frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 16 March 2023
DOI 10.3389/fmars.2023.1149186

OPEN ACCESS

EDITED BY
Hongsheng Bi,
University of Maryland, College Park,
United States

REVIEWED BY
Zhineng Chen,
Fudan University, China
Suja Cherukullapurath Mana,
Sathyabama Institute of Science and
Technology, India

*CORRESPONDENCE
Xiaodong Wang
wangxiaodong@ouc.edu.cn

SPECIALTY SECTION
This article was submitted to
Ocean Observation,
a section of the journal
Frontiers in Marine Science

RECEIVED 21 January 2023
ACCEPTED 03 March 2023
PUBLISHED 16 March 2023

CITATION
Zhai J, Han L, Xiao Y, Yan M, Wang Y
and Wang X (2023) Few-shot fine-
grained fish species classification *via*
sandwich attention CovaMNet.
Front. Mar. Sci. 10:1149186.
doi: 10.3389/fmars.2023.1149186

Few-shot fine-grained fish species classification *via* sandwich attention CovaMNet

Jiping Zhai[1], Lu Han[1], Ying Xiao[2], Mai Yan[1],
Yueyue Wang[3] and Xiaodong Wang[4]*

[1]College of Electronic Engineering, Ocean University of China, Qingdao, Shandong, China, [2]School of Science, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong SAR, China, [3]Computing Center, Ocean University of China, Qingdao, Shandong, China, [4]College of Computer Science and Technology, Ocean University of China, Qingdao, Shandong, China

The task of accurately classifying marine fish species is of great importance to marine ecosystem investigations, but previously used methods were extremely labor-intensive. Computer vision approaches have the advantages of being long-term, non-destructive, non-contact and low-cost, making them ideal for this task. Due to the unique nature of the marine environment, marine fish data is difficult to collect and often of poor quality, and learning how to identify additional categories from a small sample of images is a very difficult task, meanwhile fish classification is also a fine-grained problem. Most of the existing solutions dealing with few-shot classification mainly focus on the improvement of the metric-based approaches. For few-shot classification tasks, the features extracted by CNN are sufficient for the metric-based model to make a decision, while for few-shot fine-grained classification with small inter-class differences, the CNN features might be insufficient and feature enhancement is essential. This paper proposes a novel attention network named Sandwich Attention Covariance Metric Network (SACovaMNet), which adds a new sandwich-shaped attention module to the CovaMNet based on metric learning, strengthening the CNN's ability to perform feature extraction on few-shot fine-grained fish images in a more detailed and comprehensive manner. This new model can not only capture the classification objects from the global perspective, but also extract the local subtle differences. By solving the problem of feature enhancement, this new model can accurately classify few-shot fine-grained marine fish images. Experiments demonstrate that this method outperforms state-of-the-art solutions on few-shot fine-grained fish species classification.

KEYWORDS

fish species classification, computer vision, few-shot learning, fine-grained image classification, sandwich attention

1 Introduction

Fish species classification is critical to industry and food production as well as conservation and management of marine fisheries. However, most marine fish classification solutions still require manual classification by humans (Alsmadi et al., 2019). As fish classification is a fine-grained problem, the manual classification process is time-consuming and requires a lot of labor and material resources. Due to the dynamic changes of the marine environment, the requirements for shooting equipment are high, which means that the number of underwater images we can obtain is small. Therefore, few-shot fine-grained fish species classification has become a difficult problem to solve. At the same time, due to the absorption and scattering of light in seawater (McGlamery, 1980), as well as other impurities in seawater, most of the collected underwater fish data have poor image quality and complex background problems, which makes the task of few-shot fine-grained fish species classification even more difficult. With the rapid development of computer vision, more and more deep learning methods have appeared in our production, life and work, so the classification of marine species based on deep learning is very necessary (Zhao et al., 2021; Alsmadi and Almarashdeh, 2022; Li et al., 2022).

Few-shot learning is an emerging but important method which attempts to learn new categories from a few labeled examples (Hou et al., 2019). Commonly used methods to solve few-shot image classification mainly include transfer learning (Luo et al., 2017; Peng et al., 2019), meta-learning (Finn et al., 2017; Ren et al., 2018; Lee et al., 2019; He et al., 2023) and metric learning (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018; Li et al., 2023). The first two categories focus on finding a suitable initialization parameter model for few-shot learning networks, then using prior knowledge extracted from other tasks to prevent overfitting and improve generalization capabilities. And the last category pays attention to finding a superior similarity metric function to replace the fully connected classification layer with a large amount of parameters, where most existing methods use Euclidean distance and cosine similarity as metric function to classify images. Methods based on metric learning have achieved state-of-the-art performance in the few-shot classification field due to the strong ability of discrimination. Most of the current few-shot image classification methods focus on common classification tasks, that is, the features between categories have obvious differences. However, for fish images, the difference between sample image categories is small, which obviously makes this a fine-grained image classification problem (Zhao et al., 2021), and unfortunately the above classification methods do not take into account the difficulties raised by fine-grained classification.

For few-shot fine-grained image classification, most of the currently available methods take one of two approaches, they either attempt to make the network with a more advanced feature vector measurement module (Vinyals et al., 2016; Sung et al., 2018; Li et al., 2019) or they rely on feature reconstruction (Zhang et al., 2020; Wertheimer et al., 2021). However, they ignore the issue where fine-grained images have much higher requirements for the capabilities of feature extraction modules than general classification methods. Since the images have similar global features in different categories of fine-grained images, and only have significant differences in some subtle features, the extracted feature vectors also have a certain degree of similarity (Wei et al., 2021), which puts too much pressure on the feature measurement module. Due to the small number of samples, few-shot learning is prone to overfitting (Chen et al., 2019), and using a large feature extraction module is not a perfect solution, but through extensive research, the Attention Mechanism (AM) has been used in underwater image enhancement and underwater image dehazing (Shi et al., 2022; Liu P. et al., 2022), it was concluded that an AM may be a better solution for few-shot fish image classification.

