IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 30, 2018, accepted November 19, 2018, date of publication December 5, 2018,
date of current version December 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2885055

Improving Transfer Learning and Squeeze-and-Excitation Networks for Small-Scale Fine-Grained Fish Image Classification

CHENCHEN QIU[1], SHAOYONG ZHANG[1], CHAO WANG[1], ZHIBIN YU[1], (Member, IEEE), HAIYONG ZHENG[1,2], (Member, IEEE), AND BING ZHENG[1], (Member, IEEE)

[1] Department of Electronic Engineering, College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China
[2] Department of Mathematics, School of Science and Engineering, University of Dundee, Dundee DD1 4HN, U.K.

Corresponding authors: Haiyong Zheng (zhenghaiyong@ouc.edu.cn) and Bing Zheng (bingzh@ouc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771440 and Grant 41776113, in part by the Qingdao Municipal Science and Technology Program under Grant 17-1-1-5-jch, and in part by the China Scholarship Council under Grant 201806335022.

ABSTRACT Scientific studies on species composition and abundance distribution of fishes have considerable importance to the fishery industry, biodiversity protection, and marine ecosystem. In these studies, fish images are typically collected with the help of scuba divers or autonomous underwater vehicles. These images are then annotated manually by marine biologists. Such a process is certainly a tremendous waste of manpower and material resources. In recent years, the introduction of deep learning has helped making remarkable progress in this area. However, fish image classification can be considered as fine-grained problem, which is more challenging than common image classification, especially with low-quality and small-scale data. Meanwhile, well-known effective convolutional neural networks (CNNs) consistently require a large quantity of high-quality data. This paper presents a new method by improving transfer learning and squeeze-and-excitation networks for fine-grained fish image classification on low-quality and small-scale datasets. Our method enhances data augmentation through super-resolution reconstruction to enlarge the dataset with high-quality images, pre-pretrains, and pretrains to learn common and domain knowledge simultaneously while fine-tuning with professional skill. In addition, refined squeeze-and-excitation blocks are designed to improve bilinear CNNs for a fine-grained classification. Unlike well-known CNNs for image classification, our method can classify images with insufficient low-quality training data. Moreover, we compare the performance of our method with commonly used CNNs on small-scale fine-grained datasets, namely, Croatian and QUT fish datasets. The experimental results show that our method outperforms popular CNNs with higher fish classification accuracy, which indicates its potential applications in combination with other newly updated CNNs.

INDEX TERMS Deep learning, image classification, image recognition, transfer learning, underwater technology.

I. INTRODUCTION

With the advancement of technology in modern society, people have considerably better exploration and comprehension of our ocean. Meanwhile, abundant ocean resources, which are newly discovered, are attracting an increasing number of explorers worldwide [1]–[4]. As a result of constantly exploiting and utilizing our limited ocean resources, the biodiversity, especially fish diversity in marine ecosystem, is exposed to a tremendous threat [5]–[7]. Therefore, effective methods and techniques should be introduced for detecting and estimating fish quantitative distribution *in situ*, such as the image-based fish classification [8], to provide a good environment to the fishes, as well as the marine ecology.

Thus far, remote and diver-based videography is used by an increasing number of marine researchers to collect fish images *in situ* [9]–[11]. Traditionally, marine experts must classify and analyze each image manually, which is time-consuming while requiring professional ability. Therefore, feature extraction methods based on image processing technology have been proposed to classify fish images efficiently.

In this way, three main features, namely, color-based, geometric, and texture features, have been extensively used for fish image classification [12]–[18]. Badawi and Alsmadi [12] attempted to recognize fishes by using color-based features, which are intuitively easy to be distinguished; however, they ignored the fact that color features become distorted as depth and light change, thereby rendering their method less persuasive [19]. Meanwhile, shape and texture features, which can display the outline of fishes, are relatively stable and insensitive to the aforementioned factors. Thus, these features can be generally used in recognition. Larsen *et al.* [17] analyzed these two features to determine fish classes effectively. Rova *et al.* [18] used a deformable template object recognition method to improve the accuracy of texture-based classification.

Besides, more and more classifiers have also been developed for fish image classification. Wang *et al.* [13] attempted to classify fishes through a two-level codebook learning by using shrinking coding coefficients. Saitioh *et al.* [16] performed detailed experiments to prove that a combination of bag of visual words and geometric features could aid in obtaining accurate results. Chuang *et al.* [20], [21] proposed a hierarchical partial classification algorithm that was applied to each level of species hierarchy to recognize underwater fish species. Shiau *et al.* [22] and Hsiao *et al.* [23] adopted a maximum probability of partial ranking method based on sparse representation-based classification to identify fish species. Roberts *et al.* [8] introduced a machine learning framework, such as support vector machines (SVMs), as underlying classifiers [24], [25]. Khotimah *et al.* [26] used decision tree algorithm to establish automatic classification of tuna fish.

Despite the aforementioned handcrafted low-level features, as well as conventional machine learning tools, such as SVMs and PCA, convolutional neural networks (CNNs) composed of only several convolutional and nonlinear layers have shown many advantages on visual tracking [27], [28], saliency detection [29], and image processing [30]–[32]. AlexNet [33] with deep CNNs obtained the highest classification result in comparison with conventional methods in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 [34]. VGGNet [35], which increased the depth of a network, achieved good accuracy in the ILSVRC 2014. Meanwhile, GoogLeNet [36] achieved the highest result in the ILSVRC 2014 because of its improved utilization of computing resources inside a network. Thereafter, residual networks (ResNets) [37] with a depth of up to 152 layers won the first place in the ILSVRC 2015 image classification task. In 2017, SENets [38] refreshed the score in ILSVRC, wherein the top five errors were reduced to 2.251%. In addition, several enhancements of squeeze-and-excitation (SE) networks, such as SE_Inseption_v4, SE_Inception_resnet_v2, and SE_ResNeXt_v1_50, can achieve better results.

Recent works have also introduced CNN-based methods to address the fish image classification problem. Meng *et al.* [10] designed an underwater drone with panoramic camera, which used LeNet [39], AlexNet [33], and GoogLeNet [36] in fish recognition. Qin *et al.* [40] designed a deep architecture composed of convolutional layers, spatial pyramid pooling, and linear SVM classifier to achieve accurate real-world fish dataset recognition.

