan et al., 2023). Unfortunately, the intricate taxonomy of species typically demands a high level of expertise in the aquatic domain, meanwhile the annotation process proves to be tedious and time-consuming (Li et al., 2023).

It is estimated that more than 300,000 hours of underwater video footage have been collected worldwide so far, with only less than 15% of the data annotated by biological and ecological experts (Bell et al., 2023). As the pace of data collection accelerates annually, the substantial backlog exacerbates. Several strategies, including transfer learning (Qiu et al., 2018), data augmentation (Saleh et al., 2020), weakly supervised learning (Laradji et al., 2021), and active learning (Moller et al., 2017), have been made to tackle this problem. For example, transfer learning necessitates fine-tuning newly labeled aquatic species datasets to maximize accuracy. Weakly-supervised learning, on the other hand, relies on a limited form of supervision, where the labels may be noisy, incomplete, or imprecise. Nonetheless, these studies still require access to large-scale labeled training sets. The significance of diversity and comprehensiveness in the training dataset undoubtedly plays a pivotal role in achieving high recognition accuracy during real-world model deployment. Given the existence of unlabeled data, the marine community has emphasized the need for a powerful approach to training DL methods on vast amounts of data without annotated labels. In contrast, semi-supervised learning (SSL) can handle scenarios with both labeled and unlabeled data, providing more flexibility and potentially better performance when limited labeled data is available (Yang et al., 2022). To date, although numerous studies explore SSL to address the high cost of annotated labels in aquatic domain (Choi et al., 2021; Cai et al., 2023; Jahanbakht et al., 2023), its application in aquatic environments for species recognition remains scarce.

Two major challenges conspire to hinder the use of SSL scheme for aquatic species recognition. The first challenge stems from the unique characteristics of collected environments, including diverse lighting, variable water turbidity, and complex visual backgrounds that can obscure visual information (Ditria et al., 2020; Saleh et al., 2022). Furthermore, the movement of objects in an uncontrolled environment can introduce distortion, deformation, occlusion, and overlapping (Li et al., 2023; Ma et al., 2023). These factors increase complexities and hinder the ability of DL models to employ effectively from labeled to unlabeled data. The need for robust feature extraction methods tailored to the above challenges becomes paramount to ensure the practical applicability of the SSL scheme. The second challenge arises from the long-tailed class imbalance of aquatic species in collected images (Rubbens et al., 2023). As shown in Figure 1A, a limited subset of species are characterized by a substantial number of samples (referred to as head classes), while others are linked to only a few samples (referred to as tail classes). The limited sample information of tail classes poses a significant hurdle for SSL scheme, as there is a risk that the model being biased toward head classes due to the abundance of samples (Zhang et al., 2023).

In this work, we propose a novel SSL scheme for aquatic species recognition, which is based on the existing SSL algorithm, FixMatch (Sohn et al., 2020). Specifically, to mitigate the complexities inherent in heterogeneous collected environments, we propose a robust wavelet fusion network (WFN) equipped with wavelet transform. The proposed network comprises two frequency-aware streams, one is dedicated to capturing subtle image details by focusing on high-frequency (HF) information, while the other aims to extract high-level semantics from low-frequency (LF) information. These streams are subsequently integrated through a

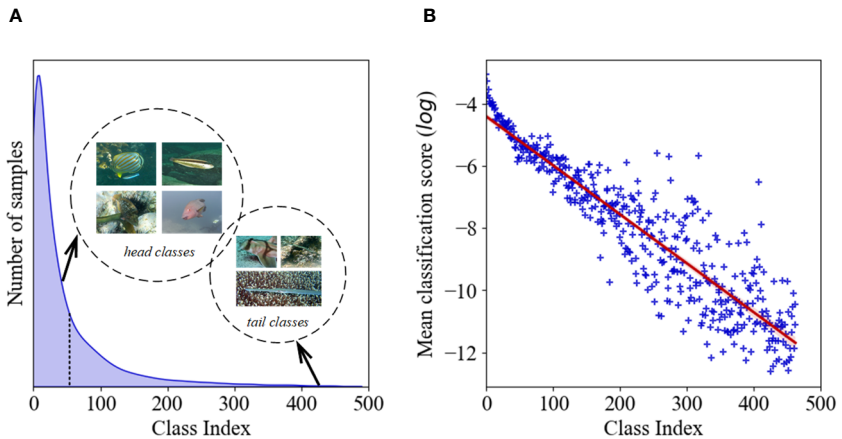

FIGURE 1
(A) The label distribution of a long-tailed aquatic species dataset (e.g., the FishNet dataset (Khan et al., 2023) with more than 450 classes). **(B)** Statistics of mean classification score for each class on the FishNet dataset. The x-axis depicts the index corresponding to the corresponding class.

FusionBlock, which facilitates attentive interactions between the LF and HF streams. Furthermore, for the problem of long-tailed nature when using unlabeled data, we design a new Consistency Equilibrium Loss (CEL) that refines the pseudo-labels and adaptively adjusts the margin for each aquatic species class. We find that replacing the unsupervised loss with CEL could ensure that the SSL algorithm achieves relative classification equilibrium, even if the collected data distribution is biased toward the head classes. Extensive experiments demonstrate the proposed method attains superior results on a large-scale aquatic species recognition dataset. In addition, the WFN and CEL are assessed to highlight their advantages over current common practices.

2 Related work

2.1 Aquatic species recognition with deep learning

In recent years, DL-based aquatic species recognition has emerged as a promising tool for assisting marine scientists and ecologists in better understanding and managing marine environments. Accurate species recognition serves as the cornerstone of aquatic biodiversity research, playing a crucial role in estimating species size and quantity. A seminal contribution in this field is the development of the filtering deep convolutional network (FDCNet) (Lu et al., 2018), which effectively classifies deep-sea objects such as sea urchins, crabs, sharks, and shrimps. Due to the complexity and dynamics of the marine environment, DL methods encounter challenges in recognizing interesting objects based on visual characteristics. To overcome this issue, the literature (Kaur and Vijay, 2023) proposes an invariant feature-based species classification method for distinguishing octopus and crabs. Similarly, the study (Liu et al., 2023) introduces an improved fish recognition network along with a novel loss function, FishFace, designs to focus more attention on fish details. More recently, automated plankton recognizing method based on DL has been developed for continuous monitoring of living plankton abundance in aquatic environments (Chen et al., 2023). A comprehensive review (Li et al., 2023) is recommended for researchers to seek an in-depth understanding of DL-based aquatic species recognition methods. However, most existing methods are constrained by their reliance on a relatively small portion of labeled data, posing a challenge to their practical application in real-world scenarios (Khan et al., 2023). Therefore, there is an urgent need to develop a new paradigm capable of effectively utilizing extensive unlabeled data with a small amount of labeled data to accurately identify a broader range of aquatic species, thereby supporting aquatic biodiversity conservation efforts.

