Heliyon 9 (2023) e16761

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Fish-TViT: A novel fish species classification method in multi water areas based on transfer learning and vision transformer

Bo Gong [a,b,c], Kanyuan Dai [c,d,e], Ji Shao [a,b,d], Ling Jing [a,b,c,d,e,*], Yingyi Chen [a,b,c,d,*]

[a] College of Information and Electrical Engineering, China Agricultural University, 100083, Beijing, China
[b] National Innovation Center for Digital Fishery, China Agricultural University, Beijing, 100083, China
[c] Key Laboratory of Smart Farming Technologies for Aquatic Animal and Livestock, Ministry of Agriculture and Rural Affairs, Beijing, 100083, China
[d] Beijing Engineering and Technology Research Center for Internet of Things in Agriculture, China Agricultural University, Beijing, 100083, China
[e] College of Science, China Agricultural University, Beijing, 100083, China

ARTICLE INFO

Keywords:
Fish species classification
Deep learning
Vision transformer
Transfer learning

ABSTRACT

The classification of fish species has important practical significance for both the aquaculture industry and ordinary people. However, existing methods for classifying marine and freshwater fishes have poor feature extraction ability and do not meet actual needs. To address this issue, we propose a novel method for multi-water fish classification (Fish-TViT) based on transfer learning and visual transformers. Fish-TViT uses a label smoothing loss function to solve the problem of overfitting and overconfidence of the classifier. We also employ Gradient-weighted Category Activation Mapping (Grad-CAM) technology to visualize and understand the features of the model and the areas on which the decision depends, which guides the optimization of the model architecture. We first crop and clean fish images, and then use data augmentation to expand the number of training datasets. A pre-trained visual transformer model is used to extract enhanced features of fish images, which are subsequently cropped into a series of flat patches. Finally, a multi-layer perceptron is used to predict fish species. Experimental results show that Fish-TViT achieves high classification accuracy on both low-resolution marine fish data (94.33%) and high-resolution freshwater fish data (98.34%). Compared with traditional convolutional neural networks, Fish-TViT has better performance.

1. Introduction

According to the latest world fish database [1], the total number of fish species present in the world is approximately 30,000. The total number of fish species in aquaculture exceeds 600 but is still dominated by a few bulk species, with 20 fish species accounting for 83.6% of farmed production, as reported in the FAO survey in 2020 [2]. The coexistence of different fish species in natural or artificial water environments makes fish species identification crucial for aquaculture, population management, and environment monitoring. Therefore, the study of this issue has significant scientific research significance. In recent years, computer vision has developed rapidly and been applied to medicine [3], transportation [4], agriculture [5], and it is widely used in aquaculture [6]. The study of automatic recognition of fish species images with the help of computer vision technology contributes to the study

* Corresponding authors at: National Innovation Center for Digital Fishery, China Agricultural University, Beijing, 100083, China.
E-mail addresses: jingling@cau.edu.cn (L. Jing), chenyingyi@cau.edu.cn (Y. Chen).

https://doi.org/10.1016/j.heliyon.2023.e16761
Received 23 December 2022; Received in revised form 25 May 2023; Accepted 26 May 2023

Available online 1 June 2023

of fish population analysis, fish quality measurement, and fish health monitoring, and is likewise an important part of intelligent aquaculture [7]. The problem of fish classification includes three main aspects: image data collection, feature extraction, and classifier construction. Classical machine learning and deep neural networks are commonly used in classification problem.

In machine learning-based fish species classification methods, some features of image, such as shape and color, are used to learn and classify the fish species. Nair et al. [8] proposed a support vector based underwater fish species identification method which used speedup robust features to extract fish image features, and then used support vector machine for classification. Islam et al. [9] proposed a method for fish identification using a decision tree and K-NN methods for prediction after acquiring local and global features of fish images. Urbanova et al. [10] used the classical method of machine learning to classify fish contours, and image preprocessing including target background segmentation and image alignment. The modified Rosenblatt algorithm was used as the training loss function for discriminant analysis. Hnin and Lynn [11] proposed a fish species classification method based on robust feature selection by machine learning techniques, using three classifiers to verify the validity of the selected feature subset. Ogunlana et al. [12] selected the shape features of fish images to construct the training set, and input these features into a support vector machine to classify fish images. The classification accuracy is significantly higher than the methods based on K-NN and K-means. Guisande et al. [13] proposed an expert system called IPez for fish species classification that was able to distinguish six species of fish, including those from marine and freshwater. AdaBoost classifier was used to classify texture and color features of fish images, where LDA was used to reduce the duality of features [14]. Hasija et al. ([15]) considered the problems of image distortion, background noise, and occlusion due to underwater photography and used graph embedding discriminant analysis to operate on a single input image, which can recognize underwater fish images. To summarize, machine learning-based fish recognition methods have limitations such as the need for manual a priori knowledge, the impact of subjective factors and external environment, time-consuming and labor-intensive, and not suitable for large-scale applications.