However, the aforementioned works did not consider that fish image classification is a fine-grained classification and that acquiring human-labeled large-scale fish dataset is difficult. Fine-grained image classification, such as of fish, dog [41], bird [42], and flower [43] species, remains a challenging task and more difficult than common image classification because objects from similar subordinate categories may have marginal visual differences that are difficult to distinguish by humans [42].

Several recent works have developed a great progress in investigation of fine-grained image classification with large-scale datasets, which benefit from the increasing emphasis on identifying critical object parts [44]–[48]. Particularly, bilinear CNNs (B-CNNs) [46] has integrated part localization into a two-stream deep learning framework and can be trained end to end.

Small-scale fine-grained image classification remains a challenging task because CNNs cannot develop professional skill from limited images, (e.g., only 10 samples per category). Furthermore, image distortion occurs when images have low quality, *e.g.*, Figure 1 from the Croatian fish dataset [49]. Hence, low-resolution images cannot achieve ideal classification results. However, many image classification tasks in real world often suffer from limited data with low quality. Different from the previous studies, this work presents a method that preferably focuses on addressing these issues.

On the basis of a B-CNN framework, our method enhances data augmentation, develops a new network block, and designs a new fine-tuning strategy to classify low-quality small-scale and fine-grained images. By comparing popular CNNs with Croatian and QUT fish datasets for fish classification, the experimental results show that our method performs better in terms of classification accuracy. Our method can be also combined with other newly updated CNNs for yielding better results in the future.

The rest of this paper is organized as follows. Section II provides the details of our method for fine-grained image classification on small-scale datasets. Section III describes our used datasets and experimental results, and Section IV concludes the paper.

II. METHODOLOGY

We improve the transfer learning by the following three parts (see Figure 2 for details about the structure of our method). Firstly, for network architecture, we combine the proposed refined SE blocks and B-CNNs to promote fine-grained feature extraction capability. Secondly, we use super-resolution reconstruction to enhance the quality of images for data augmentation. Finally, we propose pre-pretraining strategy to learn professional domain knowledge. We will

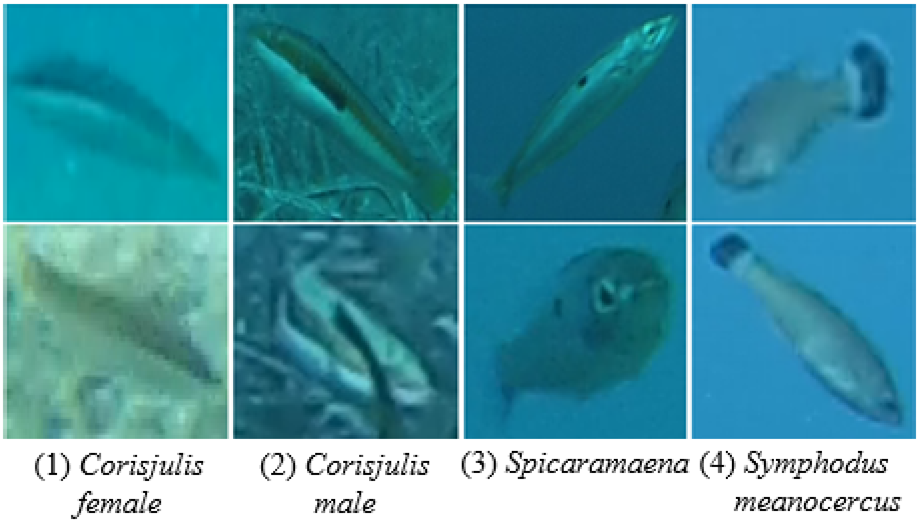


FIGURE 1. Difficulty of Croatian fish dataset [49]: low-quality and fine-grained classification with large intra-class and small inter-class variances.

describe our improved transfer learning method in detail in the remaining part of this section.

A. NETWORK ARCHITECTURE

The network architecture can be presented in quadruple as

$$N = ([\mathbf{f_a}, \mathbf{f_b}], \mathcal{B}, \mathcal{F}), \tag{1}$$

where $\mathbf{f_a}$ and $\mathbf{f_b}$ are feature functions of two CNN streams that contain our refined SE blocks, $\mathcal{B}$ indicates the bilinear pooling, and $\mathcal{F}$ indicates the classification function. In our network architecture, images preprocessed by our enhanced data augmentation are initially passed through two streams of CNN, that is, A and B, to produce localized features. Then, these features are multiplied using the outer product and pooled together to obtain the bilinear vector. Finally, the features are fed into the classification function $\mathcal{F}$ yielding predictions.

1) REFINED SE BLOCK

Generally, CNNs have strong informative feature extraction capability by combining spatial and channel-wise information. Many existing works have boosted representational power and improved classification accuracy via spatial encoding enhancement. Hu *et al.* [38] recently formulated a feature recalibration method to establish channel interdependency by designing a novel architectural unit, namely, SE block, for adaptively recalibrating channel-wise feature response. SE block represents each channel with one single point by using global average pooling. Inspired by the idea of the SE block, we propose *refined SE block* to acquire more accurate information which represents each channel with more points rather than only one point. Specifically, our refined SE block modifies the squeeze operation by dividing each channel into quadrants with crosshairs rather than pooling the entire channel (see Figure 3 and Figure 4 for details). We will explain our refined SE block in more detail below.

First, we initially *squeeze* global spatial information of each channel into a descriptor to determine the relationship among channels. The process is achieved by the SE block [38] through global average pooling of the entire channel to generate channel-wise statistics. And, as shown in Figure 4, we divide each channel into quadrants with crosshairs and represent it with four channel-wise statistics. More specifically, these four statistics, namely, $\mathbf{z_1}, \mathbf{z_2}, \mathbf{z_3}, \mathbf{z_4} \in \boldsymbol{R}^{\boldsymbol{K}}$, are created by squeezing the transformation output $\boldsymbol{U}$ through spatial dimensions $W \times H$, where the element k of $\mathbf{z}$ can be calculated by:

$$\mathbf{z_{k_1}} = \frac{1}{W \times H} \sum_{i=1}^{\frac{H}{2}} \sum_{i=1}^{\frac{W}{2}} u_k(i, j), \tag{2}$$

$$\mathbf{z_{k_2}} = \frac{1}{W \times H} \sum_{i=\frac{H}{2}+1}^{H} \sum_{i=1}^{\frac{W}{2}} u_k(i, j), \tag{3}$$

$$\mathbf{z_{k_3}} = \frac{1}{W \times H} \sum_{i=1}^{\frac{H}{2}} \sum_{i=\frac{W}{2}+1}^{W} u_k(i, j), \tag{4}$$

$$\mathbf{z_{k_4}} = \frac{1}{W \times H} \sum_{i=\frac{H}{2}+1}^{H} \sum_{i=\frac{W}{2}+1}^{W} u_k(i, j). \tag{5}$$