Considering the above problems, this paper proposes a novel AM network, named Sandwich Attention CovaMNet (SACovaMNet for short), which can effectively solve the classification problem of few-shot fine-grained fish images, and enable the CNN to more carefully and comprehensively classify marine fish. This new SACovaMNet enables the CNN to extract features from fine-grained images of marine fish in a more detailed and comprehensive manner, capturing recognition objects globally as well as extracting nuances between classes of fish samples locally, thus improving classification accuracy. The main contributions of this work are summarized as follows: 1) To solve the few-shot fine-grained fish species classification problem caused by the small number of fish images and minor differences between classes, we carefully designed a Sandwich Attention module that combines local attention and global attention on the basis of the few-shot model CovaMNet to build our SACovaMNet, which enables CovaMNet to more comprehensively extract features from fine-grained images of marine fish and expand the distance between prototype feature vectors of different categories; 2) Aiming at the problem of missing feature information in the fine-grained image of the CBAM, we improved the CBAM module so that it can weigh the feature map more completely; 3) Exhaustive experiments were conducted based on three fine-grained datasets of marine fish organisms, and experimental results demonstrate that the proposed method outperforms the state-of-the-art solutions.

The rest of this paper is as follows. Section 2 is a review of the related works for few-shot fine-grained image classification. Section 3 introduces the proposed method SACovaMNet. And Section 4 shows the experimental results. Finally, a conclusion is made in Section 5.

2 Related work

Deep learning performs very well when the amount of training data is large, but conversely training the network to perform better becomes problematic when the amount of training data is small. In recent years, few-shot learning (Chen et al., 2019) has been proposed to solve this problem. It was found that few-shot learning is better for the problem of classifying marine fish with sparse samples, and a brief review of the relevant aspects of the problem-solving approach will be given.

2.1 Fish species classification

The fish species classification task is different from general classification tasks, it is a typical fine-grained classification task (Zhao et al., 2021). In recent years, many methods for fish species classification have been proposed, and fish classification models based on biological characteristics (Kartika and Herumurti, 2016; Tharwat et al., 2018) and deep learning models (Chen et al., 2017; Zhao et al., 2021) are more popular. Kartika and Herumurti (2016) proposed a K-means segmentation background and HSV color space feature extraction method, which effectively extracted the color features of koi carp, and finally adopted NBM and SVM methods for identification and classification. Tharwat et al. (2018) took a different approach, using the fusion of Weber Local Descriptor (WLD) features and color features, and also used the LDA algorithm to reduce the dimension of the feature vector and increase the discrimination between different categories (fish species), and finally used the AdaBoost classifier for classification. Unfortunately, methods based on biometric feature extraction cannot handle complex backgrounds or a large number of images, however, deep learning can better solve this problem and achieve more accurate classification results. Rathi et al. (2017) performed classification by pre-processing images using Gaussian blur, morphological operations, Otsu's thresholding, and pyramid mean translation, and further fed the enhanced images to a convolutional neural network for classification. Prasetyo et al. (2022) proposed Multi-Level Residual (MLR) as a new residual network strategy by combining the low-level features of the initial block with the high-level features of the last block using Depthwise Separable Convolution (DSC). They used VGGNet as the backbone of the new CNN architecture by removing the fifth block and replacing it with components such as MLR, Asymmetric Convolution (AC), Batch Normalization (BN), and residual features.

Unfortunately, in reality, due to the complexity of the underwater environment (Shevchenko et al., 2018), it is impossible to obtain enough samples for traditional deep learning training. Guo et al. (2020) believed that the classic CNN model required a large amount of high-quality data to obtain excellent results. For few-shot fish images, it is difficult to obtain data diversity through image augmentation, so a generative network is used to generate realistic fake images with a small amount of training data, and the classification accuracy can be improved by making the datasets diverse and rich. However, the training method based on the generative network is complicated, so the proposed method considers building a few-shot learning method to solve this problem.

2.2 Few-shot learning

2.2.1 Meta-learning

Meta-learning (Hochreiter et al., 2001) is, as the name suggests, learning to learn; the algorithm sets up a meta-learner component and a task-specific learner component, with the training unit being the task, allowing information to cross between tasks. Meta-learning is a popular approach to tackle few-shot problems. MAML (Finn et al., 2017) proposed an algorithm for meta-learning that is model-agnostic, and trained a model's parameters such that a small number of gradient updates will lead to rapid learning on a new task. Reptile (Nichol et al., 2018) removed the re-initialization of each task in order to simplify the update process for MAML, making it a more natural choice in some settings. LEO (Rusu et al., 2019) learnt a low-dimensional latent embedding of model parameters and performed optimization-based meta-learning in this space. While meta-learning has had some success with few-shot problems, it is difficult to train due to its use of complex memory addressing structures (Li et al., 2019), therefore the proposed approach utilizes only a single CNN framework baseline which can be end-to-end trained from scratch.

2.2.2 Transfer learning

Transfer learning (Zhuang et al., 2021) is to transfer the learned model parameters from one model to a new model or task in order to achieve better training results. For datasets with fewer samples, first the model is trained on a dataset with a large number of similar data domains, and then fine-tuned, usually with good results. Compared with the complex training mode of meta-learning, transfer learning can perform simple end-to-end training. Luo et al. (2017) proposed a framework to learn representations that are transferable across different domains and tasks in a label-efficient manner. This method combats domain shift with a domain-adversarial loss and uses a metric learning-based method to generalize embeddings to new tasks. Peng et al. (2019) used the graph convolutional neural network to construct a mapping network between semantic knowledge and visual features, combined image features and semantic features through the fusion of classifier weights, and supplemented semantic features as *a priori* knowledge to a few-shot classifier.

2.2.3 Metric learning

Metric-based learning methods learn a set of item functions (embedding functions) and metrics to measure the similarity between query and sample images and classify them in a feed-forward manner. The main difference between metric-based learning methods is how they learn the metrics, hence metric learning is often referred to as similarity learning (Li et al., 2020). Matching Networks (Vinyals et al., 2016) constructed an end-to-end network architecture that uses cosine similarity to calculate distances. After training, the matching network was able to generate reasonable test labels for unobserved categories without any fine-tuning of the network. In contrast, Prototypical Networks (Snell et al., 2017) mapped the sample data in each category into a space and extracted their means to represent them as protoforms of that class, using Euclidean distance as the distance metric, they are trained so that protoforms of the same class are represented as the closest distance and that inter-class protoforms are represented as the farther distance.

2.3 Fine-grained image classification

2.3.1 Fine-grained image classification

Fine-grained image classification aims to distinguish subcategories, such as birds or dog breeds. Fish image classification also belongs to fine-grained image classification. Compared with general classification tasks, fine-grained image classification is challenging due to high intra-class and low inter-class variance (Zhao et al., 2017). Zhang et al. (2014) proposed a model utilizing deep convolutional features computed on bottom-up region proposals, which learns whole-object and part detectors, enforces learned geometric constraints between them, and predicts a fine-grained category from a pose-normalized representation. Li et al. (2021) proposed a so-called Bi-Similarity Network (BSNet) that consists of a single embedding module and a bi-similarity module of two similarity measures. After the support images and the query images pass through the convolution-based embedding module, the bi-similarity module learns feature maps according to two similarity measures of diverse characteristics.