Machine learning algorithms require more information to make predictions, while deep learning algorithms can make predictions by learning from data processing alone. Image classification methods based on deep learning have superior performance, and researchers have already proposed using them for fish species classification. Allken [16] proposed a method of using convolutional neural networks to train synthetic fish image data for the classification of marine fish species, mainly addressing the problem of insufficient training data and achieving the classification accuracy of 94%. Urbanova et al. [10] proposed to use a fine-tuned and optimized VGG16 network to classify the fish species on Verde islands. Due to the limited data, various data enhancement methods were used. But, the number of fish species categories that the method was able to classify was only 3. Ovalle [17] proposed using Mask R-CNN method to solve the fish species classification problem in a deck-specific scenario. Due to the complex background of the deck, the recognition of the model is poor when the fish bodies overlap. Xu [18] solved the fish species classification problem with a small sample category imbalance based on migration learning and the SE-ResNet152 model, using Focal-loss as the training loss function of the model, and achieved good classification results. However, this method does not consider fish images with complex backgrounds, which are simpler. Ju and Xue [19] analysed the factors of low recognition accuracy of traditional algorithms in complex backgrounds and various lighting situations and proposed an improved AlexNet model for fish species classification, which introduced a soft attention mechanism and used fewer convolutional layers, resulting in a model with high classification accuracy and low computational complexity. Rauf [20] proposed a 32-layer convolutional neural network model based on the VGG architecture for fish species recognition, which can recognize six different fish species. Khalifa [21] proposed an aquarium fish recognition system using a convolutional neural network model to recognize eight fish species with high accuracy and low computational effort. Hridayami [22] proposed a method capable of predicting 50 fish species categories based on the VGG16 network using a migration learning approach pre-trained on the ImageNet dataset. Four different types of datasets were also used, including RGB images, canny filtered images, hybrid images, and hybrid images accompanied by RGB. Villon [23] proposed a fast identification method based on deep learning for species prediction of underwater reef fish. The method first used convolutional neural networks to train the images, followed by the design of different post-processing decision rules that were able to identify 20 species of fish. Salman et al. [24] used convolutional neural networks to identify unconstrained underwater fish, considering complex underwater image backgrounds and moving fish targets, with a final model classification accuracy of 90%. [25] proposed deep learning to differentiate Individuals of undulate skate (Raja undulata). The method was based on Siamese neural networks, using statistical fundamentals as the motivation to overcome few-shot context. [26] aimed to validate CVS (Computer vision system) by deep learning to predict morphometric measurements in Tambaqui and estimate genetic parameters for growth traits and body yield. For image segmentation, a method was proposed based on Mask R-CNN (Region-based Convolutional Neural Networks). In summary, deep learning-based fish species classification methods mostly use convolutional neural networks to extract features and multi-layer perceptrons (MLPs) as classifiers, which are more effective. However, shallow networks may only extract local information.

Low-order convolution layer only extracts the local features of the image, but the vision transformer can easily derive the global information. So, the vision transformer methods have achieved good results in the computer vision tasks, such as the ViT [27] directly using the transformer encoder to encode patches of images, inspired by the transformer model in NLP [28]. The transformer has multiple advantages that CNN lacks: the ability to learn more inductive biases; the structure is applicable to most domains, such as NLP, CV, and multi-modal learning; and the learned relational representation is more general and robust than local patterns from convolutional modules. In the field of fisheries, there is a paper based on the BERT model for detecting the behavior of Atlantic salmon, which utilizes an attention mechanism to interpret sonar images and predict fish behavior [29]. Li et al. [30] proposed an improved spatial Transformer network for detecting underwater targets. However, few vision transformer-based methods for fish species classification have emerged.

Despite significant progress in deep learning-based fish image classification methods, traditional CNN structures used in these methods suffer from certain limitations such as sensitivity to input size. To overcome this challenge, we propose a novel Transformer-

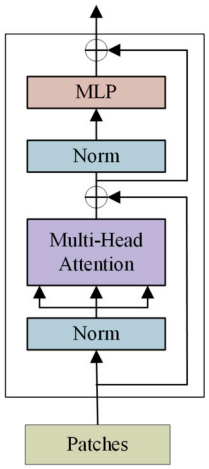

Fig. 1. Transformer encoder. The illustration of the Transformer encoder [28].

based method for fish image classification that can efficiently handle images of different resolutions and demonstrate excellent generalization ability. Our approach is highly applicable and effectively distinguishes between low-resolution marine fish images and high-resolution freshwater fish images. To further improve the model's performance, we implemented a label smoothing loss function to alleviate overfitting and overconfidence issues in the fish classifier. Moreover, we used Gradient-weighted Class Activation Mapping (Grad-CAM) techniques to visualize and understand the model's features and regions relied upon to make decisions. This insight was used to guide the optimization of the model's architecture. To enhance the model's robustness, we applied data augmentation techniques such as cropping and random flipping on the input data. We also expedited the training process of the model by utilizing pre-trained model weights. In summary, our proposed Transformer-based method for fish image classification effectively addresses the limitations of traditional CNN architectures and provides superior performance in processing fish images of different resolutions.

The rest of this paper is organized as follows: Section 2 presents our proposed method and introduces the vision transformer encoder. Section 3 describes the fish datasets used in this paper, and shows examples of the datasets and the distribution of the datasets. Section 4 shows the experimental results of our method and analyzes the impacts. Section 5 is the summary of this paper.

2. Method

Below we first introduce the structure of the transformer encoder in Fig. 1 and explain the core structure of the vision transformer. Then, we illustrate the overall framework of our proposed method, Fish-TViT, in Fig. 2. We only use the pre-trained transformer encoder to extract the fish images' features. The MLP is trained to predict the fish species.

2.1. Transformer encoder

The field of computer vision has always relied on convolutional neural networks until transformer is introduced and shows excellent performance. ViT directly uses the standard transformer encoder structure for image tasks, and the encoder structure is shown in Fig. 1. First, the image is sliced into multiple patches, then these patches are feature embedding and position embedding, and input to the Transformer encoder along with class embedding. Finally, it is input to MLP for class prediction. Multi-head attention is the core structure of the transformer encoder. Its input is three vectors query, key, and value, and the output is the weighted value of all values. By matrix Q, K, and V, the attention matrix is represented as Equation (1).

$$Attention(Q,K,V) = softmax\left(\frac{Q^T K}{\sqrt{d_k}}V\right) \tag{1}$$

The matrices Q, K, V in Equation (1) are for matrix calculation. Multi-head attention is obtained by splicing operation of multiple attention. Multi-head attention show in Equation (2) has a rich subspace representation with different positions that can be expressed as

$$\begin{aligned} MHA(Q,K,V) &= Concat\left(head_1, head_2, ..., head_h\right)W^0 \\ head_i &= Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{2}$$

where $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$.