Then, similar to [38], we implement the *excitation* operation by adopting a gating mechanism and ReLU function (represented by δ) to utilize the information produced by *squeeze* operation completely (see Figure 3 for reference),

$$\mathbf{s} = \sigma(\mathbf{L_2}\delta(\mathbf{L_1}e(\mathbf{z}))), \tag{6}$$

where $\mathbf{s}$ indicates the output of the *excitation* operation; e refers to *reshape* operation that aims to change the shape

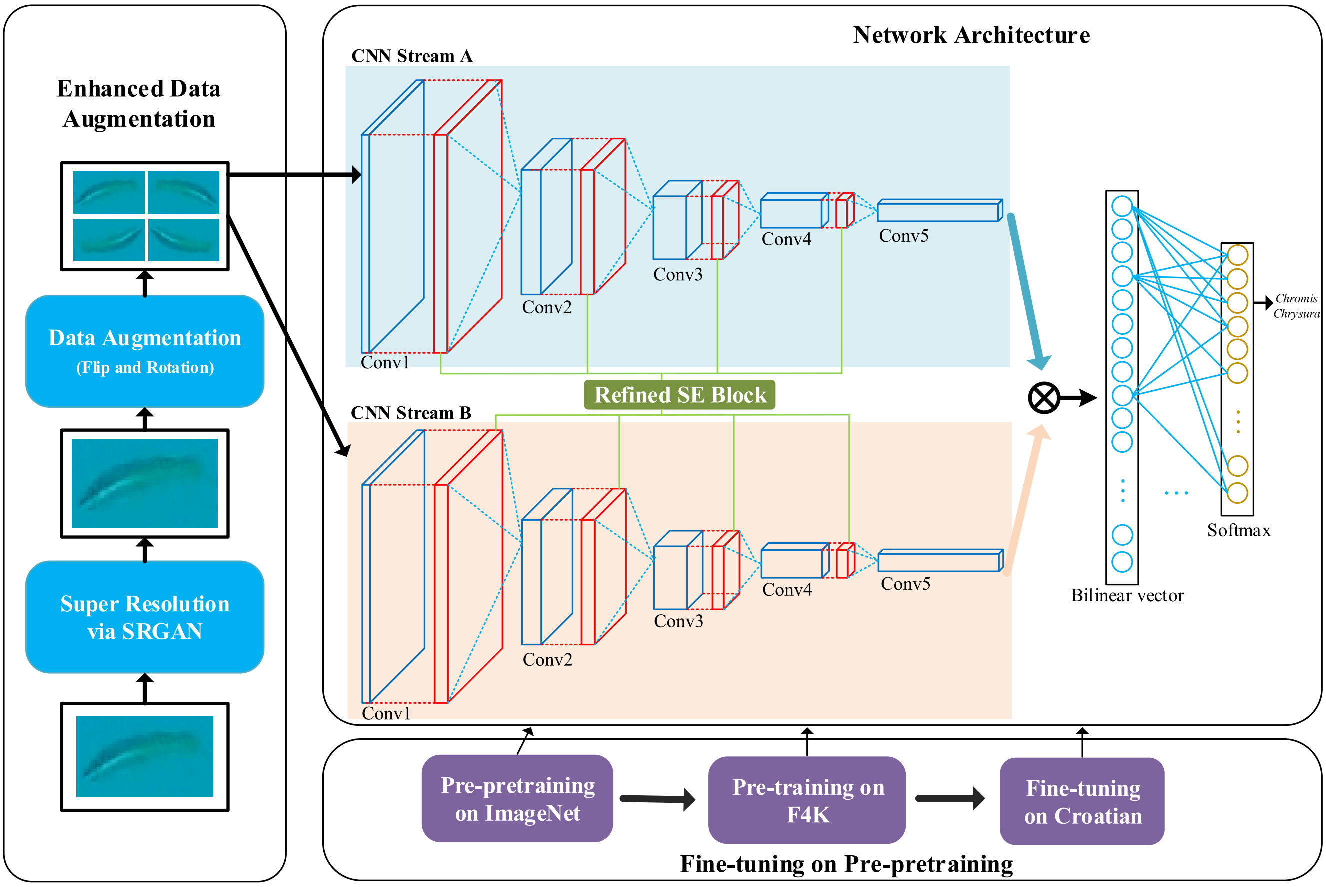


FIGURE 2. Complete classification pipeline of our improved transfer learning with refined squeeze-and-excitation networks.

of $\mathbf{z}$ (the output of *squeeze*) from $1 \times 4 \times C$ to $1 \times 1 \times 4C$ (Figure 4), δ denotes the ReLU function [50]; σ means the sigmoid activation, $\mathbf{L_1} \in \mathcal{R}^{\frac{4C}{r}\times C}$, and $\mathbf{L_2} \in \mathcal{R}^{4C\times\frac{4C}{r}}$, where reduction ratio $r = \frac{L_2}{L_1}$ (We use $r = 4$ in this paper, and the choice of the parameter r and another parameter p are discussed in Section III-B). To increase the generalization capability, we embed two fully connected (FC) layers (see Figure 3 for reference) interleaved with nonlinearity, that is, one FC layer with one-fourth of the length of $\mathbf{z}$ ($4C$) and parameters $\mathbf{L_1}$, a ReLU, and another FC layer with the same length as $\mathbf{z}$ and parameters $\mathbf{L_2}$.

Finally, we regard $\mathbf{s_k}$ (the output of *excitation*) as the most important information of each channel after feature selection. Then, we complete the recalibration of the original feature on the channel dimension by multiplying $\mathbf{s_k}$ and feature map $\mathbf{u_k}$. Finally, we reshape the output of *scale* back to the dimension $1 \times 1 \times C$ as the final output of the block.

2) REFINED SE BLOCK MEETS B-CNNs

Generally, B-CNNs [46] adopt two VGGNets (**M-Net** [51] and **D-Net** [35]) truncated at the convolutional layer as feature function. However, our network uses two **D-Net**s, which can achieve the same benefits as the second-order pooling [52], which is popularized for semantic segmentation and image classification. Moreover, we embed our refined SE block into the convolutional layers, including `conv1`, `conv2`, `conv3`, and `conv4`, to enhance the feature extraction capability of B-CNNs. Hence, a convolutional layer $\widetilde{\mathbf{X}}$ processed by the refined SE block S is calculated as follows:

$$\widetilde{\mathbf{X}} = S(\mathbf{X}), \tag{7}$$

where $\mathbf{X}$ can be any layer of the CNNs.