2.3.2 Few-shot fine-grained image classification

With the development of deep learning, fine-grained image classification has achieved remarkable achievements, but largely relies on a large number of labeled samples. However, in practical applications in some fields, it is difficult to obtain such a large amount of labeled fine-grained data. Therefore, few-shot fine-grained images classification is getting more and more attention (Liu Y. et al., 2022). CovaMNet (Li et al., 2019) proposed a deep covariance metric to measure the consistency of distributions between query samples and new concepts, and used the second-order statistics of concept representation and verified that it is more suitable to represent a concept beyond the first-order statistics, it can naturally capture the underlying distribution information of each concept (or category). Wertheimer et al. (2021) introduced a novel mechanism by regressing directly from support features to query features in closed form, without introducing any new modules or large-scale learnable parameters. Lee et al. (2022) proposed Task Discrepancy Maximization (TDM), which is a feature alignment method, to define the class-wise channel importance, and to localize the class-wise discriminative regions by highlighting channels encoding distinct information of the class. The AM can be used to make the feature vector reweight once before entering the measurement module to ensure that the feature vector pays more attention to the differences between categories, so as to solve the problem of small differences between few-shot fine-grained image samples.

2.3.3 Attention mechanism

Transformer (Vaswani et al., 2017) first achieved excellent results in natural language processing (NLP), and then researchers applied it to the field of vision (Vision Transformer, ViT) (Dosovitskiy et al., 2021; Guo et al., 2022). Dosovitskiy et al. (2021) is believed that the biggest reason for the promising results of Vision Transformer is that it uses a Multi-Headed Self-Attentive (MHSA) module and thus introduces a global attention mechanism, which has powerful representation capabilities. However, due to the image processing method of Vision Transformer, the training time and inference speed will increase quadratically when processing large scale images. To solve this problem, Srinivas et al. (2021) proposed a botnet combining CNN and transformer, in which the 3×3 convolutional layers in the bottleneck are replaced with MHSA, making the botnet achieve state-of-the-art in classification, target detection and segmentation, whilst the training time and inference speed were significantly reduced relative to (Dosovitskiy et al., 2021).

2.4 Comparison to our approach

Compared with other meta-learning based few-shot classification methods, our method SACovaMNet adopts the metric learning architecture and is based on a simple CNN network construction, which can be trained easily in an end-to-end manner from scratch. We use a second-order measurement algorithm that can compare the similarity in more detail, which improves the feature measurement capability of fine-grained images compared to other first-order metric methods. Additionally, our self-designed Sandwich Attention module strengthens the feature extraction ability of our method for fine-grained images, making our method more suitable for the few-shot fine-grained fish species classification.

3 Methodology

The proposed method utilizes episodic training as the training method, as many researchers have demonstrated it to be simple and effective for few-shot problems (Li et al., 2019). The model structure is shown in Figure 1. After the support images and the query images pass through the weight-sharing feature extraction module at the same time to obtain the feature map, the feature map then passes through the Sandwich Attention module to finally obtain the $H \times W \times C$ feature map. The measurement module uses the second-order covariance metric to measure the correlation between query features and support features.

3.1 Baseline

Various metric-based networks have achieved excellent performance in recent few-shot learning studies (Li et al., 2020). Most of the current metric learning algorithms are first-order metric methods such as Euclidean distance or cosine similarity distance. Generally speaking, before the feature map enters these measurement modules, the dimensions of the feature map need to be reduced. Obviously, there will be a large information loss due to this process, especially the spatial information of the feature map. For fish samples especially captured in situ, since the difference between categories is very small, it is very easy to lose key information in pooling and dimensional reduction, so the above approach is unacceptable in fine-grained fish image classification.

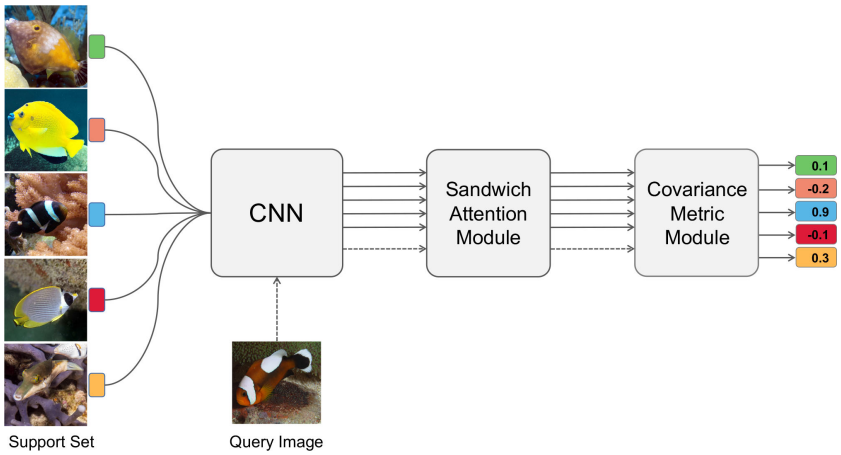

FIGURE 1
Architecture of the proposed SACovaMNet model. The support images and the query image are simultaneously passed through a weight-sharing CNN network to obtain the feature map, and the output feature map is then fed into our Sandwich Attention module to produce the feature map ($H \times W \times C$), which is finally passed through the second-order covariance metric module for similarity calculation.

Recently, (Li et al. 2019) proposed a method based on the second-order local covariance metric. The covariance matrix is the original second-order statistic of the sample set. Since the number of images in each category is very small under the few-shot settings, it is impossible to accurately learn the covariance matrix to describe the data distribution. So the baseline introduces local covariance, expressed as follows:

$$\Sigma_c^{local} = \frac{1}{MK-1}\sum_{i=1}^{K}(\boldsymbol{X}_i - \boldsymbol{\tau})(\boldsymbol{X}_i - \boldsymbol{\tau})^{\top}, \quad (1)$$

where Σ_c^{local} represents the local covariance representation of the c-th class, K is the total number of samples of the c-th class, usually is set as 1 or 5, and X_i is the input sample image, M represents the M local depths of the sample, and τ is an average vector matrix.