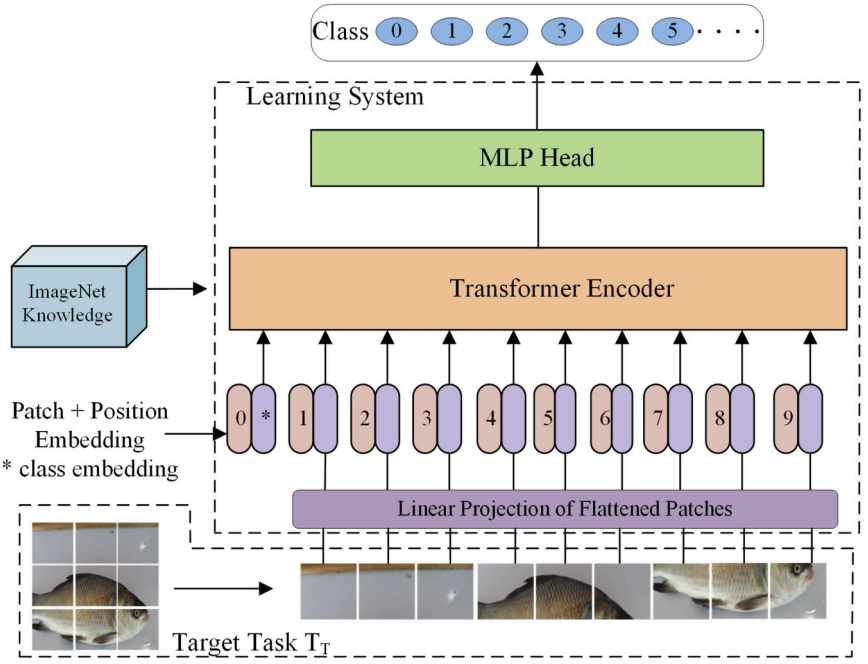

Fig. 2. The structure of Fish-TViT. Initially, the image is partitioned into multiple patches, which are then fed into the encoder following position encoding and linear projection. The encoder employs pre-trained weights from ImageNet. Eventually, the predicted category is produced by the MLP.

2.2. Fish-TViT

The fish species identification method proposed in this paper is based on transfer learning and vision transformer model. In conventional deep learning models, the strategy of training neural networks from scratch is computationally expensive and has a high demand for data. To reduce the computational cost and data usage, we use the Vision Transformer model pre-trained on the ImageNet dataset for fine-tuning, which is used to extract the generalized features of fish images and continue to train the MLP for classification.

Considering the large amount of data required to train a vision transformer model from a new one, the computational resource consumption is too high, and the number of fish currently is small, for this reason, we consider using transfer learning to improve the feature extraction capability of the model for fish data. Therefore, we use the vision transformer model trained on the ImageNet dataset, and after obtaining the pre-trained model weights, we fine-tune them for extracting the features in this study. The specific process is shown in the Fig. 2. The ViT is trained on the source task ImageNet data in advance to obtain the pre-trained model weights. Then the pre-trained model is used on the target dataset, i.e., fish datasets. Finally, the encoder layer of ViT is frozen and only the MLP layer is trained for fish species image classification.

2.3. Loss function

The loss function is one of the essential components in deep learning. Equation (3) shows the loss function of the cross-entropy.

$$loss(\theta) = -\sum_{i=1}^{K} y_i \log \hat{y}_i \tag{3}$$

The loss function measures the error between an output of model $\hat{y}_i$ and a label y_i, and the hyperparameter θ is found to minimize the loss function. In the real world, the networks usually become overfitting and overconfident. Therefore, we use label smoothing to stimulate the activations of the penultimate layer to be close to the template of the correct class and keep some distance from the template of the wrong class. It improves accuracy by computing cross-entropy, using a weighted mixture of these targets with a uniform distribution. The formula of the label smoothing loss function is Equation (4):

$$\begin{aligned} L_{labelsmooth} &= -\sum_{i}^{n} y_k \log\left(p_k\right) \\ y_k &= \begin{cases} \frac{\varepsilon}{n}, & others \\ 1-\varepsilon+\frac{\varepsilon}{n}, & correct\ class \end{cases} \end{aligned} \tag{4}$$