This combination allows the network to recalibrate each channel of any convolutional layer before being sent to the subsequent convolutional layer. In this way, B-CNNs can balance the contribution of each channel in the convolutional layer. Similar to [46], we adopt bilinear pooling to combine feature outputs of each location. The bilinear pooling operation of the input image $\mathcal{I}$ at location l is defined as follows:

$$\boldsymbol{B}(l, \mathcal{I}, f_a, f_b) = f_a(l, \mathcal{I})^T f_b(l, \mathcal{I}), \tag{8}$$

where f_a and f_b are the outputs of two streams of CNN, that is, A and B. Thereafter, the bilinear features will be initially aggregated by sum pooling, then passed through a signed square root and l_2 normalization, and finally fed into the classification function $\mathcal{F}$ yielding predictions.

Similar to most image classification works, B-CNNs are also pretrained on the ImageNet dataset [34] when domain-specific data are scarce. Furthermore, we use the CNNs that are pre-pretrained on the dataset composed of

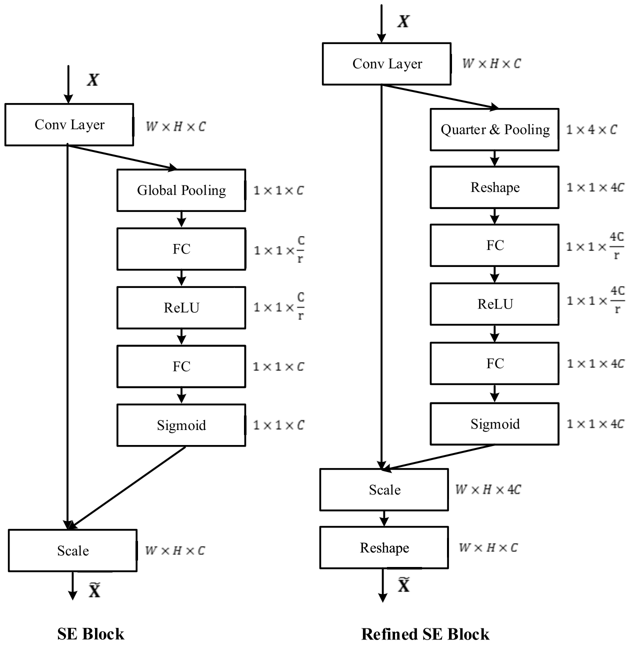


FIGURE 3. SE and refined SE blocks.

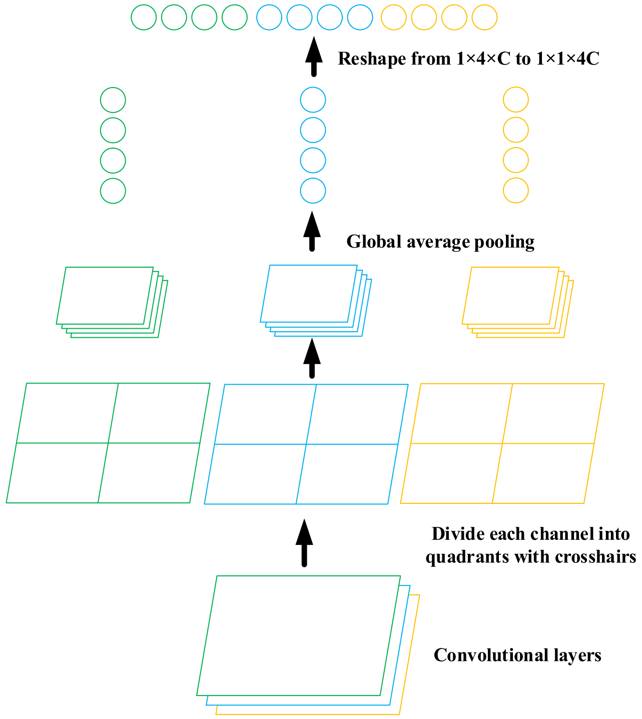

Reshape from 1×4×C to 1×1×4C
Global average pooling
Divide each channel into quadrants with crosshairs
Convolutional layers

FIGURE 4. The "Quarter & Pooling" and "Reshape" of refined SE block.

Fish4Knowledge (F4K) project [53] to learn professional domain knowledge (see Section II-B).

B. FINE-TUNING ON PRE-PRETRAINING

Many existing works for image classification often use pretrained model on the large-scale dataset (*e.g.*, ImageNet [34]) when domain-specific data are scarce. However, pretrained method remains limited in learning professional skill on several computer vision tasks, such as fine-grained image classification. To further improve the performance of CNNs on small-scale datasets, we explore a "pre-pretraining" strategy to learn professional domain knowledge from small-scale datasets.

We initially pre-pretrain our network on the ImageNet dataset, then pretrain it on the F4K dataset, and finally finetune it on a small-scale fine-grained dataset (*i.e.*, Croatian or QUT fish dataset). In this way, the network learns the common classification information during the first prepretraining process, acquires domain knowledge during the second pretraining process, and masters the fine-grained discriminative information during the fine-tuning process. This strategy enables the network to learn the features of the target dataset accurately and comprehensively, which can effectively improve the representation performance of neural networks on small-scale datasets.

C. ENHANCED DATA AUGMENTATION

In addition to the proposed pre-pretraining strategy, we also propose our enhanced data augmentation to enlarge the dataset through super-resolution reconstruction with high quality.

Generally, we need to resize the input image before sending it to the network. Hence, the more complex the network is, the greater size we need. Nevertheless, the method based on the linear interpolation adopted by most image classification works may result in image distortion, especially when using images with low resolution. To address this problem, we formulate a method dubbed enhanced data augmentation which consists of super-resolution reconstruction and general data augmentation.

Specifically, we use the method of super-resolution reconstruction based on generative adversarial network (GAN) to enhance image quality. Ledig *et al.* [54] presented a super-resolution generative adversarial network (SRGAN), which is feasible for improving image quality. To achieve better super-resolution results, SRGAN used a perceptual loss function composed of adversarial and content losses. The adversarial loss was generated by a discriminator to render the generated image close to the natural image. Meanwhile, the content loss was generated from the perceptual similarity of an image, not the similarity in pixel space. By using SRGAN, we can effectively improve the resolution of the dataset and reduce the image distortion problem.