The covariance measure is to measure the relationship between a sample and a category, and the measure function named Covariance Metric is as follows:

$$f(\boldsymbol{x}, \Sigma) = \boldsymbol{x}^{\top}\Sigma x\,. \quad (2)$$

The above mentioned Covariance Metric can directly describe the underlying distribution of a concept, and it can fully take into account the local similarity information of the feature map. Since the fish images are fine-grained dataset, and one of the key issues for the classification is to distinguish the local subtle differences between fish categories so as to achieve the more accurate classification. The proposed method has opted to use CovaMNet (Li et al., 2019) which has achieved promisingresults in a series of experimental settings meeting the requirements.

The whole network framework is simple and compact due to it being based on a single end-to-end CNN, a local covariance representation to represent the underlying distribution of each category, and a new covariance metric that is embedded into the network to measure the relationship between query images and categories. The 5-way 1-shot and 5-way 5-shot episodic training mechanism are considered to measure the few-shot classification method under different few-shot situations.

3.2 Sandwich attention

Although the baseline solves some problems in fish classification to a certain extent, the measurement method can only solve the issues in the process of comparing the similarity of feature maps. However, by analyzing the fish image datasets, it was found that most of the images collected in real time cannot correctly reflect the feature information of fish samples due to a variety of problems. In the face of complex fish images, it is expected that feature maps will better reflect the differences between different categories, thereby improving the accuracy of classification, so it was decided to leverage the attention, with a novel attention module designed as shown in Figure 2.

Firstly, in most fish images, the object to be classified is usually only part of the whole image, and there is a lot of interference from the background and other creatures on the seabed, which is also reflected in the feature map extracted by the backbone, making the

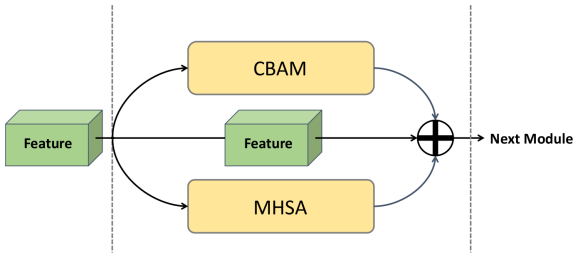

FIGURE 2
Architecture of the proposed attention mechanism, which has been named Sandwich Attention due to it being shaped like a sandwich.

feature map full of useless spatial information. If a manual process was used to increase the proportion of objects identified by manual culling, this would increase the human and financial investment. Therefore it is believed that spatial attention is the most "cost effective" approach to this problem. To this end, a Convolutional Block Attention Module (CBAM) (Woo et al., 2018) module was added to the network, so that the network can correctly locate the position and key feature information of the recognized categories. There are two main tandem sub-modules in CBAM, the channel attention module and the spatial attention module, which perform channel and spatial attention respectively.

In the channel attention module in Figure 3A, the input feature map F ($H\times W\times C$) is subjected to global max pooling and global average pooling to obtain two $1\times1\times C$ feature maps, which are then fed into a two-layer neural network (MLP). Then, the features output by MLP are summed based on element-wise, and activated by sigmoid to generate the final channel attention feature. In the spatial attention module in Figure 3A, the output channel attention and the input feature map F are multiplied element-wise to generate the input of spatial attention module. Next, channel-based global maximum pooling and global average pooling are performed, and then the two feature maps are channel-based splicing operations, one $H\times W\times1$ feature map is obtained through a convolution operation. Finally, the spatial attention feature is generated through the sigmoid function.

At the same time, as fish images are inherently fine-grained, and the difficulty with fine-grained image classification is that the differences between recognized objects are very small and only vary in subtle ways, so the difficulty lies in making the network more accurate in classifying fine-grained images in a few-shot setting. With the rise of ViT in recent years, it is believed that the biggest reason for the promising results achieved by Vision Transformer is because of its powerful representation capabilities using a Multi-Headed Self-Attention module (MHSA) and introducing a global attention mechanism. In Srinivas et al. (2021), the proposed MHSA also introduces Relative Position Encodings, as shown in Figure 3B, thus taking into account the relative distances between features at different locations and being able to effectively relate cross-object

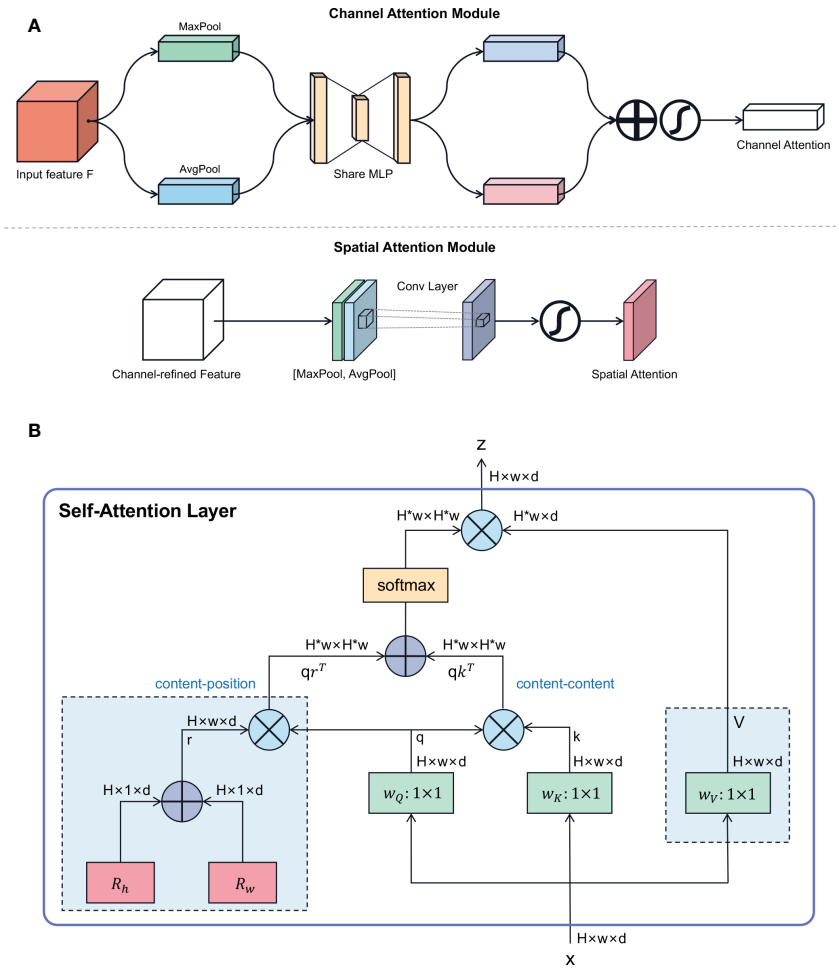

FIGURE 3
The details of AM modules employed in our SACovaMNet model. **(A)** Schematic diagram of each attention sub-module of CBAM (Woo et al., 2018). **(B)** Network structure of multi-head self-attention (MHSA) (Srinivas et al., 2021).

information to location awareness, so this attention mechanism is used in the proposed model.