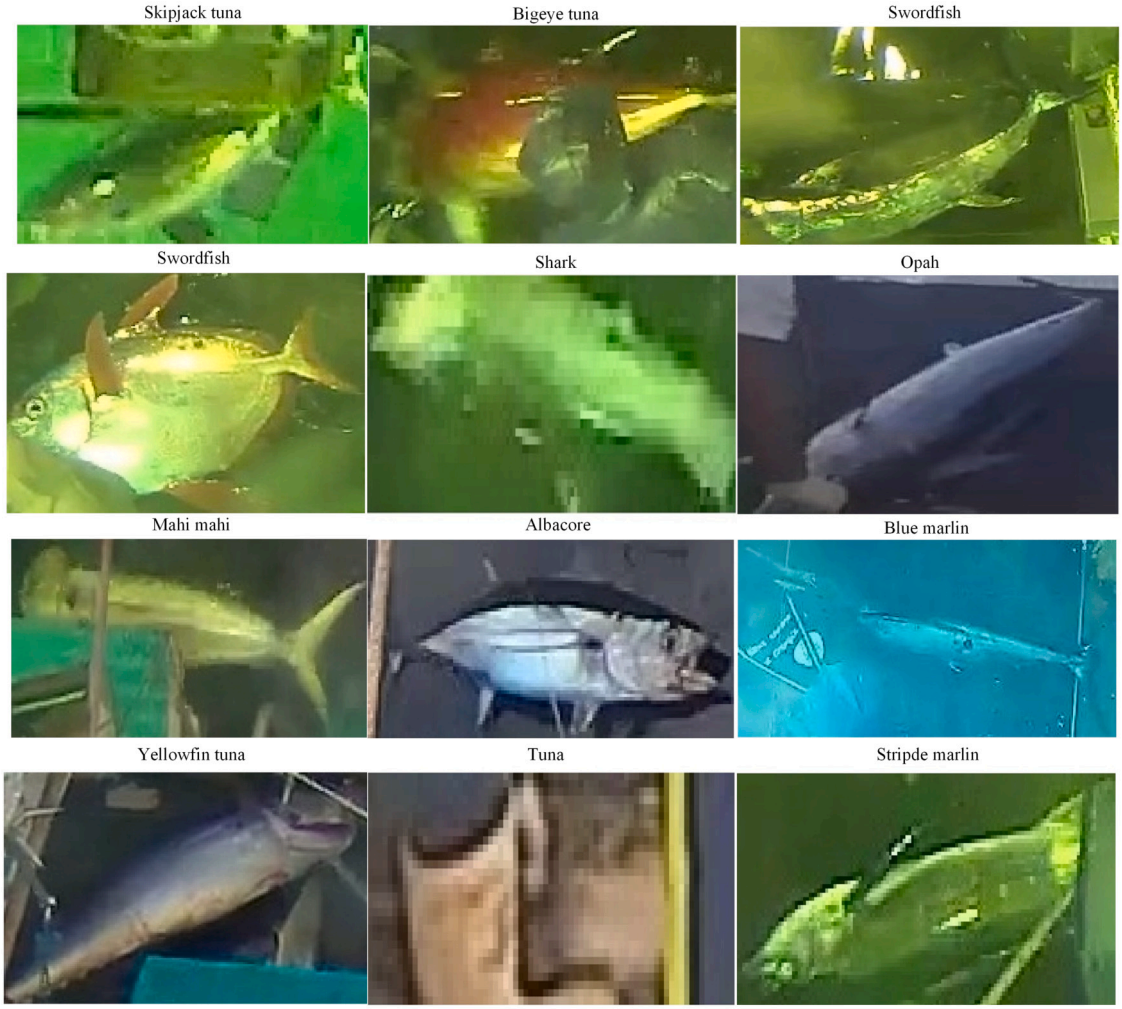

Fig. 3. The 12 categories of marine fish images.

where p_k is the network's output, y_k represents the true targets. A label smoothing factor of $\varepsilon = 0.1$, and n is the number of categories. In this paper, the LabelSmoothing is used to improve the performance of proposed method.

3. Training and evaluation indicators

3.1. Datasets

In this paper, we used two kinds of datasets, the low-resolution images of marine fishes and high-resolution images of freshwater fishes. The marine fish image dataset we used was cropped from the public dataset, Fishnet [31] based on the bounding box annotations, containing about 19 thousand low-resolution images with 12 categories of marine fish, e.g. Yellowfin tuna, Shortbill spearfish, and Bigeye tuna, as shown in Fig. 3. Fishnet vision1.0.0, an open fish images dataset, containing 30 objects, used for marine fish fine grained classification and detection, approximately 34 GB. It is an electronic monitoring (EM) data. We split the dataset into a training set and a test set, with a data set split ratio of 7:3. Fig. 4 shows the distribution of different categories of data. For freshwater fish images, we used the Fish-Pak dataset [32] which contains six categories freshwater fish images, 3 views ($head, body, and scales$), with a resolution of 5184×3456. All six species are freshwater fish, and there are some similarities between the images, depicted in Fig. 5. There is a small amount of data and unbalanced between each category. The fish scale view of these datasets are not used, the details of the head view and body view shown in Fig. 6.

To improve the classification performance of the model, we processed the images data including image scaling and image augmentation. A random crop was used to adjust the pixels of the image to 224×224, followed by a random flip. In this paper, the datasets follow international animal testing welfare guidelines.

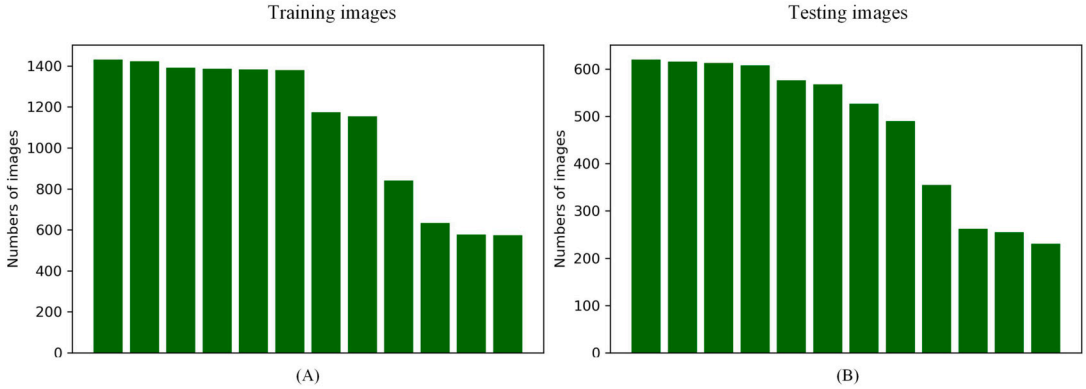

Fig. 4. The distributions of low resolution marine fish images cropped from the Fishnet.

3.2. Experimental setup

All experiments were conducted using an NVIDIA GeForce RTX 3090 24 GB GPU and an Intel i9-12900K CPU with 64 GB RAM, utilizing open-source Python and PyTorch frameworks. The two test sets comprised 5721 marine fish samples and 181 freshwater fish samples, respectively. Our proposed model was trained for 100 epochs using the AdamW optimization strategy with a learning rate of 0.003 and a weight decay factor of 0.3. Prior to input into our Fish-TViT model, image sizes were scaled and data augmentation techniques such as random cropping and flipping were applied. The model utilized pre-trained weights from the ViT architecture on the ImageNet dataset. We conduct three repeated runs and calculate the average.

3.3. Evaluation metrics

To evaluate the effectiveness of fish species classification model, we use several evaluation metrics, including accuracy, precision, recall, and F1-score. The calculation procedures are outlined by Equation (5) and are described below:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ Precision &= \frac{TP}{TP+F} \\ Recall &= \frac{TP}{TP+FN} \\ F1-score &= \frac{2\times Recall\times Precision}{RecallPrecision} \end{aligned} \tag{5}$$

where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively.

4. Results and discussion

4.1. Methods comparison

To illustrate the performance of the Fish-TViT, we conduct comparative experiments, and the experimental results as follows.

Low-resolution marine fish datasets: The Fish-TViT is compared with several CNN-based methods on the low-resolution marine fish datasets. The experimental results are shown in Table 1 where the best results are in bold. The performance of our proposed method is superior, and the classification accuracy reaches 94.33% better than the CNN-based methods. In contrast to VAN [33], our proposed method has achieved a 19.66% improvement in accuracy. Compared our proposed method with ConvMixer [34] the accuracy of Fish-TViT exceeds 7.811%. This is due to the global feature representation ability of the model in this paper. To illustrate the effect of different patch size and image size on the proposed method, we set two kinds of 16 and 32, and the input image size contains 224 and 384, the experimental performance is shown in Table 2. The proposed method achieves the maximum classification accuracy of 94.33% with 16 patch size and 224×224. Increasing the patch size reduces the detection accuracy of the model.