Moreover, on the basis of the super-resolution reconstruction, we employ two types of general data augmentation, namely, flip and rotation. For the flip type, we flip each image horizontally and vertically. For the rotation type, we rotate each image in 90°, 180°, and 270° clockwise. Thus, we achieve additional high-quality images through the enhanced data augmentation. Our network architecture can obtain improved generalization by the enhanced data augmentation.

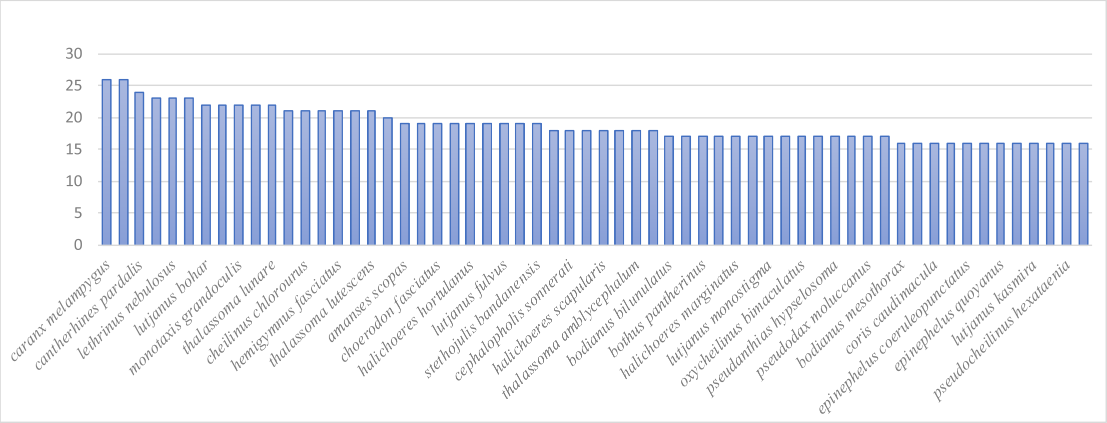


FIGURE 5. Data distribution per category of QUT fish dataset.

TABLE 1. Data distribution per category of the Google Flowers dataset for training and testing.

Species	*Training set*	*Testing set*
(1) *Daisy*	509	126
(2) *Dandelion*	720	179
(3) *Roses*	514	128
(4) *Sunflowers*	560	140
(5) *Tulips*	640	160
Total	2493	733

III. EXPERIMENTAL EVALUATION

We experimentally compare the popular benchmark CNNs to verify the effectiveness of our method. Section III-A describes the data distribution of the dataset and evaluation protocol used in our experiments. The evaluation of the refined SE block is provided in Section III-B. Section III-C provides the details of our compared experiments with popular CNNs to validate the effectiveness of the proposed pre-pretraining strategy. Section III-D presents the compared experiments of various types of B-CNNs with different data augmentation methods to prove the validity of the proposed network architecture and enhanced data augmentation.

A. DATASET AND EVALUATION PROTOCOL

We implement a series of experiments by using five datasets for evaluation.

First, we use the CIFAR-10 [55] and Google Flowers datasets, which are commonly used in computer vision and image classification, to illustrate the effectiveness of our refined SE block. The CIFAR-10 dataset consists of 50,000 32×32 training images and 10,000 test images in 10 classes. Each class of Google Flowers is randomly assigned to training and test sets with the proportion of $4 : 1$. The data distribution of the Google Flowers dataset is listed in Table 1. In these two datasets, we use a fivefold cross-validation scheme with the mean classification accuracy results in all the comparison experiments.

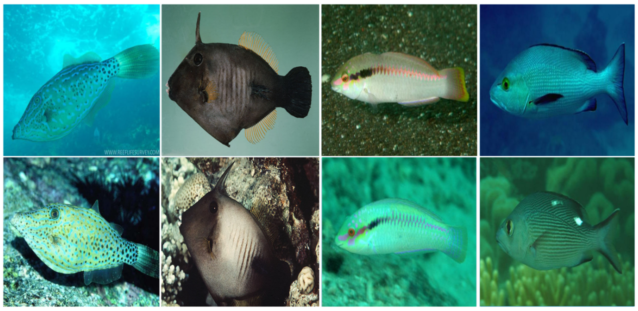
TABLE 2. Data distribution per category of the Croatian fish dataset for training and testing.

Species	*Training set*	*Testing set*
(1) *Chromis chromis*	10	96
(2) *Coris julis female*	10	47
(3) *Coris julis male*	10	47
(4) *Diplodus annularis*	10	84
(5) *Diplodus vulgaris*	10	101
(6) *Oblada melanura*	10	47
(7) *Serranus scriba*	10	46
(8) *Spondyliosoma cantharus*	10	41
(9) *Spicara maena*	10	39
(10) *Symphodus melanocercus*	10	95
(11) *Symphodus tinca*	10	24
(12) *Sarpa salpa*	10	7
Total	120	674

We adopt the Croatian fish [49] and QUT fish datasets [56] to verify the effectiveness of our pre-pretraining strategy. The Croatian fish dataset contains 794 images of 12 fish species for image classification experiments. The samples of this dataset are shown in Figure 1. Data distribution per category of this dataset for training and testing is shown in Table 2. This dataset is suitable for validating the effectiveness of the improved transfer learning with refined squeeze-and-excitation networks for the classification of low-quality small-scale and fine-grained images. Moreover, we select the top 60 largest classes from the QUT fish dataset to further validate the generalization of our pre-pretraining on the fish image classification. The number of images per category varies from 16 to 26. Figure 5 shows the data distribution per category of the QUT fish dataset with $1 : 1$ splitting for training and testing, and Figure 6 shows several samples of this dataset. We adopt a twofold cross-validation scheme to verify the pre-pretraining strategy and use average accuracy as our final results.

TABLE 3. Experiments of popular CNNs embedded with SE and refined SE blocks for image classification on CIFAR-10 and Google Flowers datasets. The values in bold font indicate the best results.