Based on the above thinking, both MHSA and CBAM were fused into the proposed network. To demonstrate that this approach works and the use of the attentional connectivity, ablation experiments were also conducted in Section 4.4. The final network is based on a simple end-to-end framework using a single CNN with a compact training simple network structure, and the experimental results are presented in Section 4.3.

3.3 Improved CBAM

Although the new model can achieve promising classification results on few-shot fish datasets, fish classification is more difficult due to the difference between fish datasets and general datasets, so it is believed that while CBAM can be applied to fish classification it is still not a perfect solution. More specifically, it is thought that the application of CBAM in fish species classification still has the following problems: 1) The channel attention of CBAM uses global pooling to process the feature map, which obviously does not take into account the importance of different spatial regions of the feature map, resulting in a deviation in the weight calculation of the channel, which is very important for classification, especially that, the difficult fish classification task will obviously have a greater impact; 2) The CBAM uses the feature map of channel attention after global average pooling and maximum pooling to calculate the channel weight through weight-sharing MLP, obviously, there are some differences in the feature map information saved by these two different pooling methods, and using the same MLP cannot fully mine all the information it contains.

Based on the above considerations, we improved the channel attention module of the CBAM module, as shown in Figure 4, both adaptive average pooling and maximum pooling were performed on the feature map (64×21×21) output by the CNN, and it was divided into 7×7 spatial areas, then the MLP module was removed from the CBAM, and two small CNN networks were employed to perform weight calculations respectively, in which the convolution kernel of the first layer of CNN has a large receptive field convolution kernel of 7×7, the second layer is a CNN for dimensionality reduction, the third layer is a Rectified Linear Unit (ReLU) activation function, and the fourth layer is a CNN for dimensionality increase, so we call it DualPath Channel Attention CBAM (DPCACBAM). The importance of different regions of the feature map is calculated not only to ensure that the contribution of different spatial regions of the feature map can be comprehensively considered in the channel attention, but also to fully mine the hidden information in the feature map.

4 Experiments

In this section, extensive experiments were conducted on three fish datasets under corresponding few-shot settings to evaluate the proposed SACovaMNet.

4.1 Datasets

4.1.1 WildFish

This dataset was first proposed in Zhuang et al. (2018), which is a large-scale benchmark dataset for wild fish identification. And it is the largest wild fish recognition image dataset, which contains 1000 fish categories and 54,459 unconstrained images, according to our statistics, the number of images per category varies between 30 and 167. In this work, we randomly split the dataset by categories, where 550, 150, and 300 categories are used for training, validation, and testing, respectively.

4.1.2 Fishclassifierfinal

This dataset is a dataset on the Kaggle website[1], which contains 30 kinds of fish. The dataset has been divided into a train set and a test set. We merge the images of the same fish, and the number of fish images in each category is about 300. We randomly split the dataset by category, where 17, 6, and 7 categories are used for training, validation, and testing, respectively.

4.1.3 QUT fish dataset

This dataset is a dataset also published on the Kaggle website (Anantharajah, 2014), which contains about 4,000 images of 468 fish species. After we classify the given raw images, according to our statistics, the number of each category is between 3 and 26. In this paper, we randomly split this dataset by the number of categories, where 280, 80, and 123 categories are used for training, validation, and testing, respectively.

4.2 Experimental settings

The 5-way 1-shot and 5-way 5-shot classification experiments were conducted on WildFish and fishclassifierfinal datasets. During the training process, episodic training mechanism was used to learn the model parameters, and a total of 250,000 episodes were trained. Each episode contained a query set and a support set. For the 5-way 1-shot classification task, 5 different categories of images were required. Each category of images needed 1 support image and 15 query images. For the 5-way 5-shot classification task, 5 different categories of images were required, and each category of images needed 5 support images and 15 query images. The optimization algorithm Adam (Kingma and Ba, 2014) was used, the initial learning rate was set to 0.0001, and every 10,000 episodes the learning rate would be reduced. During the testing process, 600 episodes were randomly constructed from the testing set, and the top-1 mean accuracy and 95% confidence intervals (model's skill having a 95% probability to correctly generalize) were calculated. Note that the proposed SACovaMNet model was trained from scratch in an end-to-end manner and did not require fine-tuning.

1 https://www.kaggle.com/datasets/khaledelsayedibrahim/fishclassifierfinal

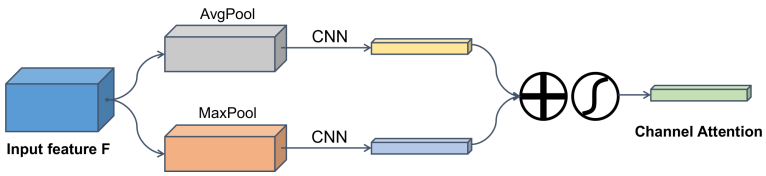

FIGURE 4
Architecture of the proposed DPCACBAM. The input feature map is subjected to local average pooling and maximum pooling, and then the features obtained after passing through two CNN networks are summed element-wise, and finally a channel attentionfeature is generated through a sigmoid.

For QUT fish dataset, due to the small sample size, only the 5-way 1-shot classification experiment was conducted. In the episodic training mechanism, in each category of each episode, there was 1 support image and 2 query images. Other experimental settings remained unchanged.

In order to evaluate the performance of our model on the fish datasets, a selection of state-of-the-art methods commonly used in few-shot fine-grained images were considered for comparison, including baseline CovaMNet (Li et al., 2019), Matching Nets (Vinyals et al., 2016), Prototypical Nets (Snell et al., 2017), MAML (Finn et al., 2017), FRN (Wertheimer et al., 2021), and TDM (Lee et al., 2022). MAML and FRN use the method of meta-learning, Matching Nets, Prototypical Nets and CovaMNet use the method of metric learning, and TDM uses a transferable attention module. We use the TDM method with both FRN and Prototypical Net. For these comparative models, their experimental setup followed the settings from their original work. The SACovaMNet model employed a four-layer convolutional network with a kernel size 64 of each convolutional layer as an embedding module.