High-resolution freshwater fish datasets: The Fish-TViT is compared with CNN-based methods on the high-resolution freshwater fish. The details of the experiment are shown in Table 3. The accuracy of Fish-TViT is 98.34%, 8.84% larger than ResNet50. In order to investigate the impact of the number of patch size, we conducted experiments with different patch sizes. VAN combines convolution and self-attention, but Fish-TViT surpasses VAN by 0.55%. In contrast to ConvMixer, the accuracy of our method is increased by 1.65%. We discuss that under the high-resolution fish dataset, the detection accuracy of our model is the highest. However, the best performer among other metrics is the VAN model. This benefits from the introduction of large kernel attention. Table 4 shows the results of the tests with various sizes of patches. It can be seen that Fish-TViT with 16 patch size has a higher accuracy when the input pixel size is 224×224, which is 0.55% better than the model with 32 patches. The comparison also shows that the

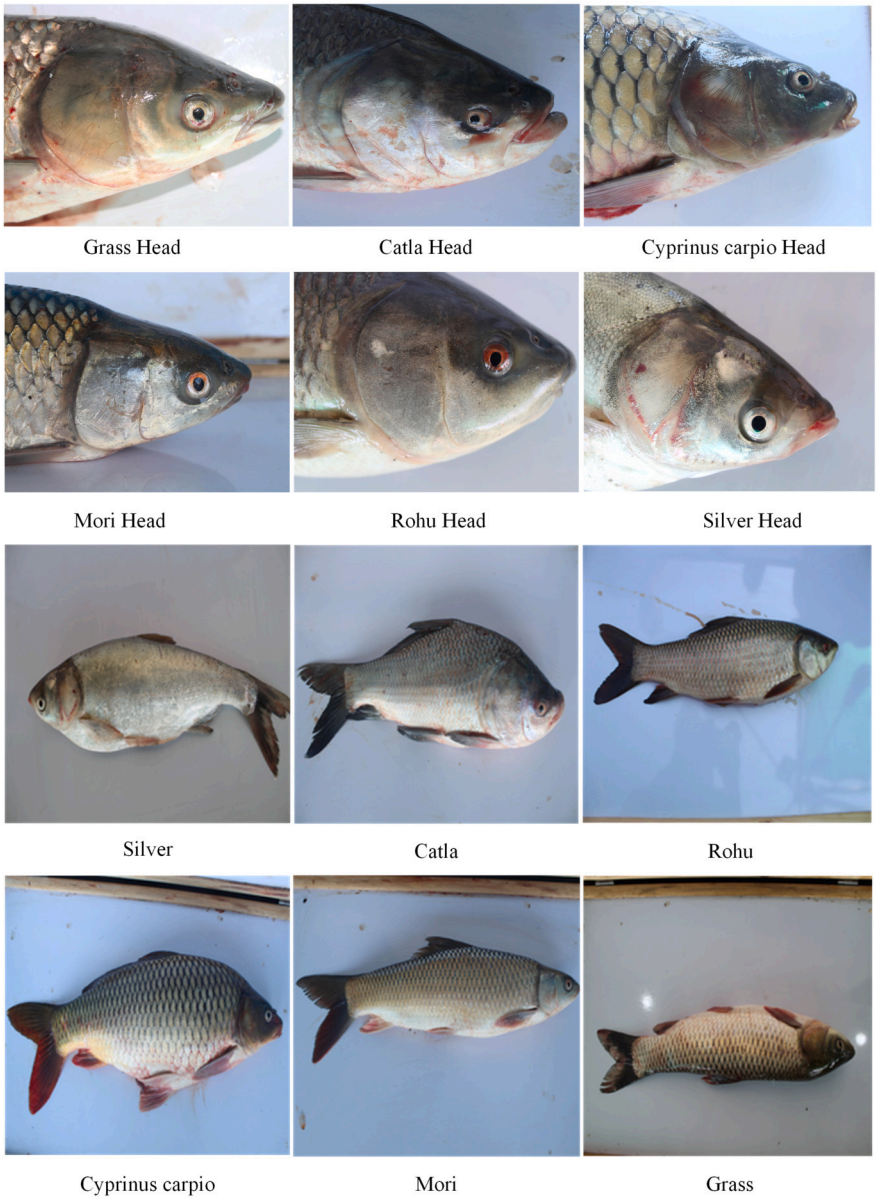

Fig. 5. Example images of six fish species with head view and body view.

accuracy of Fish-TViT with a pixel size of 384×384 is 92.33% with a reduction of 6.01%, and the F1-score is reduced by 3.43% when the numbers of patches are 16. Patch size has a great influence on the performance of the model and should be determined according to the resolution and size of the input image. We compare the precision, recall, and F1-score on each category, as shown in Table 5. Compared with ResNet50, the F1-score of Fish-TViT is improved by 5.79% and 2.70% in the categories Grass carp and Catla, respectively. In the category of Cyprinus Carpio, compared with ResNet50, the precision value of proposed method decreases by 1.96%, and the Recall value is reduced by 0.99%. In the category of Rohu, compared with ResNet50, the Recall value of the proposed method decreases by 3.03%, and the F1 value drops by 1.54%. The performance degradation of the proposed method on these two categories is that our method does not adequately represent the local information of the image. The transformer neural network relies on the self-attention mechanism, which is global and can learn the global dependencies of features, but cannot capture the local information in the image. We use the confusion matrix to compare the test results of the models. The darker color on the diagonal line indicates that Fish-TViT predicts more number of correct categories, as shown in Fig. 7. First, the category Cyprinus carpio has the highest number of categories in the whole test set, and the category with the lowest number of categories is Grass carp. In addition, the category Grass carp has more correct predictions than other two models, and this category has a small number of samples in the dataset, but Fish-TViT still achieves better classification results.