2.2 Semi-supervised learning

SSL methods have garnered significant attention from both industry and academia for use unlabeled data during the training process, particularly when the amount of labeled data is scarce. Recent SSL research has generally been categorized into two main groups. The first category of consistency regularization methods imposes a classification invariance loss on unlabeled data following perturbation (Miyato et al., 2018; Xie et al., 2020a). In the second category, pseudo-labeling extends model training data beyond labeled samples to contain additional unlabeled data, augmented with credible pseudo-labels (Berthelot et al., 2019b; Xie et al., 2020b). Techniques like FixMatch (Sohn et al., 2020) and RemixMatch (Berthelot et al., 2019a) combine pseudo-labeling with consistency regularization, yielding superior performance compared to many other SSL algorithms in image recognition tasks. Furthermore, several studies have been conducted experiments on long-tailed SSL. For example, DARP (Kim et al., 2020) proposes eliminating biased pseudo-labels through distribution alignment, which refines the pseudo-labels based on the labeled data distribution. Additionally, an auxiliary balanced classifier learned by down-sampling the head class is used to enhance generalization capabilities (Lee et al., 2021). The above designs largely promote the overall performance of long-tailed semi-supervised methods, but the performance of natural long-tailed SSL problems in aquatic species recognition is still unsatisfactory, and no research has been found that effectively addresses this issue.

2.3 Wavelet-based deep learning

The integration of wavelet transform with deep neural networks (DNNs) has gained traction due to its robust frequency and spatial representation capabilities. Common strategies involve utilizing wavelet transform as either a pre-processing or post-processing step (Huang et al., 2017; Yin and Xu, 2021), as well as substituting specific layers in DNNs (Li et al., 2021). Previous research has also explored the application of the dual-tree complex wavelet transform to extract robust features from Synthetic Aperture Radar images (Duan et al., 2017). More recently, Wave-ViT (Yao et al., 2022) uses the wavelet transform to down-sample keys/values in a Transformer (Vaswani et al., 2017). The Multi-level Wavelet CNN (Liu et al., 2018) integrates wavelet package transform into the DNN to concatenate the LF and HF components and process them in a unified manner, despite the notable disparity between these components. In contrast, we employ wavelet transform as an effective approach to tackle image complexity. Further, none of these studies has attempted to design a fusion block specially tailored for the wavelet transform paradigm to obtain attentive feature representation.

2.4 Loss function for long-tailed learning

Re-weighting and Re-margining loss functions serve as key components in tackling long-tailed class imbalanced challenges (Zhang et al., 2023). These methods are primarily implemented by adjusting margins or loss weights based on the distribution of training data. For instance, seminal works (Cui et al., 2019; Ren et al., 2020) reweight the loss functions according to the sampling

frequency of each class. Recent literature (Lai et al., 2022) enhances the robustness of SSL to long-tailed class imbalanced problems by designing weights in the unsupervised loss based on estimating the learning difficulty of each class. In contrast, several studies (Cao et al., 2019; Menon et al., 2020; Tan et al., 2020) have attempted to adjust the loss margins of each class. The Label-Distribution-Aware-Margin (Cao et al., 2019) approach motivates tail classes to have larger margins based on label frequencies. Additionally, the study (Feng et al., 2021) replaces the margin term with mean classification score for long-tailed object detection. While our CEL function is inspired by the above pioneer studies, it differs significantly in two aspects. Firstly, to the best of our knowledge, the CEL function is the first to utilize the mean classification score to extend the existing consistency loss in SSL. Secondly, our key idea involves refining pseudo-labels via the mean classification score to match the true data distribution. With the proposed CEL function, our approach demonstrates superior performance in aquatic species recognition based on the SSL scheme.

3 Method

In this section, we first revisit the formulation of the SSL scheme in Section 3.1. After that, we illustrate the process of generating LF and HF entities using wavelet transform in Section 3.2, and provide detailed insights into our FusionBlock in Section 3.3. Lastly, along with the SSL scheme, we introduce the CEL function for unlabeled samples in Section 3.4. An overview of the framework is shown in Figure 2.

3.1 Semi-supervised learning setup

The basic technique utilized in FixMatch (Sohn et al., 2020) revolves around pseudo-labeling and consistency-regularization, where unlabeled samples with high confidence are selected as training samples. Suppose we have a labeled dataset $X_L = \{(x_i, y_i)\}_{(i=1)}^{L}$, where x_i is the i^{th} training sample $y_i \subseteq \{0,1\}^C$ is the corresponding label with C classes, and L is the number of labeled samples. $X_U = \{(x_i)\}_{i=L+1}^{L+U}$ represents a dataset comprising unlabeled samples, where U is the number of unlabeled samples. Both X_L and X_U share identical semantic labels. The loss function is composed of two terms: $L = L_s + \lambda_u \times L_u$, where L_s denotes the supervised loss applied to labeled data, L_u is the consistency loss for unlabeled data, and λ_u is a scalar hyperparameter.

The supervised loss L_s is defined as: $L_s = \frac{1}{B}\sum_{i=1}^{B} H(y_i, p(\eta(x_i)))$, where η denotes the weak augmentation, B is the batch size, H is the cross-entropy loss, and $p(\cdot)$ is the output of logits in DNN. Pseudo-labels $\hat{y}_i = argmax(softmax(p(\eta(x_i))))$ are generated from weakly augmented unlabeled samples, guiding the prediction of model on strongly augmented samples. The consistency loss L_u can be formally expressed as: $L_u = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathrm{II}[max(softmax(p(\eta(x_i)))) \geq \tau] H(\hat{y}_i, p(\phi(x_i)))$, where ϕ represents strong augmentation, μ governs the proportion of labeled to unlabeled samples in a minibatch, and II is the indicator function; 0 if the highest probability of unlabeled samples is below the confidence threshold τ and 1 otherwise.

3.2 Wavelet transform

The wavelet transform serves as an effective frequency analysis tool, establishing extensive applications in signal processing (Mallat, 1989). A wavelet is linked with wavelet and scaling functions, which establish a relationship with the low-pass and high-pass filters to facilitate data decomposition. In practice, the images represent discrete non-stationary signals, involving various frequency intervals and spatial location information. Single-level 2D discrete wavelet transform (Equation 1) with four filters (f_{LF},$f_{HF_{horizontal}}$,$f_{HF_{vertical}}$,and $f_{HF_{diagonal}}$) are often used to decompose an image x to obtain its LF component LF and three HF components $HF_{horizontal}$, $HF_{vertical}$, $HF_{diagonal}$,

$$i = Conv_{f_i}(x) \downarrow_2, \quad i \in \{LF, HF_{horizontal}, HF_{vertical}, HF_{diagonal}\}, \tag{1}$$

where $Conv_{f_i}$, $\downarrow_2$ denote the convolution operation with the typical filter f_i and downsampling operation, receptively. The components acquired through wavelet transform contain distinct information about the raw images of aquatic species (see Figures 3A, B). Our method strives to leverage wavelet transform to generate robust information as the input of DNN to extract LF and HF features. As such, a LF entity is represented solely by a LF component (Equation 2), while a HF entity is represented as a set of HF components in

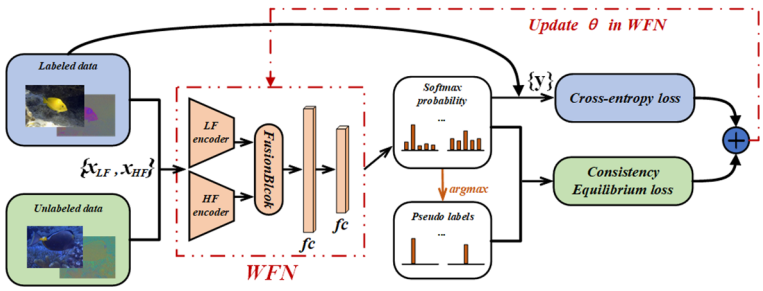

FIGURE 2
Illustration of the overall framework for aquatic species recognition. The proposed WFN (Wavelet Fusion Network) and CEL (Consistency Equilibrium Loss) are added into the exiting SSL scheme FixMatch (Khan et al., 2023).