Dataset	SE_Inception_v4	rSE_Inception_v4	SE_Inception_resnet_v2	rSE_Inception_resnet_v2	SE_ResNext_v1_50	rSE_ResNeXt_v1_50
CIFAR-10	79.35%	**83.24%**	83.01%	**86.67%**	92.40%	**93.52%**
Flowers	92.75%	**94.20%**	92.71%	**93.80%**	34.27%	**94.82%**



(1) *Aluterus scriptus* (2) Amanses scopas (3) *Halichoeres scapularis* (4) *Lutjanus bohar*

FIGURE 6. Samples of the QUT fish image dataset.

In addition, we use a middle-scale dataset, namely, F4K dataset [53], to pretrain our pre-pretrained network. Different from Croatian and QUT fish datasets, the F4K dataset contains 22,370 fish images with relatively high quality; thus, it is feasible for a pre-pretrained model of the Croatian or QUT fish dataset. In the experiments, the total images are divided into two subsets, namely, 4/5 for training and 1/5 for testing.

B. EVALUATION OF THE REFINED SE BLOCK

We provide a detailed explanation related to the motivation and structure of our refined SE block in Section II-A1. We implement three pairs of comparison experiments on two commonly used datasets, namely, CIFAR-10 and Google Flowers datasets, of image classification in a computer vision system to validate the generalization of the refined SE block.

As shown in Table 3, three popular neural networks embedded with SE blocks are used as baselines. We replace the SE blocks with our refined SE blocks. The results show that the accuracy of the three networks is improved on both datasets, thereby validating the effectiveness of our refined SE block.

Besides, we test other forms for refining SE block based on SE-ResNet-34. Specifically, the refined SE block can be changed by two factors of r and p: r represents the reduction ration and p indicates the number of points extracted from the channel. To explore the power of different kinds of refined SE block, we designed a series of experiments using CIFAR-10 validation set to find the optimal form listed in Table 4, and it can be seen that, with the increase of p, the amount of parameter size increases significantly while the accuracy keeps stable relatively. Therefore, we consider that the choice of $r = 4$ and $p = 4$ is optimal in our work.

TABLE 4. Accuracy on CIFAR-10 validation set and parameter sizes of refined SE-ResNet-34 at different r and p (accuracy / parameter size).

r	p		
	4	16	64
4	**93.7% / 0.68M**	93.58% / 3.92M	93.54% / 55.56M
8	93.56% / 0.58M	93.20% / 2.20M	93.54% / 28.03M
16	93.60% / 0.52M	93.50% / 1.34M	93.77% / 14.27M
32	93.41% / 0.50M	93.31% / 0.91M	93.42% / 7.39M

C. EVALUATION OF IMPROVED TRANSFER LEARNING

We perform comparison experiments on various popular CNNs, including AlexNet [33], VGG-16 [35], Inception-v4 [36], and ResNet-50 [37], to prove the effectiveness of our pre-pretraining strategy.

We divide four groups of comparison experiments on each CNN model (Table 5) as follows: (1) training from scratch without any transfer learning on the Croatian or QUT dataset (the blank ImageNet and F4K items in Table 5); (2) pretraining the model on ImageNet and fine-tuning the parameters on the Croatian or QUT dataset (the marked ImageNet and blank F4K items in Table 5); (3) pretraining the model on the F4K and fine-tuning the parameters on the Croatian or QUT dataset (the blank ImageNet and marked F4K items in Table 5); and (4) pre-pretraining the model on ImageNet (the proposed strategy), then pre-training the model on F4K, and finally fine-tuning the parameters on the Croatian or QUT dataset (the marked ImageNet and F4K items in Table 5).

As shown in Table 5, we enumerate the results of the four groups of comparison experiments on four popular CNN models. Besides, we also list the results from Croatian [49] and QUT [56] datasets as baselines as well as the state-of-the-art results for comparison [57], [58] in Table 5. On the basis of these experiments, the pre-pretraining strategy effectively improves the accuracy on Croatian and QUT fish datasets by mastering the fine-grained discriminative information of species. Therefore, this strategy can be helpful for small-scale image classification.

D. EVALUATION OF OUR METHOD ON B-CNNS

To prove the effectiveness of our network architecture (see Section II-A) and enhanced data augmentation (see Section II-C), we implement three groups of comparison experiments on various types of B-CNNs by using different data augmentation methods and SRGAN, as presented in Table 6.

Table 6 shows the compared image classification results of B-CNNs, B-CNNs plus SE blocks, and B-CNNs plus

TABLE 5. Experiments of popular CNNs with and without pre-training and pre-pretraining for the image classification on Croatian and QUT fish datasets. The results in bold font indicate the best performance of a specific CNN model.

Network	Pre-trained model		**Accuracy**	
	ImageNet	F4K	Croatian	QUT
AlexNet			40.56%	22.52%
	√		61.46%	39.55%
		√	45.01%	30.88%
	√	√	**62.35%**	**45.57%**
VGG-16			44.25%	24.21%
	√		67.10%	49.27%
		√	64.47%	40.45%
	√	√	**72.07%**	**52.15%**
Inception-v4			40.97%	30.45%
	√		74.26%	50.15%
		√	64.37%	44.56%
	√	√	**78.25%**	**56.22%**
ResNet-50			68.52%	32.22%
	√		76.02%	52.24%
		√	79.24%	48.84%
	√	√	**80.10%**	**58.56%**
Baseline [49], [56]			66.78%	49.3%
State-of-the-art [57], [58]			82.95%	57.0%

TABLE 6. Results of the compared image classification of B-CNNs, B-CNNs plus SE blocks, and B-CNNs plus refined SE blocks on the Croatian fish dataset. The results in bold font indicate the best performance of a specific CNN model.

Network	Enhanced data augmentation		**Accuracy**
	Data augmentation	SRGAN	Croatian
B-CNNs			81.30%
	√		83.36%
		√	82.56%
	√	√	**83.52%**
B-CNNs+ SE blocks			82.26%
	√		83.52%
		√	82.92%
	√	√	**83.78%**
B-CNNs+ refined SE blocks			82.50%
	√		83.56%
		√	83.30%
	√	√	**83.92%**

refined SE blocks on the Croatian fish dataset. Using either the data augmentation or SRGAN has a limited role in the classification performance. However, when they are combined, the improvements are remarkable, thereby indicating that data augmentation and SRGAN are complementary. Furthermore, the combination of B-CNNs and refined SE block by using the enhanced data augmentation achieves the best performance with a **2-3%** improvement by enhancing the feature extraction capability.

TABLE 7. Results of the compared image classification of B-CNNs, B-CNNs plus SE blocks, and B-CNNs plus refined SE blocks on the QUT fish dataset.