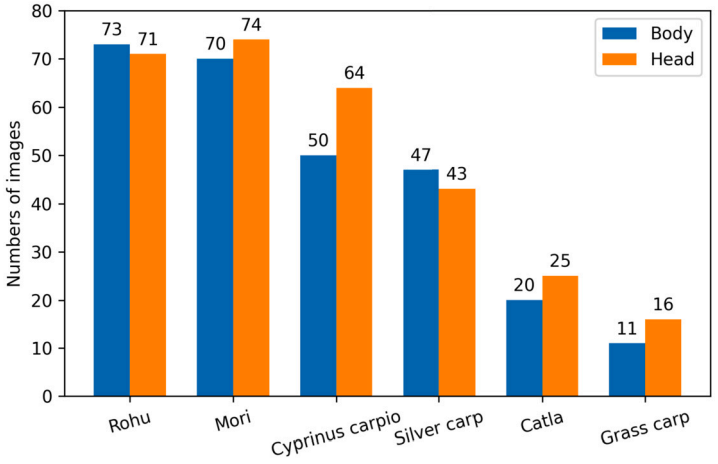

Fig. 6. The distribution of freshwater fish images, with body and head views.

Table 1
Our proposed method classification accuracy on low resolution marine fish images.

Methods	Acc(Top1)	Acc(Top5)	Precision	Recall	F1-score
ResNet50 [35]	92.13	99.09	91.58	91.86	91.70
DenseNet121 [36]	76.98	95.63	75.91	76.09	75.47
VGG19 [37]	78.08	95.98	78.42	76.81	76.86
VAN [33]	74.67	94.72	73.78	72.78	72.78
ConvMixer [34]	86.52	98.02	85.62	85.84	85.63
Fish-TViT(our)	**94.33**	**99.37**	**93.75**	**93.98**	**93.85**

Table 2
The classification performance comparisons of the proposed method with two patch sizes on low-resolution marine fish images.

Patch size	Image-size	Acc(Top1)	Acc(Top5)	Precision	Recall	F1-score
Patch-16	224*224	**94.33**	**99.37**	**93.75**	**93.98**	**93.85**
Patch-32	224*224	66.26	91.45	66.32	64.44	63.98
Patch-16	384*384	89.79	98.32	89.37	89.51	89.37

Table 3
Performance comparisons of classification methods on high-resolution fresh water fish images.

Model	Acc(Top1)	Precision	Recall	F1-score
ResNet50 [35]	89.50	91.16	84.53	84.80
VGG19 [37]	77.9	67.58	67.48	66.4
DenseNet121 [36]	69.06	73.79	62.82	64.23
VAN [33]	97.79	**97.90**	**96.23**	**96.90**
ConvMixer [34]	96.69	97.59	94.32	95.58
Fish-TViT	**98.34**	96.20	94.32	95.12

Table 4
Classification performance comparison of the proposed method with two patch sizes on high-resolution freshwater fish images.

Patch size	Image-size	Acc(Top1)	Precision	Recall	F1-score
Patch-32	224×224	97.79	**98.58**	**96.12**	**97.10**
Patch-16	224×224	**98.34**	96.20	94.32	95.12
Patch-16	384×384	92.33	91.67	91.81	91.69

Table 5
Comparisons of different classification methods for each class of high-resolution freshwater fish images.

Model	Classes	Precision	Recall	F1-Score
ReNet50 [35]	Catla	94.74	100.00	97.30
	Cyprinus Carpio	97.96	97.96	97.96
	Grass carp	100.00	72.73	84.21
	Mori	95.45	100.00	97.67
	Rohu	100.00	100.00	100.00
	Silver carp	100.00	100.00	100.00
Fish-TViT	Catla	100.00	100.00	100.00
	Cyprinus Carpio	96.00	97.96	96.97
	Grass carp	100.00	81.82	90.00
	Mori	95.45	100.00	97.67
	Rohu	100.00	96.97	98.46
	Silver carp	100.00	100.00	100.00

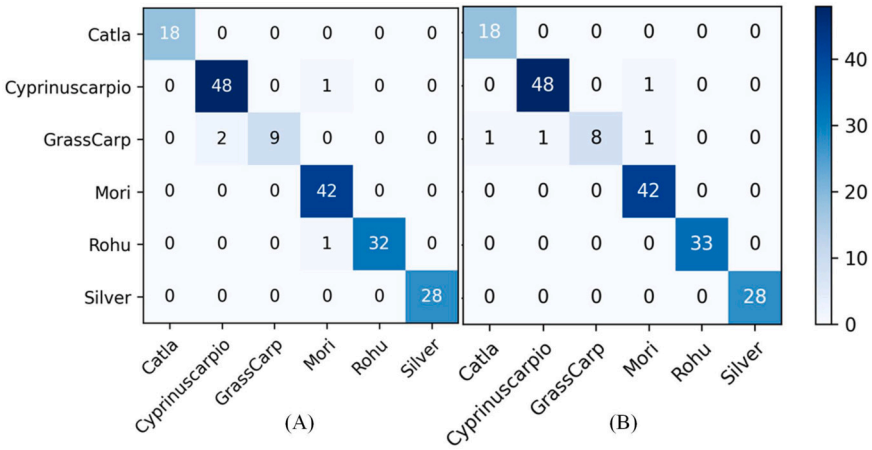

Fig. 7. The confusion matrix of (*A*) Fish-TViT and (*B*) ResNet50 on the low-resolution freshwater fish images.

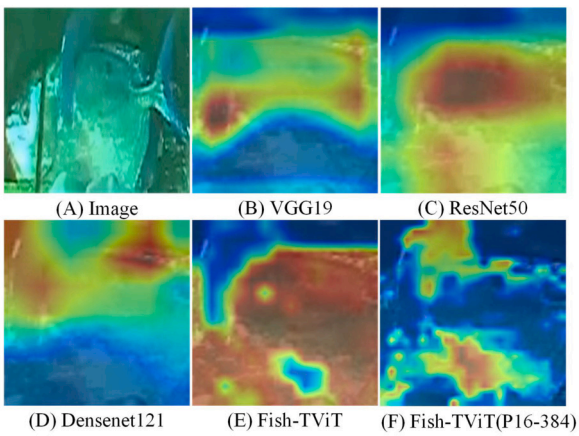

Fig. 8. The Grad-CAM visualizations from Fish-TViT and CNN-based methods for low-resolution marine fish classification.