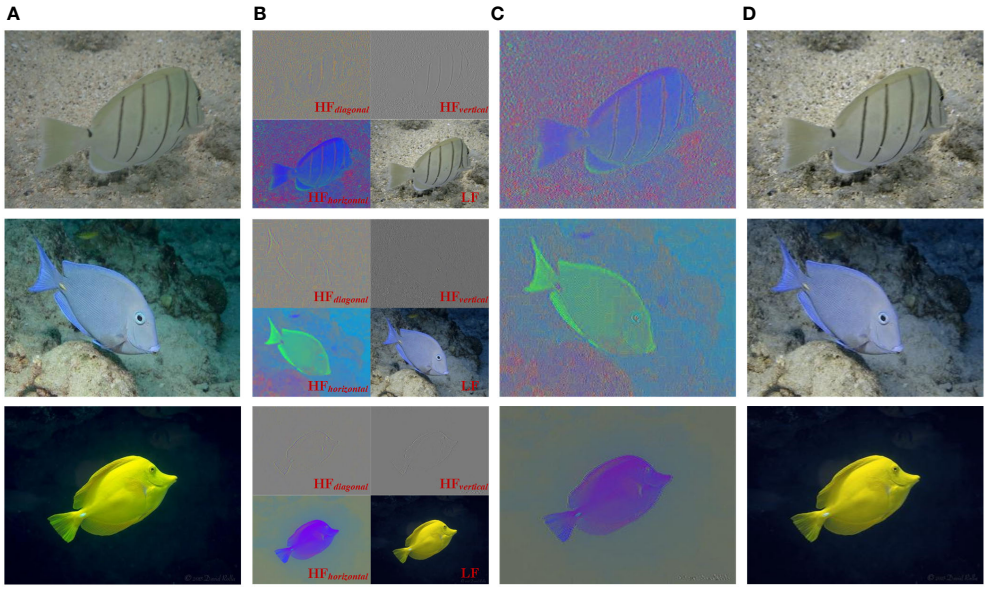

FIGURE 3
Taking FishNet (Khan et al., 2023) as an example, visualize LF and HF results. **(A)** Raw image. **(B)** Wavelet transform results. **(C)** the HF entity. **(D)** the LF entity.

various directions (Equation 3) similar to those used in (Zhou et al., 2023):

$$LF = LF, \tag{2}$$

$$HF = average(HF_{horizontal} + HF_{vertical} + HF_{diagonal}). \tag{3}$$

Note that our average HF components aim to reduce computational costs by decreasing the number of subsequent encoders. Ideally, each HF component would be feature-extracted by a specific encoder, but this is computationally expensive. In contrast, our average strategy is orthogonal and complements previous practical approaches for handling HF components, such as element-wise addition (Zhou et al., 2023), concatenation (Liu et al., 2018; de Souza Brito et al., 2021; Liu et al., 2021), and maximum (Ramamonjisoa et al., 2021). We refer to Section 4.5 for further details on these.

Figures 3C, D visually illustrate the LF and HF entities as defined above. By using the LF entity as input, DNN can focus more on LF semantics due to its less noise. In contrast, the HF entity, while exhibiting more noise, offers clearer object boundaries and shapes, enabling DNN to concentrate on HF details. A similar perspective has been adopted by Zhou et al. (2023) who argues that HF information typically represents image details, while LF information often embodies abstract semantics.

3.3 FusionBlock

Given the entities processed by wavelet transform, we employ parallel encoders equipped with ResNet-50 (He et al., 2016) to respectively generate high-level LF and HF features. These features are then passed through the FusionBlock to generate attentive features from one stream to another. We argue that applying a cross-stream attention strategy to high-level features can capture the connection between conceptual entities in the LF and HF streams, helping subsequent modules in recognizing aquatic objects in images.

Taking $f_{LF} \in \mathbb{R}^{w \times h \times c \times b}$ and $f_{HF} \in \mathbb{R}^{w \times h \times c \times b}$ as example to illustrate the details of FusionBlock (where w, h, c and b denote width, height, channel number, and batch size), we use a cross-stream attention strategy to explore correlations between the two streams. Specifically, as shown in Figure 4, the two features are passed through four 1×1 convolutional layers to generate query and key matrices. We reshape the query and key matrices into 3D spatial feature maps ($w \times h \times cb$), and then concatenate them to obtain the fused key and query as (Equations 4, 5):

$$f_{fused-key} = concate(R(conv_{1\times1}(f_{LF})), R(conv_{1\times1}(f_{HF}))), \tag{4}$$

$$f_{fused-query} = concate(R(conv_{1\times1}(f_{LF})), R(conv_{1\times1}(f_{HF}))), \tag{5}$$

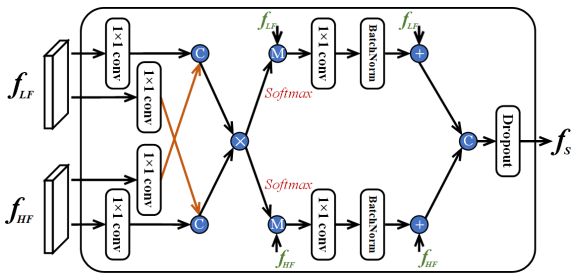

FIGURE 4
Diagram of the proposed FusionBlock. Here, "C" signifies feature concatenation, while "+" represents element-wise addition, "×" denotes dot-product, and "M" signifies element-wise multiplication.

where R denotes the reshape operation, $f_{fused-key} \in \mathbb{R}^{w\times 2h\times cb}$, and $f_{fused-query} \in \mathbb{R}^{w\times 2h\times cb}$. After that, the attention map $A \in \mathbb{R}^{w\times h\times c\times b}$ is computed by performing a dot-product and applying the softmax activation function (Equation 6).

$$A = R(\sigma(f_{fused-query} \times f^{T}_{fused-key}), \tag{6}$$

where σ is the softmax activation function. In this way, the feature from one stream could serve to augment another stream. Additionally, to preserve the original information of each stream, a residual connection is employed to fuse the enhanced features with their original counterparts. As such, we obtain the cross-stream attentive features for the two streams as (Equation 7):

$$\begin{cases} f'_{LF} = f_{LF} + Bconv_{1\times 1}(f_{LF} \text{Ⓜ} \text{A}) \\ f'_{HF} = f_{HF} + Bconv_{1\times 1}(f_{HF} \text{Ⓜ} \text{A}) \end{cases}, \tag{7}$$

where Ⓜ denotes element-wise multiplication, $Bconv_{1\times 1}(\cdot)$ represents a sequential operation combining a 1×1 convolutional layer and batch normalization. Once obtaining the cross-stream feature representation, we concatenate these features and apply the dropout operator to the fused feature f_S. Finally, two fully connected layers are utilized to output the final logits.