Method	QUT
B-CNNs	53.01%
B-CNNs+SE	69.53%
B-CNNs+rSE	71.80%

The comparison experiments of B-CNNs, B-CNNs plus SE blocks, and B-CNNs plus refined SE blocks are performed on the QUT fish dataset. Notably, we do not use the SRGAN on this dataset because of the relatively high image quality of the QUT fish dataset. As shown in Table 7, the accuracy of the refined SE blocks improves from 2.27% to 71.8% in comparison with B-CNNs plus SE blocks, which demonstrates the generalization capability of our proposed method.

IV. CONCLUSION

In this study, we propose an improved transfer learning method with refined squeeze-and-excitation networks for fine-grained fish image classification on small-scale datasets. This method enables the network to learn the features of the target dataset accurately and comprehensively. Thus, it achieves better performance in fish image classification. The experimental results show that our method outperforms other popular CNNs with the highest classification accuracy. Moreover, the accuracy can also be improved further if better CNN network architecture is adopted. In future study, we will combine our method with the newly updated deep CNNs for other fine-grained image classification applications.

REFERENCES

[1] J. T. Cobb, K. C. Slatton, and G. J. Dobeck, "A parametric model for characterizing seabed textures in synthetic aperture sonar images," *IEEE J. Ocean. Eng.*, vol. 35, no. 2, pp. 250–266, Apr. 2010.

[2] H. Eriksson *et al.*, "Contagious exploitation of marine resources," *Frontiers Ecol. Environ.*, vol. 13, no. 8, pp. 435–440, 2015.

[3] P. M. Cury *et al.*, "The ecosystem approach to fisheries: Reconciling conservation and exploitation," *Tools Oceanogr. Ecosyst. Model.*, pp. 221–311, Jul. 2016.

[4] J. Ramos-Muñoz *et al.*, "Early use of marine resources by Middle/Upper Pleistocene human societies: The case of Benzú rockshelter (northern Africa)," *Quaternary Int.*, vol. 407, pp. 6–15, Jul. 2016.

[5] L. N. K. Davidson, M. A. Krawchuk, and N. K. Dulvy, "Why have global shark and ray landings declined: Improved management or overfishing?" *Fish Fisheries*, vol. 17, no. 2, pp. 438–458, 2016.

[6] O. Le Pape, S. Bonhommeau, A.-E. Nieblas, and J.-M. Fromentin, "Overfishing causes frequent fish population collapses but rare extinctions," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 31, p. E6274, 2017.

[7] B. Meissa and D. Gascuel, "Overfishing of marine resources: Some lessons from the assessment of demersal stocks off mauritania," *ICES J. Mar. Sci.*, vol. 72, no. 2, pp. 414–427, 2014.

[8] P. L. D. Roberts, J. S. Jaffe, and M. M. Trivedi, "Multiview, broadband acoustic classification of marine fish: A machine learning framework and comparative analysis," *IEEE J. Ocean. Eng.*, vol. 36, no. 1, pp. 90–104, Jan. 2011.

[9] S. B. Williams *et al.*, "Monitoring of benthic reference sites: Using an autonomous underwater vehicle," *IEEE Robot. Autom. Mag.*, vol. 19, no. 1, pp. 73–84, Mar. 2012.

[10] L. Meng, T. Hirayama, and S. Oyanagi, "Underwater-drone with panoramic camera for automatic fish recognition based on deep learning," *IEEE Access*, vol. 6, pp. 17880–17886, 2018.

[11] S. Hasija, M. J. Buragohain, and S. Indu, "Fish species classification using graph embedding discriminant analysis," in *Proc. IEEE Int. Conf. Mach. Vis. Inf. Technol.*, Feb. 2017, pp. 81–86.

[12] U. A. Badawi and M. K. Alsmadi, "A general fish classification methodology using meta-heuristic algorithm with back propagation classifier," *J. Theor. Appl. Inf. Technol.*, vol. 66, no. 3, pp. 803–812, 2014.

[13] G. Wang, J.-N. Hwang, K. Williams, F. Wallace, and C. S. Rose, "Shrinking encoding with two-level codebook learning for fine-grained fish recognition," in *Proc. IEEE ICPR Workshop Comput. Vis. Anal. Underwater Imag.*, Dec. 2016, pp. 31–36.

[14] A. Baareh, "A hybrid memetic algorithm (genetic algorithm and tabu local search) with back-propagation classifier for fish recognition," *Int. Rev. Comput. Softw.*, vol. 8, no. 6, pp. 1287–1293, 2013.

[15] M. K. Alsmadi, K. B. Omar, and S. A. Noah, "Fish classification based on robust features extraction from color signature using back-propagation classifier," *J. Comput. Sci.*, vol. 7, no. 1, p. 52, 2011.

[16] T. Saitoh, T. Shibata, and T. Miyazono, "Feature points based fish image recognition," *Int. J. Comput. Inf. Syst. Ind. Manage. Appl.*, vol. 8, pp. 12–22, 2016.

[17] R. Larsen, H. Olafsdottir, and B. K. Ersbøll, "Shape and texture based classification of fish species," in *Proc. Scand. Conf. Image Anal.* Berlin, Germany: Springer, 2009, pp. 745–749.

[18] A. Rova, G. Mori, and L. M. Dill, "One fish, two fish, butterfish, trumpeter: Recognizing fish in underwater video," in *Proc. IAPR Conf. Mach. Vis. Appl.*, 2007, pp. 404–407.

[19] R. Schettini and S. Corchs, "Underwater image processing: State of the art of restoration and image enhancement methods," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, 2010, Art. no. 746052.

[20] M.-C. Chuang, J.-N. Hwang, and K. Williams, "Supervised and unsupervised feature extraction methods for underwater fish species recognition," in *Proc. IEEE ICPR Workshop Comput. Vis. Anal. Underwater Imag.*, Aug. 2014, pp. 33–40.

[21] M.-C. Chuang, J.-N. Hwang, F.-F. Kuo, M.-K. Shan, and K. Williams, "Recognizing live fish species by hierarchical partial classification based on the exponential benefit," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5232–5236.

[22] Y.-H. Shiau, S.-I. Lin, Y.-H. Chen, S.-W. Lo, and C.-C. Chen, "Fish observation, detection, recognition and verification in the real world," in *Proc. Int. Conf. Image Process., Comput. Vis., Pattern Recognit. (WorldComp)*, 2012, pp. 1–6.