4.2. Visual interpretation

Gradient-weighted Category Activation Mapping (Grad-CAM) [38] is a visual interpretation technique that can make the network more transparent and interpretable. It helps to identify the region a method is looking at while classifying an image. The penultimate layer of transformer encoder is used as the target in Fish-TViT, with the high score categories corresponding to the red areas. As shown in Fig. 8 and Fig. 9, the feature map of Fish-TViT contains more comprehensive category information and engages the regions

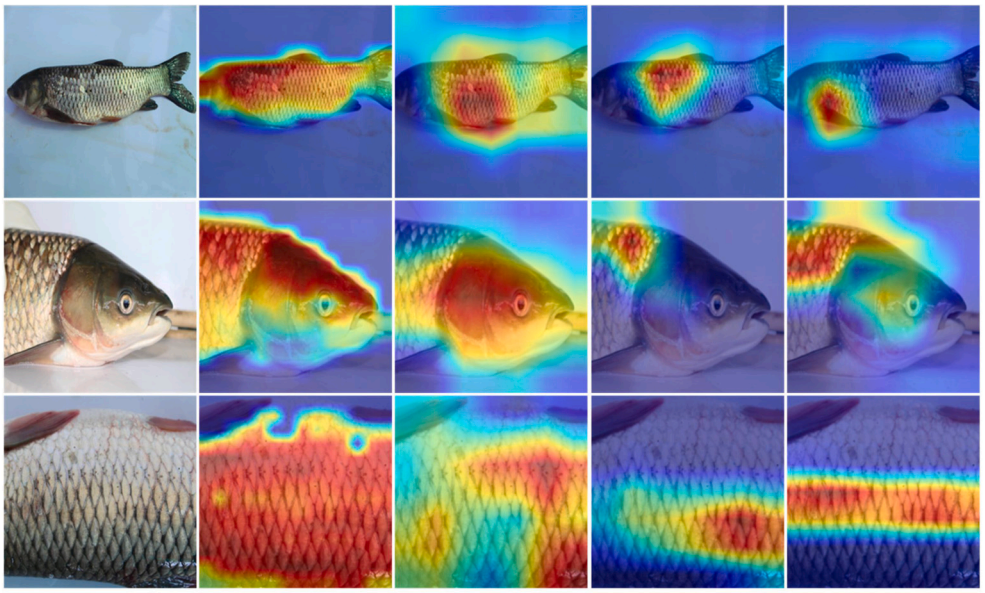

Fig. 9. The Grad-CAM visualizations from Fish-TViT and CNN-based methods for high-resolution freshwater fish classification.

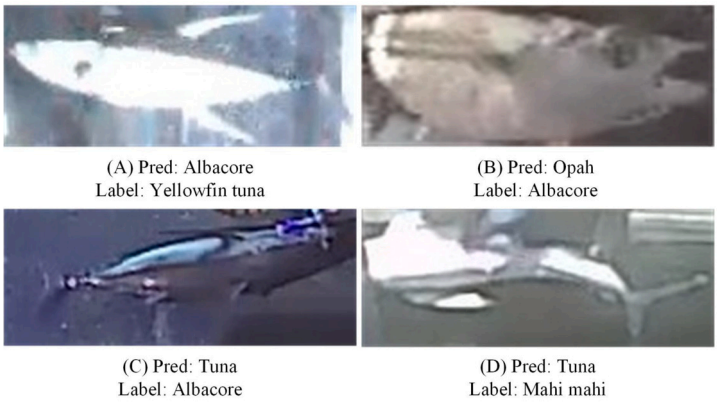

Fig. 10. The examples of error classification of the proposed method on the low-resolution marine fish. The predicted classes and the true labels are below the subplots.

with semantic relevance to classification. In contrast, other CNN-based methods only focus on part of the category region information, and there are still blue regions on the target categories.

The error classified pictures of the proposed Fish-TViT are presented in Figs. 10 and 11. The failure case shown in Fig. 10 is resulted from the blurry and noisy appearance of the target, which makes it difficult to distinguish the images even with the naked eye. In Fig. 11, misclassification occurs due to incomplete freshwater fish images: the upper image has only some scales, and the lower image is the fish's head. These misclassifications are mainly caused by poor test datasets. Due to the difficulty of collecting images of different fishes, we will consider including more datasets in future work to further validate the generalizability of Fish-TViT.

5. Conclusion

In this paper, we proposed a model capable of automatically identifying the multi-water areas fish species, applying the transfer learning and transformer encoder. First, we cropped out the low-resolution marine fish dataset and applied high-resolution freshwater fish, processing and analyzing these datasets. Then, we constructed the Fish-TViT model, using the ViT model pre-trained on the ImageNet dataset, and transferred it to the target dataset, which was used to extract image features. The proposed features were input to the MLP for classifier training. At last, we tried various combinations of patches and image sizes to verify the models. The experimental results showed that the Fish-TViT model can solve the multi-water, multi-resolution fish species classification problem and outperforms the traditional convolution-based model. In aquaculture, our proposed method can accurately identify fish species,

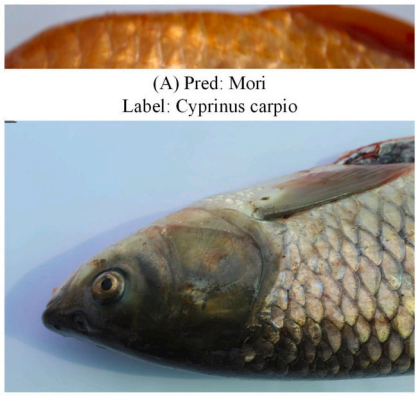

Fig. 11. The examples of error classification of the proposed method on the high-resolution freshwater fish.

improve production efficiency and management, and help promote intelligent aquaculture. However, the total number of trainable parameters in the proposed method is 85.81 M and the number of Flops is 16.86 GFlops. These metrics show that the model is resource-intensive and may pose challenges for deployment and scalability. Besides, the experiment results show that the patch size affects the model accuracy, so how to provide appropriate hyperparameter value is also worth studying. The future research plan focuses on how to obtain more images of fish and proposes new data augmentation techniques to increase the diversity of data. Fish species classification tasks involve multiple levels of information, such as appearance features, behavioral features, and environmental features. Multi-task learning methods can be attempted to learn different tasks simultaneously in order to improve the accuracy and robustness of fish species classification.