3.4 Consistency equilibrium loss

In a recent study (Feng et al., 2021), it was demonstrates that the learning status of a class can be inferred through the mean classification scores. When we take a deeper look into Figure 1B, it is evident that the head classes exhibit higher mean classification scores, whereas tail classes illustrate lower mean classification scores. Based on this observation, we follow the finding of utilizing the mean classification score to adjust the learning effectiveness of each class throughout the training process. The update process of the mean classification score during training can be illustrated as (Equation 8):

$$s = m \times s + (1 - m) \times p_y, \tag{8}$$

where $s \in \mathbb{R}^C$ denotes the mean classification score, initialized for each class using $\frac{1}{C} \cdot p_y$ is the mean predicted probability of the sample in a mini-batch, and m is a hyper-parameter.

Previous research (Kim et al., 2020) has revealed that the performance of SSL scheme is highly sensitive to the quality of pseudo-label, and a long-tailed data distribution leads to biased predictions favoring head classes. Utilizing these pseudo-labels in the SSL scheme can be harmful for tail classes. Instead of solely adjusting the class-dependent margin by deriving the mean classification score, the confirmation bias in pseudo-labels should be alleviated at the same time. To this end, we first refine the original pseudo-labels via mean classification score so that match the true data distribution (Equation 9):

$$\hat{y}_i = argmax\Big(softmax\big(p\big(\eta(x_i)\big) - \theta log(s)\big)\Big), \tag{9}$$

where θ is a hyper-parameter. Simultaneously, we adaptively adjust the margin by encouraging the tail classes to have larger margins. According to the mean classification score, we add a tunable term to balance the classification, similar to the previous study (Feng et al., 2021). As such, the CEL can be written as (Equation 10):

$$CEL = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \text{II}\Big[max\Big(softmax\big(p\big(\eta(x_i)\big) - \theta log(s)\big)\Big) \geq \tau\Big] H\big(\hat{y}_i, p\big(\phi(x_i)\big) + \theta log(s)\big). \tag{10}$$

We can control the training process through hyper-parameter θ to ensure the model remains unbiased towards the head classes and does not neglect tail classes. In particular, we increase the larger margin with lower mean classification scores for tail classes, mitigating the suppression of head classes over tail classes to balance the consistency loss.

4 Results

4.1 Dataset and evaluation metrics

Extensive experiments are conducted using the large-scale FishNet dataset (Khan et al., 2023), comprising 94,532 images encompassing 17,357 distinct species. Each species is represented by at least one associated image, which span 8 taxonomic classes, 83 orders, 463 families, and 3,826 genera. To validate the effectiveness and universality of our proposed method, we focus on the family classification task. The FishNet dataset categorizes family classes into three groups based on the class frequencies: common, medium, and rare. There are a total of 6 categories in the common group, 52 categories in the medium group, and 405 categories in the rare group. In our experiments, we report the class average accuracy for each group, as well as the overall accuracy over all categories followed as official metrics. The FishNet contains 75,631 images in the training set and 18,901 images in the test set. Unless otherwise stated, we conduct the experiments with a ratio of 20% labeled samples in the training set as labeled data, and the remaining 80% data in the training set as unlabeled data, adhering to the common semi-supervised experimental partition.

4.2 Implementation details

We implement our model using PyTorch (Paszke et al., 2019), with both training and inference procedures conducted on the NVIDIA GeForce RTX 3090 GPU. We use 200 epochs in the training process. In each training step, our batch contains 12 labeled examples and 48 unlabeled examples, maintaining a ratio $\mu = 4$ to support the SSL scheme. To ensure a smooth start, we incorporate linear learning rate warm-up for the first 50 steps, progressively increasing the initial value to 0.004. Subsequently, we decay the learning rate at epochs 30, 60, 100, and 150 by multiplying it by 0.1. For all experiments, the two-stream encoders are initialized with weights pre-trained on the ImageNet dataset (Deng et al., 2009). We adapt a relatively larger coefficient $m = 0.99$ for the mean classification score update. For the unsupervised loss function

CEL, the weight parameter (λ) increases linearly per epoch according to $\lambda = \lambda_u \times \frac{epoch}{epoch_{max}}$, and the confidence threshold τ is set to 0.95. As in previous works (Sohn et al., 2020; Lai et al., 2022), we employ an exponential moving average of model parameters to generate the final performance. We keep other hyper-parameters the same as the ImageNet experiments in FixMatch, except for those mentioned above.

4.3 Comparison of aquatic species recognition performance

Several experiments are conducted to elaborate the findings: (a) the baseline utilizing only labeled images for aquatic species recognition based on ResNet-50 (He et al., 2016); (b) an improved version of (a) incorporating our proposed WFN to enhance image features with wavelet transform; (c) the baseline utilizing both labeled images and unlabeled images based on the representative SSL scheme FixMatch (Sohn et al., 2020); (d) the proposed WFN integrated into the FixMatch scheme; (e-i) evaluation of state-of-the-art methods designed for long-tailed SSL on the FishNet dataset; (j) utilization of CEL combined with FixMatch; (k) is the final version of our proposed methods incorporating both CEL and WFN into the FixMatch scheme. Table 1 presents the performance of different methods on the FishNet dataset. Based on these results, several observations emerge regarding the overall progress of the proposed method and variations among different supervised types.

From a → b, it is evident that the WFN significantly improves overall performance. WFN achieves competitive performance, with average classification accuracy of 72.73%, 58.60%, 25.47%, and 29.80%, surpassing the ResNet-50 by 2.1%, 0.95%, 4.06%, and 3.68% over four metrics. The experiment demonstrates that WFN equipped with wavelet transform and FusionBlock, has better generalization than previous ResNet-50 architecture, which allows the model to tackle the challenges posed by the heterogeneous aquatic environment. From a → c, we can observe that the use of SSL yields a notable enhancement compared to the model trained solely using labeled data. The gain from unlabeled data becomes evident in the aquatic species recognition. SSL enables the DL model to leverage the abundance of unlabeled images, further refining its understanding of various species and environmental conditions.

From b → d, we can infer a similar conclusion to a → c. Furthermore, the combination of WFN and SSL yields a synergistic effect, tackling the challenges posed by the heterogeneity of aquatic environments while leveraging the benefits afforded by unlabeled data. Incorporating WFN into the SSL scheme enables the DL model to acquire robust features across diverse aquatic conditions. In other words, it is crucial to acknowledge that enhanced performance of the robust feature extraction method within SSL extends beyond the initial finding observed in a to c.

Table 1 also compares the proposed CEL function with several other methods: CReST (Wei et al., 2021), which oversample tail classes generated by pseudo-labels, ABC (Lee et al., 2021), utilizing an auxiliary balanced classifier of a single layer, DARP (Kim et al., 2020), refining pseudo-labels to match the true distribution of unlabeled data, SAW Lai et al. (2022), adjusting weights based on the estimated learning difficulty of each class in unsupervised loss, and DASO (Oh et al., 2022), employing a blending pseudo-labels strategy to mitigate the overall bias. Since these methods were originally experiment with in the long-tailed SSL domain, we evaluate their performance on the FishNet dataset. We utilize publicly available code to train each method and report the best results obtained from multiple runs, fine-tuning their hyper-parameters to ensure optimal performance. From (e, f, g, h, i) → j, we observe that our CEL achieves competitiveness with other methods on the FishNet dataset. From c → (e, f, g, h, i, j), the long-tailed extensions yield performance gains of varying degrees for all methods, such as a notable 4.36% increase in average classification accuracy for DASO, demonstrating the importance of long-tailed distribution as a general issue for the task of aquatic species recognition.