[23] Y.-H. Hsiao, C.-C. Chen, S.-I. Lin, and F.-P. Lin, "Real-world underwater fish recognition and identification, using sparse representation," *Ecol. Inform.*, vol. 23, pp. 13–21, Sep. 2014.

[24] S. O. Ogunlana, O. Olabode, S. A. A. Oluwadare, and G. B. Iwasokun, "Fish classification using support vector machine," *Afr. J. Comput. ICT*, vol. 8, no. 2, pp. 75–82, 2015.

[25] P. Bosch, J. López, H. Ramírez, and H. Robotham, "Support vector machine under uncertainty: An application for hydroacoustic classification of fish-schools in Chile," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4029–4034, 2013.

[26] W. N. Khotimah, A. Z. Arifin, A. Yuniarti, A. Y. Wijaya, D. A. Navastara, and M. A. Kalbuadi, "Tuna fish classification using decision tree algorithm and image processing method," in *Proc. IEEE Int. Conf. Comput., Control, Inform. Appl.*, Oct. 2015, pp. 126–131.

[27] Y. Zhou, X. Bai, W. Liu, and L. J. Latecki, "Similarity fusion for visual tracking," *Int. J. Comput. Vis.*, vol. 118, no. 3, pp. 337–363, Jul. 2016.

[28] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognit.*, vol. 48, no. 9, pp. 2885–2905, 2015.

[29] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.

[30] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[31] A. Mahmood *et al.*, "Deep image representations for coral image classification," *IEEE J. Ocean. Eng.*, Jan. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8255621

[32] C. Bentes, D. Velotto, and B. Tings, "Ship classification in TerraSAR-X images with convolutional neural networks," *IEEE J. Ocean. Eng.*, vol. 43, no. 1, pp. 258–266, Jan. 2018.

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

[34] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[35] K. Simonyan and A. Zisserman. (Apr. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[36] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

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

[38] J. Hu, E. Wu, L. Shen, and G. Sun. (Sep. 2017). "Squeeze-and-excitation networks." [Online]. Available: https://arxiv.org/abs/1709.01507

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[40] H. Qin, X. Li, J. Liang, Y. Peng, and C. Zhang, "DeepFish: Accurate underwater live fish recognition with a deep architecture," *Neurocomputing*, vol. 187, pp. 49–58, Apr. 2016.

[41] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, vol. 2, Jun. 2011, pp. 1–2.

[42] S. Branson, G. Van Horn, S. Belongie, and P. Perona. (Jun. 2014). "Bird species categorization using pose normalized deep convolutional nets." [Online]. Available: https://arxiv.org/abs/1406.2952

[43] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, May 2014, pp. 806–813.

[44] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 834–849.

[45] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5546–5555.

[46] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1449–1457.

[47] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1173–1182.

[48] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognit.*, vol. 76, pp. 704–714, Apr. 2018.

[49] J. Jaeger, M. Simon, J. Denzler, V. Wolff, K. Fricke-Neuderth, and C. Kruschel, "Croatian fish dataset: Fine-grained classification of fish species in their natural habitat," in *Proc. Mach. Vis. Animals Behav.*, 2015, pp. 1–7.

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

[51] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (Nov. 2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531

[52] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 430–443.

[53] B. J. Boom, P. X. Huang, J. He, and R. B. Fisher, "Supporting ground-truth annotation of image datasets using clustering," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 1542–1545.

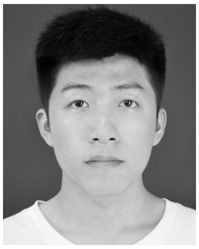
[54] C. Ledig *et al.* (Sep. 2016). "Photo-realistic single image super-resolution using a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.04802

[55] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

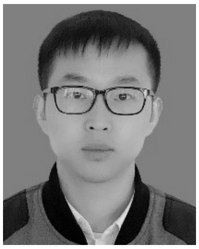
[56] K. Anantharajah *et al.*, "Local inter-session variability modelling for object classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 309–316.

[57] J. Zhao *et al.*, "Semi-supervised learning-based live fish identification in aquaculture using modified deep convolutional generative adversarial networks," *Amer. Soc. Agricult. Biol. Eng.*, vol. 61, no. 2, pp. 699–710, 2018.

[58] Z. Ge, C. McCool, C. Sanderson, and P. Corke, "Modelling local deep convolutional neural network features to improve fine-grained image classification," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 4112–4116.



CHENCHEN QIU received the B.E. degree from the Guilin University of Electronic Technology in 2017. He is currently pursuing the master's degree with the College of Information Science and Engineering, Ocean University of China. His research interests include deep learning and image classification.

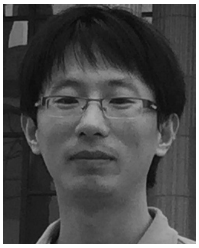


SHAOYONG ZHANG received the B.S. degree from Henan Polytechnic University in 2016. He is currently pursuing the master's degree with the College of Information Science and Engineering, Ocean University of China. His research interests include deep learning and depth map prediction.

CHAO WANG received the B.E. degree in communication engineering from the Ocean University of China in 2016, where he is currently pursuing the master's degree with the College of Information Science and Engineering. His research interests include deep learning and computer vision.

ZHIBIN YU (M'16) received the B.S. degree in thermal energy and power engineering from the Harbin Institute of Technology, Harbin, China, in 2005, and the M.S. degree in computer engineering and the Ph.D. degree in electrical engineering from Kyungpook National University, Daegu, South Korea, in 2011 and 2016, respectively.

In 2016, he joined the Department of Electronic Engineering, Ocean University of China, where he is currently a Lecturer. His research interests include underwater image processing, artificial neural networks, and underwater 3D reconstruction.



HAIYONG ZHENG (M'12) received the B.S. degree in electronic information engineering and the Ph.D. degree in ocean information sensing and processing from the Ocean University of China, Qingdao, China, in 2004 and 2009, respectively.

In 2009, he joined the Department of Electronic Engineering, Ocean University of China, where he is currently an Associate Professor. His research interests include image processing, computer vision, and machine learning.

BING ZHENG (M'07) received the B.S. degree in electronics and information system, the M.S. degree in marine physics, and the Ph.D. degree in computer application technology from the Ocean University of China, Qingdao, China, in 1991, 1995, and 2013, respectively.

He is currently a Professor with the Department of Electronic Engineering, Ocean University of China. His research interests include ocean optics, underwater imaging, and optical detection.