4.3 Comparison with state-of-the-arts

The experimental results are shown in Table 1, where, the second column indicates whether the method needs to be fine-tuned; the third and the fourth columns indicate the 5-way 1-shot and the 5-way 5-shot classification accuracies on the WildFish dataset, with 95% confidence intervals; the fifth and the sixth columns represent the 5-way 1-shot and the 5-way 5-shot classification accuracies on the fishclassifierfinal dataset, with 95% confidence intervals; the seventh column represent the 5-way 1-shot classification accuracies on the QUT fish dataset, with 95% confidence intervals. SACovaMNet indicates the method proposed in Section 3.2, and SACovaMNet* indicates the method proposed in Section 3.3. From Table 1, it can be seen that the baseline is more suitable for the fish datasets than other methods, which appears to prove that it was the correct choice for the baseline method to utilize the second-order covariance metric measure. Experimental results have shown that the proposed method outperforms state-of-the-art methods with higher accuracies in all

TABLE 1 The 5-way 1-shot and the 5-way 5-shot classification accuracies on the three datasets, *i.e.*, WildFish, fishclassifierfinal, and QUT fish dataset, with 95% confidence intervals.

Model	Fine-tuning	5-Way Accuracy(%)				
		WildFish		fishclassifierfinal		QUT fish dataset
		1-shot	5-shot	1-shot	5-shot	1-shot
Matching Nets (2016)	N	49.37	56.76	39.84	43.64	60.40
Prototypical Nets (2017)	N	49.81	79.87	51.55	75.49	67.11
MAML (2017)	Y	61.93	76.40	47.73	64.45	74.06
CovaMNet (2019)	N	70.87	84.33	54.54	68.52	66.86
FRN (2021)	N	64.12	80.81	45.42	66.41	61.05
FRN+TDM (2022)	N	43.71	81.66	41.92	69.03	37.03
ProtoNet+TDM (2022)	N	60.23	78.79	52.51	73.03	61.05
SACovaMNet	N	71.44	85.88	58.89	69.01	68.85
SACovaMNet*	N	72.68	86.12	59.28	73.82	70.52

cases. Matching Nets and Prototypical Nets are the earliest few-shot learning methods, and the network structure is simple, so the performance in few-shot fine-grained image classification is not satisfactory; and MAML uses a strategy of meta-learning and fine-tuning, so the effect has been improved. CovaMNet does not adopt the common first-order metric, but uses the second-order metric method, because the details of fine-grained images are preserved, resulting in higher accuracy. FRN achieves better classification results by reconstructing the feature space. The effect of TDM on FRN is not as good as that on Prototypical Nets. This is because FRN itself has more parameters than Prototypical Nets. After adding TDM, overfitting occurs when the number of samples is set to be very small, resulting in unsatisfactory results. Compared to the meta-learning-based MAML that needs to be fine-tuned, our method not only has a simple network structure, but also has a simple training process and short training time, additionally in this case it also achieves high accuracy. And the recent TDM has poor performance mainly because there are very few training samples, with the unsatisfactory results especially on the QUT fish dataset. Compared with other methods, the proposed method demonstrates state-of-the-art capabilities, which validates that the novel AM module, namely Sandwich Attention, can better solve the problem of few-shot fine-grained fish image classification.

4.4 Ablation study

We then conducted ablation study to experimentally demonstrate the effectiveness of our different design choices. For this ablation study, the three datasets mentioned in 4.1 were used and the same convolutional layers as the baseline architecture were also employed. The experimental settings were consistent with those in 4.2. The proposed module design process was divided into two parts, the first part to be examined was to add effective attention to solve the problem which was encountered on the fish datasets, and the second part considered how to incorporate attention modules that were effective for problem solving. The details of each experiment are explained below.

The first line of experimental results in Table 2 is the 5-way 1-shot and 5-way 5-shot classification accuracies obtained by the baseline method CovaMNet on three datasets (i.e., WildFish, fishclassifierfinal, QUT fish dataset); the second line of results shows where after the features were extracted through the convolutional layer of the baseline, the features were passed through the CBAM (Woo et al., 2018) module to obtain the accuracy results of the 5-way 1-shot and 5-way 5-shot; the third line of results shows where the feature was extracted by the convolutional layer on the baseline, and then was passed through the CBAM module (Woo et al., 2018) and the MHSA module (Srinivas et al., 2021), and the features obtained through the two AM modules were paralleled before finally being sent to the classification network to obtain the 5-way 1-shot and 5-way 5-shot accuracy results; the results in the fourth line show where the features were passed through the CBAM module (Woo et al., 2018) and the MHSA module (Srinivas et al., 2021) after the features were extracted in the convolutional layer on the baseline, the three features obtained by the two AM modules and the features obtained by the original extraction were paralleled and then sent to the classification network to obtain the 5-way 1-shot and 5-way 5-shot accuracy results.

Through the comparison of experimental results, it can be found that the original feature map extracted by the convolutional layer has been paralleled with the CBAM and MHSA modules, forming our Sandwich Attention module, such a network structure can allow the network to more comprehensively consider the importance of different regions and channels of the fish image feature map, weight the feature map more accurately, parallel connection with the feature map can effectively ensure the integrity of the original information, so that our experimental results are significantly higher than our baseline.

4.5 Results visualization

For qualitative analysis, the results are presented in the form of t-SNE diagram (Van der Maaten and Hinton, 2008), which is a machine learning algorithm for nonlinear dimensionality reduction, and usually reduces high-dimensional data to 2 dimensions or 3 dimensions for visualization. Here we show the output visualization results of the baseline CovaMNet, SACovaMNet mentioned in 3.2, and SACovaMNet* mentioned in 3.3, on the fishclassifierfinal dataset for 5-way 5-shot classification

TABLE 2 Ablation study on different choices and connections of AM modules, in terms of the 5-way 1-shot and 5-way 5-shot classification accuracies on the three datasets, *i.e.*, WildFish, fishclassifierfinal, and QUT fish dataset, with 95% confidence intervals.