Lastly, group (k) demonstrates that integrating WFN and CEL within SSL enhances overall performance for the aquatic species

TABLE 1 Comparison with supervised, semi-supervised, and long-tailed semi-supervised methods on the FishNet dataset.

	Method	SSL	LT	Common	Medium	Rare	All
a)	ResNet-50 (He et al., 2016)	–	–	70.63	57.65	21.41	26.12
b)	WFN	–	–	72.73	58.60	25.47	29.80
c)	FixMatch (Sohn et al., 2020)	✓	–	79.07	64.94	22.24	27.77
d)	FixMatch + WFN	✓	–	81.65	67.61	27.99	33.13
e)	Fixmatch+CReST (Wei et al., 2021)	✓	✓	68.19	67.26	24.93	30.24
f)	Fixmatch+ABC (Lee et al., 2021)	✓	✓	69.14	66.71	24.98	30.24
g)	Fixmatch+DARP (Kim et al., 2020)	✓	✓	69.74	67.42	26.19	31.38
h)	FixMatch+SAW (Lai et al., 2022)	✓	✓	64.54	67.18	27.31	32.27
i)	FixMatch+DASO (Oh et al., 2022)	✓	✓	65.74	67.70	27.07	32.13
j)	FixMatch+CEL	✓	✓	67.75	68.83	28.30	33.36
k)	FixMatch+CEL+WFN	✓	✓	69.58	68.36	32.61	37.11

recognition task. The collaborative integration of WFN and CEL could leverage the strengths of each component. WFN enhances the feature extraction capabilities of the model, enabling better handling of the complexities of the aquatic environment. Meanwhile, CEL guides the training process, ensuring that the model benefits from unlabeled data and mitigating long-tailed class imbalanced problems. Through rigorous evaluation, we demonstrate that the combined strength of WFN and CEL contributes to a more robust and accurate aquatic species recognition system, paving the way for advancements in the field of aquatic biodiversity research and conservation.

4.4 Ablation study

4.4.1 Impact of different wavelet bases in WFN

Table 2 presents an analysis of various wavelet bases trained on labeled data, including Dmey, Haar, Daubechies 2, Coiflets, Biorthogonal 1.5, and Biorthogonal 2.4. The results we obtained show that the Daubechies 2 wavelet has better classification accuracy, and the Haar wavelet presents better border accuracy. As such, we select the Daubechies 2 wavelet basis as the default for our experiments.

4.4.2 Ablation study of different coefficient θ in CEL

We perform an ablation study on the CEL function with various values of θ to evaluate the impact of model performance. As shown in Figure 5, an improper proportion of the term, either too large or too small, impedes the attainment of optimal performance. Observing the CEL function reveals a significant variation in the impact of θ. When the value of θ is set to 0, the CEL is equivalent to the consistency loss of FixMatch. However, excessively large values of θ may hinder the ability of model to focus attention on the data, whereas too small values inadequately addresses the bias in long-tailed SSL problem. The trade-off between model performance and CEL when $\theta = 0.4$ achieves the relatively best performance.

4.4.3 Comparison of fusion strategies for WFN

We further examine the effectiveness of the proposed FusionBlock in Table 3. We utilize different feature fusion strategies to train the DNNs on the labeled images combined with wavelet transform. The proposed FusionBlock achieves better performance on the FishNet test set compared with element-wise add operation and concatenate features along with channel dimension. We believe that the cross-stream attention fusion strategy is more effective for learning interactive features, making it well-suited for the diverse and challenging environment in aquatic species recognition.

4.5 Analysis of different frequency components

Since the main semantic information is conveyed in the LF component, previous studies have often used the LF component alone in certain tasks (Li et al., 2020; Zhao et al., 2023). However, researchers have attempted to aggregate HF components with methods such as concatenation (Liu et al., 2018; de Souza Brito et al., 2021; Liu et al., 2021), maximum (Ramamonjisoa et al., 2021), or element-wise addition (Zhou et al., 2023), and incorporate them into DNNs to improve model performance. To verify the effectiveness of our WFN, we compare the performance of experiments conducted by using different components alone and the ways to connect the HF components. We report the results on FishNet's labeled data in Table 4.

The results show that both HF and LF entities are important for aquatic species recognition, as both HF and LF only attain relatively good performance. From Table 4, we find that LF alone achieves better performance than that of only using raw images. One reason for this phenomenon may be the LF entity has less data noise, which enhances the noise-robustness of the DNN by neglecting HF components (Li et al., 2020). The results also demonstrate despite the noise-robustness in LF leads to quite high performance, the details information conveyed in the HF entity is critical for aquatic species recognition. Furthermore, we compare the strategy of averaging HF components with strategies such as maximum, addition, and concatenation. As illustrated, the model with an averaging connection for HF components used in WFN achieves better performance.

4.6 Sensitivity analysis of dataset partition

As shown in Table 5, we examine the impact of varying number of labeled and unlabeled data. We set the ratios of labeled data in the training set to 10%, 20%, 30%, and 100%, thereby determining the corresponding ratios of arbitrary unlabeled data. With the entire training dataset labeled (100% labeled data) in supervised learning,

TABLE 2 Ablation study for the wavelet bases in WFN.

Wavelet bases	Common	Medium	Rare	All
Dmey	58.86	44.36	18.48	21.91
Haar	70.84	55.90	25.08	29.13
Coiflets	60.57	46.59	21.44	24.77
Biorthogonal 1.5	58.35	45.43	18.98	22.46
Biorthogonal 2.4	60.18	44.65	19.38	22.74
Daubechies 2	72.73	58.60	25.47	29.80

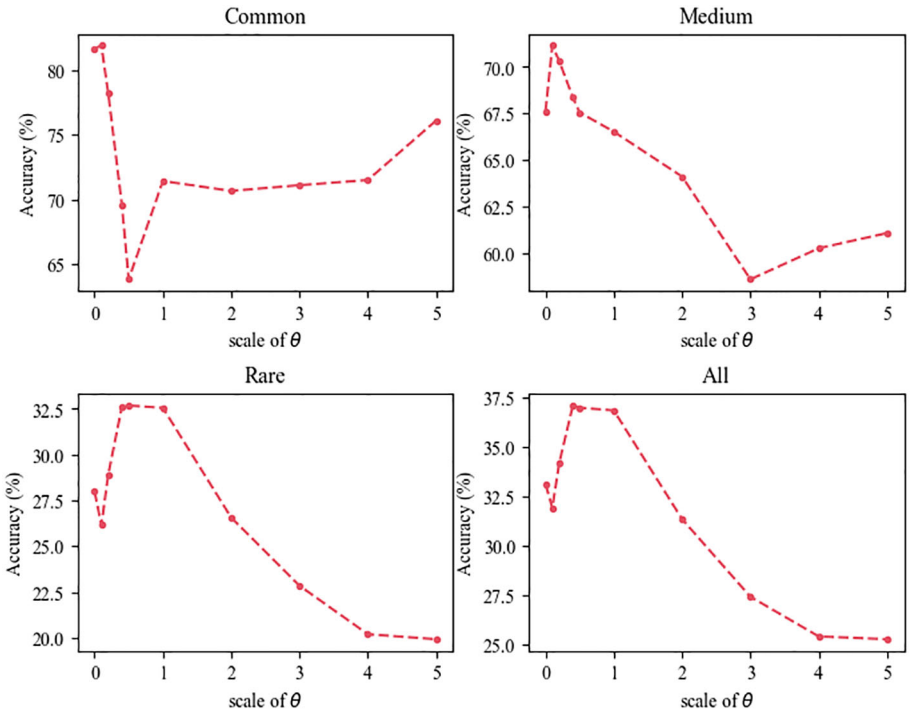

FIGURE 5
Ablation study for the hyper-parameter θ in CEL.

the WFN achieves an average classification accuracy of 49.41% across all aquatic species. Furthermore, the overall average classification accuracy of SSL increases by 8.52%, 7.31%, and 10.06% compared to supervised methods when using 10%, 20%, and 30% labeled data and the remaining unlabeled data. Moreover, our method exhibits improved performance with increasing amounts of unlabeled data. Training with 20% labeled data and 40%, 60%, and 80% unlabeled data result in overall average classification accuracy improvements of 5.97%, 6.82%, and 7.31% over the baseline. The empirical results confirm the proficiency of our method in generating pseudo-labels using arbitrary quantities of labeled data. Additionally, the robustness of the proposed method under diverse conditions has been comprehensively validated.

4.7 Replacing HF entity with edge information

Wavelet transform and edge detectors such as Canny and Sobel serve similar purposes in extracting detailed information within images. To further demonstrate the effectiveness of the HF entity, we replace the HF entity with the information generated by the edge detector. As shown in Table 6, we can see using HF entity outperforms the previous edge detection algorithm by a large margin. To be specific, WFN improves the optimal classification accuracy by over 2.46% in the Canny edge detector, and 1.96% in the Sobel edge detector, respectively. The performance degradation of both experiments illustrates the HF entity extracted by wavelet transform contains rich information about fine details and textures in the image. Besides, Canny and Sobel detectors can be sensitive to noise, especially in low-quality underwater images or those with uneven illumination and complex visual backgrounds, which might lead to false edge detection or noisy information. Through the experiments, we also conclude that WFN has a stronger ability for feature extraction than using raw images with edge information, which can be beneficial for heterogeneous image-collected environments in aquatic species recognition.

4.8 Comparison of model size and computation cost

We showcase the performance of models trained on the labeled images along with model size and computational cost. Given that

TABLE 3 Ablation study for feature fusion strategies in WFN.

Fusion strategy	Common	Medium	Rare	All
concatenate	75.14	58.78	22.88	27.59
Element-wise add	72.68	58.93	23.42	28.05
FusionBlock	72.73	58.60	25.47	29.80

TABLE 4 Analysis of different frequency components.

Raw	LF	HF_{Sum}	HF_{Max}	$HF_{Concate}$	$HF_{Average}$	Common	Medium	Rare	All
✓						70.63	57.65	21.41	26.12
	✓					69.18	56.12	22.52	26.90
		✓				56.25	40.65	14.28	17.78
			✓			45.22	30.88	11.17	13.83
				✓		48.95	29.72	12.85	15.22
					✓	57.19	41.09	14.14	17.72
	✓	✓				69.10	55.48	23.60	27.77
	✓		✓			71.71	56.25	23.16	27.50
	✓			✓		68.90	56.30	24.41	28.57
	✓				✓	72.73	58.60	25.47	29.80

the proposed CEL function is designed for pseudo-labels, its computational complexity is negligible compared to that of fully-supervised training methods. As shown in Table 7, WFN requires two encoders for various frequency awareness, which significantly increased the computation cost as the acquired information increased. Furthermore, to illustrate that the performance enhancement stems from well-designed components, we expand ResNet-50 to match the number of parameters and computational costs of WFN. Our results indicate that while increased computational complexity yields positive effects, it still falls shorts of matching the performance of WFN.

5 Discussion

A previous study (Torney et al., 2019) demonstrates that recognition task typically requiring four ecologists approximately 3 to 6 weeks for manual analysis can be completed in just 24 hours using DL methods. Their research also concludes that this accelerated approach does not compromise accuracy, as abundance estimates obtained through DL were within 1% of those derived from manual analysis by experts. Computer analysis has the potential to substantially streamline the investigative analysis process. Our study concurs with this point but also underscores the significant challenges in data labeling, as evidenced by previous representative studies (Li et al., 2023; Rubbens et al., 2023). This paper introduces a novel technique based on a SSL scheme, where DNN are learned using a limited number of labeled data and extensive unlabeled data, thereby alleviating the burden of manually labeling large dataset for researchers. This is enabled by two simple-to-implement but crucial modifications (1) using a robust feature extraction method, (2) replacing original consistency loss with CEL function. These modifications enable the DL method, trained on a limited amount of labeled data, to effectively address a diverse aquatic environment, as well as the long-tailed distribution of aquatic species.

Aquatic species recognition based on DL serves as a foundation for specific application, particularly biomass estimation and species habitat monitoring (Li et al., 2023). This information is crucial for informed decision-making in conservation management, including the establishment of protected areas, restoration effort, and mitigation of anthropogenic impacts. Furthermore, our research presents promising applications for long-tailed distribution of

TABLE 5 Sensitivity analysis results of dataset partition strategies.

Labeled	Unlabeled	Common	Medium	Rare	All
10%	0%	66.41	47.35	14.97	19.28
	90%	59.78	59.72	23.23	27.80
20%	0%	72.73	58.60	25.47	29.80
	40%	69.27	67.48	31.30	35.77
	60%	67.10	68.23	32.11	36.62
	80%	69.58	68.36	32.61	37.11
30%	0%	75.93	62.35	28.05	32.53
	70%	73.05	71.42	38.43	42.59
100%	0%	83.31	74.49	45.68	49.41

TABLE 6 Ablation on effectiveness of various information, including raw image, LF entity, HF entity, and information generated by edge detector.

Raw	LF	HF	Canny	Sobel	Common	Medium	Rare	All
✓					70.63	57.65	21.41	26.12
✓			✓		67.92	54.55	19.56	24.11
✓				✓	69.66	56.01	20.27	24.93
	✓		✓		70.33	56.34	22.97	27.34
	✓			✓	70.95	56.05	23.58	27.84
	✓	✓			72.73	58.60	25.47	29.80

TABLE 7 Comparison of model sizes and computational cost.

Method	Params (M)	Flops (G)	Common	Medium	Rare	All
ResNet-50	24.46	4.14	70.63	57.65	21.41	26.12
ResNet-50*	59.09	11.63	72.70	57.35	23.09	27.58
WFN	64.11	9.11	72.73	58.60	25.47	29.80

* indicates increasing the number of convolutions and channels.

aquatic species in the natural world, which can significantly contribute to marine biodiversity conservation efforts. Species distribution and abundance follow a highly skewed rule, with a small number of species exhibiting high abundant, while numerous species are present in relatively low numbers (Villon et al., 2022; Saleh et al., 2023). The complex image collection environment poses challenges for commonly used methods such as data augmentation or data generation to be effective, particularly when labeled data is limited. As recommendations for improvement concerning existing conservation measures, we propose integrating our method into established monitoring frameworks. We have conducted both quantitative and qualitative experiments demonstrating the utility of the our method across a variety of diverse aquatic environments using large-scale species recognition datasets. The results of the above experiments instill confidence in our ability to collaborate with existing conservation monitoring programs.