	5-Way Accuracy(%)				
	WildFish		fishclassifierfinal		QUT fish dataset
	1-shot	5-shot	1-shot	5-shot	1-shot
Baseline	70.87	84.33	54.54	68.52	66.86
CBAM	73.63	85.41	57.23	68.52	68.06
CBAM+MHSA	72.97	85.29	57.61	68.21	67.04
CBAM+feature+MHSA (Ours)	71.44	85.88	58.89	69.01	68.85

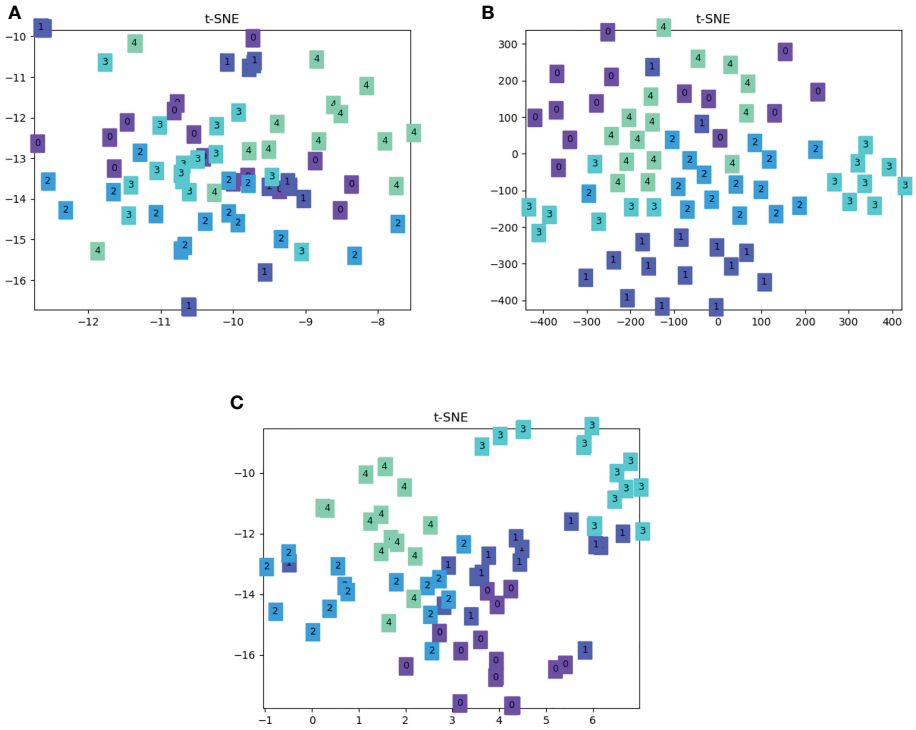

FIGURE 5
Visualization comparison of the t-SNE on baseline CovaMNet, SACovaMNet, and SACova-MNet*. The same color represents one category. **(A)** Visualization of t-SNE on baseline CovaMNet. **(B)** Visualization of t-SNE on SACova-Mnet. **(C)** Visualization of t-SNE on SACova-MNet*.

tasks. The same color in the figure represents the data of the same category. It can be seen from Figure 5A that there is a problem of overlap between different categories, and the boundary of each category is unclear, which will lead to poor classification effects. In Figure 5B, the situation where there is overlap between different categories is reduced, however the data between the same category is relatively scattered. In comparison, the clustering effect in Figure 5C is better, and the boundaries between categories are clearer. The results indicate that our method can make the classification more accurate.

5 Conclusion

In this paper, an approach called SACovaMNet was proposed for few-shot fine-grained marine fish species classification to address the problems caused by a lack of marine fish data and difficulties in classification. The proposed SACovaMNet can extract fish features in detail by fusing CBAM and MHSA in the case of few-shot settings. At the same time, DPCACBAM is proposed to correctly locate the identified objects and key feature information to improve the accuracy of the fine-grained classification, while also applying a second-order covariance metric for similarity comparison that fully takes into account the local similarity information of the feature maps. Based on extensive experiments, the proposed method is shown to be superior to the state-of-the-art methods and the training process is much simpler, providing a basis for research in marine life conservation and marine production.

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 author.

Author contributions

JZ, LH, YW, and XW designed the study and wrote the draft of the manuscript with contributions from YX and MY. YX and MY collected the marine fish image datasets. YW and XW devised the method. JZ and LH performed the experiments. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 32073029) and the Key Project of Shandong Provincial Natural Science Foundation (No. ZR2020KC027).

Acknowledgments

We thank the Intelligent Information Sensing and Processing Lab at Ocean University of China for their computing servers and collaboration during experiments. We also thank Leon Bevan Bullock for his suggestions on manuscript writing. We kindly thank the Editor Dr. Hongsheng Bi for his efforts to handle this manuscript

and all the reviewers for their constructive suggestions that helped us to improve our present manuscript.

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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alsmadi, M. K., and Almarashdeh, I. (2022). A survey on fish classification techniques. *J. King Saud Univ. - Comput. Inf. Sci.* 34, 1625–1638. doi: 10.1016/j.jksuci.2020.07.005

Alsmadi, M. K., Tayfour, M., Alkhasawneh, R. A., Badawi, U., Almarashdeh, I., and Haddad, F. (2019). Robust feature extraction methods for general fish classification. *Int. J. Electrical Comput. Eng.* 9, 5192–5204. doi: 10.11591/ijece.v9i6

Anantharajah, K., Ge, Z., McCool, C., Denman, S., Fookes, C. B., 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*. 309–316.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. (2019). "A closer look at few-shot classification," in *Proceedings of the International Conference on Learning Representations*. 1–17.

Chen, G., Sun, P., and Shang, Y. (2017). "Automatic fish classification system using deep learning," in *Proceedings of the International Conference on Tools for Artificial Intelligence*. 24–29.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*. 1–22.

Finn, C., Abbeel, P., and Levine, S. (2017). "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the International Conference on Machine Learning*.

Guo, Z., Gu, Z., Zheng, B., Dong, J., and Zheng, H. (2022). Transformer for image harmonization and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–19. doi: 10.1109/TPAMI.2022.3207091

Guo, Z., Zhang, L., Jiang, Y., Niu, W., Gu, Z., Zheng, H., et al. (2020). "Few-shot fish image generation and classification," in *Proceedings of the Global Oceans 2020: Singapore–US Gulf Coast*. 1–6.

He, J., Kortylewski, A., and Yuille, A. (2023). "CORL: Compositional representation learning for few-shot classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3890–3899.

Hochreiter, S., Younger, A. S., and Conwell, P. R. (2001). "Learning to learn using gradient descent," in *Proceedings of the International Conference on Artificial Neural Networks*. 87–94.