While this work represents progress in developing a robust and effective SSL scheme for real-world aquatic species recognition applications, it has also revealed some limitations that future research should address. Firstly, the CEL enhances the performance of tail-class at the expense of lower performance for head-class. Given the importance of all aquatic species in real environments, it is worthwhile to explore strategies for significantly improving the performance of tail species while maintaining or even enhancing the performance of head species. Secondly, while our study has confirmed the effectiveness of WFN combined with single-level 2D discrete wavelet transform for aquatic species recognition, it is worth developing a DNN equipped with multilevel wavelet packet transform in future research because it could benefit from the hierarchical representation. Lastly, it would be interesting to apply our algorithm to more practical task, such as aquatic species detection, behavior analysis, and trait prediction. By deploying these application in real-world aquatic environments, we can develop increasingly intelligent solutions to address some of the most pressing issues of our time.

6 Conclusion

In this work, we have introduced a robust feature extractor, WFN, and a novel loss function, CEL, based on the SSL scheme FixMatch, for aquatic species recognition. Our proposed methods have demonstrated effectiveness in addressing the challenges of high-quality recognition in complex image-collected environments and the long-tailed class imbalanced nature of aquatic species, even with a limited number of labeled data. This is achieved through dedicated components, using the output of wavelet transform of one to train the DNN, and applying the CEL function at the stage where pseudo-labels come into play. The proposed method has consistently shown performance gains in both quantitative and qualitative experiments. We thus believe that our study can serve as a valuable resource for future research efforts in aquatic species recognition.

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

DM: Writing – original draft. JW: Visualization, Writing – original draft. LZ: Writing – review & editing. FZ: Writing – review

& editing, Supervision. HW: Writing – original draft. XC: Methodology, Writing – review & editing. YL: Funding acquisition, Writing – original draft. ML: 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 by the International Research Center of Big Data for Sustainable Development Goals (No. CBAS2022GSP07), and the National Natural Science Foundation of China (No. 42230505, 42206148).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bell, K. L. C., Quinzin, M. C., Amon, D., Poulton, S., Hope, A., Sarti, O., et al. (2023). Exposing inequities in deep-sea exploration and research: results of the 2022 global deep-sea capacity assessment. *Front. Mar. Sci.* 10, 1217227. doi: 10.3389/fmars.2023.1217227

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., et al. (2019a). Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019b). "Mixmatch: A holistic approach to semi-supervised learning," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS '19)* (NY, USA), 454, 5049–5059. doi: 10.5555/3454287.3454741

Cai, L., McGuire, N. E., Hanlon, R., Mooney, T. A., and Girdhar, Y. (2023). Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. *Int. J. Comput. Vision* 131, 1406–1427. doi: 10.1007/s11263-023-01762-5

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS'19)* (NY, USA). 140, 1567–1578. doi: 10.5555/3454287.3454427

Chen, Z., Du, M., Yang, X.-D., Chen, W., Li, Y.-S., Qian, C., et al. (2023). Deep-learning-based automated tracking and counting of living plankton in natural aquatic environments. *Environ. Sci. Technol.* 57 (46), 18048–18057. doi: 10.1021/acs.est.3c00253

Choi, C., Kampffmeyer, M., Handegard, N. O., Salberg, A.-B., Brautaset, O., Eikvil, L., et al. (2021). Semi-supervised target classification in multi-frequency echosounder data. *ICES J. Mar. Sci.* 78, 2615–2627. doi: 10.1093/icesjms/fsab140

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Long Beach, CA, USA), 9268–9277. doi: 10.1109/CVPR.2019.00949

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Miami, FL, USA), 248–255. doi: 10.1109/CVPR.2009.5206848

de Souza Brito, A., Vieira, M. B., De Andrade, M. L. S. C., Feitosa, R. Q., and Giraldi, G. A. (2021). Combining max-pooling and wavelet pooling strategies for semantic image segmentation. *Expert Syst. Appl.* 183, 115403. doi: 10.1016/j.eswa.2021.115403

Ditria, E. M., Lopez-Marcano, S., Sievers, M., Jinks, E. L., Brown, C. J., and Connolly, R. M. (2020). Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. *Front. Mar. Sci.* 7, 429. doi: 10.3389/fmars.2020.00429

Duan, Y., Liu, F., Jiao, L., Zhao, P., and Zhang, L. (2017). Sar image segmentation based on convolutional-wavelet neural network and markov random field. *Pattern Recognition* 64, 255–267. doi: 10.1016/j.patcog.2016.11.015

Feng, C., Zhong, Y., and Huang, W. (2021). "Exploring classification equilibrium in long-tailed object detection," in *Proceedings of the IEEE/CVF International conference on computer vision (ICCV)* (Montreal, QC, Canada), 3417–3426. doi: 10.1109/ICCV48922.2021.00340

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (Las Vegas, NV, USA), 770–778. doi: 10.1109/CVPR.2016.90

Huang, H., He, R., Sun, Z., and Tan, T. (2017). "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE international conference on computer vision (ICCV)* (Venice, Italy), 1689–1697. doi: 10.1109/ICCV.2017.187

Irfan, S., and Alatawi, A. M. M. (2019). Aquatic ecosystem and biodiversity: a review. *Open J. Ecol.* 9, 1–13. doi: 10.4236/oje.2019.91001

Jahanbakht, M., Azghadi, M. R., and Waltham, N. J. (2023). Semi-supervised and weakly-supervised deep neural networks and dataset for fish detection in turbid underwater videos. *Ecol. Inf.* 78, 102303. doi: 10.1016/j.ecoinf.2023.102303

Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., et al. (2022). Fathomnet: A global image database for enabling artificial intelligence in the ocean. *Sci. Rep.* 12, 15914. doi: 10.1038/s41598-022-19939-2

Kaur, M., and Vijay, S. (2023). Deep learning with invariant feature based species classification in underwater environments. *Multimedia Tools Appl.*, 1–22. doi: 10.1007/s11042-023-15896-8

Khan, F. F., Li, X., Temple, A. J., and Elhoseiny, M. (2023). "Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France), 20496–20506. doi: 10.1109/ICCV51070.2023.01874

Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S. J., and Shin, J. (2020). Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *Adv. Neural Inf. Process. Syst.* 33, 14567–14579.

Lai, Z., Wang, C., Gunawan, H., Cheung, S.-C. S., and Chuah, C.-N. (2022). "Smoothed adaptive weighting for imbalanced semi-supervised learning: Improve reliability against unknown distribution data," in *International Conference on Machine Learning (PMLR)*. 11828–11843.

Laradji, I. H., Saleh, A., Rodriguez, P., Nowrouzezahrai, D., Azghadi, M. R., and Vazquez, D. (2021). Weakly supervised underwater fish segmentation using affinity lcfcn. *Sci. Rep.* 11, 17379. doi: 10.1038/s41598-021-96610-2

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature* 521, 436–444. doi: 10.1038/nature14539

Lee, H., Shin, S., and Kim, H. (2021). Abc: Auxiliary balanced classifier for class-imbalanced semisupervised learning. *Adv. Neural Inf. Process. Syst.* 34, 7082–7094.

Li, Q., Shen, L., Guo, S., and Lai, Z. (2020). "Wavelet integrated cnns for noise-robust image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Seattle, WA, USA), 7245–7254. doi: 10.1109/CVPR42600.2020.00727

Li, Q., Shen, L., Guo, S., and Lai, Z. (2021). Wavecnet: Wavelet integrated cnns to suppress aliasing effect for noise-robust image classification. *IEEE Trans. Image Process.* 30, 7074–7089. doi: 10.1109/TIP.2021.3101395

Li, J., Xu, W., Deng, L., Xiao, Y., Han, Z., and Zheng, H. (2023). Deep learning for visual recognition and detection of aquatic animals: A review. *Rev. Aquaculture* 15, 409–433. doi: 10.1111/raq.12726

Liu, L., Meng, L., Peng, Y., and Wang, X. (2021). A data hiding scheme based on u-net and wavelet transform. *Knowledge-Based Syst.* 223, 107022. doi: 10.1016/j.knosys.2021.107022

Liu, L., Wu, J., Zheng, T., Zhao, H., Kong, H., Qu, B., et al. (2023). Fish recognition in the underwater environment using an improved arcface loss for precision aquaculture. *Fishes* 8, 591. doi: 10.3390/fishes8120591

Liu, P., Zhang, H., Zhang, K., Lin, L., and Zuo, W. (2018). "Multi-level wavelet-cnn for image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (Salt Lake City, UT, USA), 773–782. doi: 10.1109/CVPRW.2018.00121

Lu, H., Li, Y., Uemura, T., Ge, Z., Xu, X., He, L., et al. (2018). Fdcnet: filtering deep convolutional network for marine organism classification. *Multimedia Tools Appl.* 77, 21847–21860. doi: 10.1007/s11042-017-4585-1

Ma, D., Wei, J., Li, Y., Zhao, F., Chen, X., Hu, Y., et al. (2023). Mldet: Towards efficient and accurate deep learning method for marine litter detection. *Ocean Coast. Manage.* 243, 106765. doi: 10.1016/j.ocecoaman.2023.106765

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 674–693. doi: 10.1109/34.192463

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. (2020). Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1979–1993. doi: 10.1109/TPAMI.34

Moller, T., Nilssen, I., and Nattkemper, T. W. (2017). "Active learning for the classification of species in underwater images from a fixed observatory," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)* (Venice, Italy), 2891–2897. doi: 10.1109/ICCVW.2017.341

Oh, Y., Kim, D.-J., and Kweon, I. S. (2022). "Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA, USA), 9786–9796. doi: 10.1109/CVPR52688.2022.00956

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32.

Qiu, C., Zhang, S., Wang, C., Yu, Z., Zheng, H., and Zheng, B. (2018). Improving transfer learning and squeeze-and-excitation networks for small-scale fine-grained fish image classification. *IEEE Access* 6, 78503–78512. doi: 10.1109/Access.6287639

Ramamonjisoa, M., Firman, M., Watson, J., Lepetit, V., and Turmukhambetov, D. (2021). "Single image depth prediction with wavelet decomposition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Nashville, TN, USA), 11089–11098. doi: 10.1109/CVPR46437.2021.01094

Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al. (2020). "Balanced meta-softmax for long-tailed visual recognition." in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (NY, USA), 4175–4186. doi: 10.5555/3495724.3496075

Rubbens, P., Brodie, S., Cordier, T., Destro Barcellos, D., Devos, P., Fernandes-Salvador, J. A., et al. (2023). Machine learning in marine ecology: an overview of techniques and applications. *ICES J. Mar. Sci.* 80, 1829–1853. doi: 10.1093/icesjms/fsad100

Sala, E., Mayorga, J., Bradley, D., Cabral, R. B., Atwood, T. B., Auber, A., et al. (2021). Protecting the global ocean for biodiversity, food and climate. *Nature* 592, 397–402. doi: 10.1038/s41586-021-03371-z

Saleh, A., Laradji, I. H., Konovalov, D. A., Bradley, M., Vazquez, D., and Sheaves, M. (2020). A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Sci. Rep.* 10, 14671. doi: 10.1038/s41598-020-71639-x

Saleh, A., Sheaves, M., Jerry, D., and Azghadi, M. R. (2023). Applications of deep learning in fish habitat monitoring: A tutorial and survey. *Expert Syst. Appl.*, 121841.

Saleh, A., Sheaves, M., and Rahimi Azghadi, M. (2022). Computer vision and deep learning for fish classification in underwater habitats: A survey. *Fish Fisheries* 23, 977–999. doi: 10.1111/faf.12666

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., et al. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* 33, 596–608.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., et al. (2020). "Equalization loss for long-tailed object recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Seattle, WA, USA), 11662–11671. doi: 10.1109/CVPR42600.2020.01168

Torney, C. J., Lloyd-Jones, D. J., Chevallier, M., Moyer, D. C., Maliti, H. T., Mwita, M., et al. (2019). A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods Ecol. Evol.* 10, 779–787. doi: 10.1111/2041-210X.13165

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (NY, USA), 6000–6010. doi: 10.5555/3295222.3295349

Villon, S., Iovan, C., Mangeas, M., and Vigliola, L. (2022). Confronting deep-learning and biodiversity challenges for automatic video-monitoring of marine ecosystems. *Sensors* 22, 497. doi: 10.3390/s22020497

Visbeck, M. (2018). Ocean science research is key for a sustainable future. *Nat. Commun.* 9, 690. doi: 10.1038/s41467-018-03158-3

Wei, C., Sohn, K., Mellina, C., Yuille, A., and Yang, F. (2021). "Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Nashville, TN, USA), 10857–10866. doi: 10.1109/CVPR46437.2021.01071

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020a). "Unsupervised data augmentation for consistency training," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)* (NY, USA), 6256–6268. doi: 10.5555/3495724.3496249

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020b). "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Seattle, WA, USA), 10687–10698. doi: 10.1109/CVPR42600.2020

Yang, X., Song, Z., King, I., and Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Trans. Knowledge Data Eng.* 35 (9), 8934–8954. doi: 10.1109/TKDE.2022.3220219

Yao, T., Pan, Y., Li, Y., Ngo, C.-W., and Mei, T. (2022). "Wave-vit: Unifying wavelet and transformers for visual representation learning," in *European Conference on Computer Vision (ECCV)* (Berlin, Heidelberg), 328–345. doi: 10.1007/978-3-031-19806-9_19

Yin, X., and Xu, X. (2021). "A method for improving accuracy of deeplabv3+ semantic segmentation model based on wavelet transform," in *International Conference in Communications, Signal Processing, and Systems* (Singapore), 315–320. doi: 10.1007/978-981-19-0390-8_85

Zhang, Y., Kang, B., Hooi, B., Yan, S., and Feng, J. (2023). Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 10795–10816. doi: 10.1109/TPAMI.2023.3268118

Zhao, Y., Wang, S., Zhang, Y., Qiao, S., and Zhang, M. (2023). Wranet: wavelet integrated residual attention u-net network for medical image segmentation. *Complex intelligent Syst.* 9, 6971–6983. doi: 10.1007/s40747-023-01119-y

Zhou, Y., Huang, J., Wang, C., Song, L., and Yang, G. (2023). "Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris, France), 21085–21096. doi: 10.1109/ICCV51070.2023.01928

Zhuang, P., Wang, Y., and Qiao, Y. (2020). Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Trans. Multimedia* 23, 3603–3617. doi: 10.1109/TMM.2020.3028482