Hou, R., Chang, H., MA, B., Shan, S., and Chen, X. (2019). "Cross attention network for few-shot classification," in *Proceedings of the Advances in Neural Information Processing Systems*. 4005–4016.

Kartika, D. S. Y., and Herumurti, D. (2016). "Koi fish classification based on HSV color space," in *Proceedings of the International Conference on Information & Communication Technology and Systems*. 96–100.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*. doi: 10.48550/arXiv.1412.6980

Lee, K., Maji, S., Ravichandran, A., and Soatto, S. (2019). "Meta-learning with differentiable convex optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10657–10665.

Lee, S., Moon, W., and Heo, J.-P. (2022). "Task discrepancy maximization for fine-grained few-shot classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5331–5340.

Li, X., Li, Y., Zheng, Y., Zhu, R., Ma, Z., Xue, J.-H., et al. (2023). ReNAP: Relation network with adaptiveprototypical learning for few-shot classification. *Neurocomputing* 520, 356–364. doi: 10.1016/j.neucom.2022.11.082

Li, X., Wu, J., Sun, Z., Ma, Z., Cao, J., and Xue, J.-H. (2021). BSNet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Trans. Image Process.* 30, 1318–1331. doi: 10.1109/TIP.2020.3043128

Li, J., Xu, W., Deng, L., Xiao, Y., Han, Z., and Zheng, H. (2022). Deep learning for visual recognition and detection of aquatic animals: A review. *Rev. Aquac.* 15, 1–25. doi: 10.1111/raq.12726

Li, W., Xu, J., Huo, J., Wang, L., Gao, Y., and Luo, J. (2019). "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*. 8642–8649.

Li, X., Yu, L., Fu, C.-W., Fang, M., and Heng, P.-A. (2020). Revisiting metric learning for few-shot image classification. *Neurocomputing* 406, 49–58. doi: 10.1016/j.neucom.2020.04.040

Liu, Y., Bai, Y., Che, X., and He, J. (2022). "Few-shot fine-grained image classification: A survey," in *Proceedings of the International Conference on Natural Language Processing*. 201–211.

Liu, P., Zhang, C., Qi, H., Wang, G., and Zheng, H. (2022). Multi-attention DenseNet: A scattering medium imaging optimization framework for visual data pre-processing of autonomous driving systems. *IEEE Trans. Intelligent Transport. Syst.* 23, 25396–25407. doi: 10.1109/TITS.2022.3145815

Luo, Z., Zou, Y., Hoffman, J., and Fei-Fei, L. F. (2017). "Label efficient learning of transferable representations acrosss domains and tasks," in *Proceedings of the Advances in Neural Information Processing Systems*. 165–177.

McGlamery, B. L. (1980). "A computer model for underwater camera systems," in *Proceedings of the Ocean Optics VI*. 221–231.

Nichol, A., Achiam, J., and Schulman, J. (2018). On first-order meta-learning algorithms. *arXiv*. doi: 10.48550/arXiv.1803.02999

Peng, Z., Li, Z., Zhang, J., Li, Y., Qi, G.-J., and Tang, J. (2019). "Few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 441–449.

Prasetyo, E., Suciati, N., and Fatichah, C. (2022). Multi-level residual network VGGNet for fish species classification. *J. King Saud University-Computing Inf. Sci.* 204, 5286–5295. doi: 10.1016/j.jksuci.2021.05.015

Rathi, D., Jain, S., and Indu, S. (2017). "Underwater fish species classification using convolutional neural network and deep learning," in *Proceedings of the International Conference on Advances in Pattern Recognition*. 1–6.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., et al. (2018). "Meta-learning for semi-supervised few-shot classification," in *Proceedings of the International Conference on Learning Representations*. 1–15.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., et al. (2019). "Meta-learning with latent embedding optimization," in *Proceedings of the International Conference on Learning Representations*. 1–17.

Shevchenko, V., Eerola, T., and Kaarna, A. (2018). "Fish detection from low visibility underwater videos," in *Proceedings of the International Conference on Pattern Recognition*. 1971–1976.

Shi, Z., Guan, C., Li, Q., Liang, J., Cao, L., Zheng, H., et al. (2022). Detecting marine organisms *via* joint attention-relation learning for marine video surveillance. *IEEE J. Ocean. Eng.* 47, 959–974. doi: 10.1109/JOE.2022.3162864

Snell, J., Swersky, K., and Zemel, R. (2017). "Prototypical networks for few-shot learning," in *Proceedings of the Advances in Neural Information Processing Systems*. 4077–4087.

Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., and Vaswani, A. (2021). "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16519–16529.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018). "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1199–1208.

Tharwat, A., Hemedan, A. A., Hassanien, A. E., and Gabel, T. (2018). A biometric-based model for fish species classification. *Fish. Res.* 204, 324–336. doi: 10.1016/j.fishres.2018.03.008

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

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 Advances in Neural Information Processing Systems*. 5998–6008.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. (2016). "Matching networks for one shot learning," in *Proceedings of the Advances in Neural Information Processing Systems*. 3630–3638.

Wei, X.-S., Song, Y.-Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., et al. (2021). Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 8927–8948. doi: 10.1109/TPAMI.2021.3126648

Wertheimer, D., Tang, L., and Hariharan, B. (2021). "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8012–8021.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*. 3–19.

Zhang, C., Cai, Y., Lin, G., and Shen, C. (2020). "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12203–12213.

Zhang, N., Donahue, J., Girshick, R., and Darrell, T. (2014). "Part-based r-CNNs for fine-grained category detection," in *Proceedings of the European Conference on Computer Vision*. 834–849.

Zhao, B., Feng, J., Wu, X., and Yan, S. (2017). A survey on deep learning-based fine-grained object classification and semantic segmentation. *Int. J. Automation Comput.* 14, 119–135. doi: 10.1007/s11633-017-1053-3

Zhao, S., Zhang, S., Liu, J., Wang, H., Zhu, J., Li, D., et al. (2021). Application of machine learning in intelligent fish aquaculture: A review. *Aquaculture* 540, 736724. doi: 10.1016/j.aquaculture.2021.736724

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2021). A comprehensive survey on transfer learning. *Proc. IEEE* 109, 43–76. doi: 10.1109/JPROC.2020.3004555

Zhuang, P., Wang, Y., and Qiao, Y. (2018). "WildFish: A large benchmark for fish recognition in the wild," in *Proceedings of the ACM International Conference on Multimedia*. 1301–1309.