Research ethics

We further confirm that any aspect of the work covered in this manuscript that has involved human patients has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript. Written consent to publish potentially identifying information, such as details or the case and photographs, was obtained from the patient(s) or their legal guardian(s). The experiments in this paper, fellow such guidelines, the ARRIVE guideline, the U.K. Animals (Scientific Procedures) Act, 1986, and the National Research Council's Guide for the Care and Use of Laboratory Animals.

Funding

The work was supported by the National Natural Science Foundation of China [12071024, 62076244]; Beijing Digital Agriculture Innovation Consortium Project [BAIC10-2023].

CRediT authorship contribution statement

Bo Gong: Conceived and designed the experiments; Wrote the paper.
Ling Jing: Contributed reagents, materials, analysis tools or data; Wrote the paper.
Ji Shao: Analyzed and interpreted the data.
Kanyuan Dai: Performed the experiments.
Yingyi Chen: Contributed reagents, materials, analysis tools or data.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] R. Froese, D. Pauly, et al., Fishbase, 2010.

[2] FAO, The State of World Fisheries and Aquaculture 2020: Sustainability in Action, The State of World Fisheries and Aquaculture (SOFIA), vol. 2020, FAO, Rome, Italy, 2020, https://www.fao.org/documents/card/en/c/CA9229EN.

[3] L.A. Lundervold, Alexander Selvikvag, An overview of deep learning in medical imaging focusing on mri, Z. Med. Phys. 29 (2) (2019) 102–127, https://doi.org/10.1016/j.zemedi.2018.11.002, https://www.webofscience.com/wos/alldb/full-record/WOS:000469897200003.

[4] Y. Wang, D. Zhang, Y. Liu, B. Dai, L.H. Lee, Enhancing transportation systems via deep learning: a survey, Transp. Res., Part C, Emerg. Technol. 99 (2019) 144–163, https://doi.org/10.1016/j.trc.2018.12.004, https://www.webofscience.com/wos/alldb/full-record/WOS:000459367900009.

[5] Andreas Kamilaris, F.X. Prenafeta-Boldú, Deep learning in agriculture: a survey, Comput. Electron. Agric. 147 (2018) 70–90, https://doi.org/10.1016/j.compag.2018.02.016, https://www.webofscience.com/wos/alldb/full-record/WOS:000428828200010.

[6] S. Zhao, S. Zhang, J. Liu, H. Wang, R. Zhao, Application of machine learning in intelligent fish aquaculture: a review, Aquaculture 540 (99–115) (2021) 736724.

[7] W. Vásquez-Quispesivana, M. Inga, I. Betalleluz-Pallardel, Inteligencia artificial en acuicultura: Fundamentos, aplicaciones y perspectivas futuras, Sci. Agropecuaria 13 (1) (2022) 79–96, https://doi.org/10.17268/sci.agropecu.2022.008, https://revistas.unitru.edu.pe/index.php/scientiaagrop/article/view/4350.

[8] S.S. Nair, F.A. Mon, K. Suthendran, Under water fish species recognition, Int. J. Pure Appl. Math. 118 (7) (2018).

[9] S. Islam, S.I. Bani, R. Ghosh, Content-based fish classification using combination of machine learning methods 1 (2021), MECS Publisher.

[10] P. Urbanova, V. Bozhynov, P. Císař, M. Železný, Classification of fish species using silhouettes, in: I. Rojas, O. Valenzuela, F. Rojas, L.J. Herrera, F. Ortuño (Eds.), Bioinformatics and Biomedical Engineering, Springer International Publishing, Cham, 2020, pp. 310–319.

[11] T.T. Hnin, K.T. Lynn, Fish classification based on robust features selection using machine learning techniques, in: T.T. Zin, J.C.-W. Lin, J.-S. Pan, P. Tin, M. Yokota (Eds.), Genetic and Evolutionary Computing, vol. 387, Springer International Publishing, Cham, 2016, pp. 237–245, http://link.springer.com/10.1007/978-3-319-23204-1_24.

[12] S.O. Ogunlana, O. Olabode, S.A.A. Oluwadare, G.B. Iwasokun, Fish classification using support vector machine 8 (2) (2015) 8.

[13] C. Guisande, A. Manjarrés-Hernández, P. Pelayo-Villamil, C. Granado-Lorencio, I. Riveiro, A. Acuña, E. Prieto Piraquive, E. Janeiro, J. Matias, C. Patti, B. Patti, S. Mazzola, Ipez: an expert system for the taxonomic identification of fishes based on machine learning techniques, Fish. Res. 102 (2010) 240–247, https://doi.org/10.1016/j.fishres.2009.12.003.

[14] A. Tharwat, A.A. Hemedan, A.E. Hassanien, T. Gabel, A biometric-based model for fish species classification, Fish. Res. 204 (2018) 324–336, https://doi.org/10.1016/j.fishres.2018.03.008, https://linkinghub.elsevier.com/retrieve/pii/S0165783618300821.

[15] S. Hasija, M.J. Buragohain, S. Indu, Fish species classification using graph embedding discriminant analysis, in: 2017 International Conference on Machine Vision and Information Technology (CMVIT), IEEE, Singapore, Singapore, 2017, pp. 81–86, http://ieeexplore.ieee.org/document/7878719/.

[16] V. Allken, N.O. Handegard, S. Rosen, T. Schreyeck, T. Mahiout, K. Malde, Fish species identification using a convolutional neural network trained on synthetic data, ICES J. Mar. Sci. 76 (1) (2019) 342–349.

[17] J.C. Ovalle, C. Vilas, L.T. Antelo, On the use of deep learning for fish species recognition and quantification on board fishing vessels, Mar. Policy 139 (2022) 105015, https://doi.org/10.1016/j.marpol.2022.105015, https://linkinghub.elsevier.com/retrieve/pii/S0308597X22000628.

[18] X. Xu, W. Li, Q. Duan, Transfer learning and se-resnet152 networks-based for small-scale unbalanced fish species identification, Comput. Electron. Agric. 180 (2021) 105878, https://doi.org/10.1016/j.compag.2020.105878, https://linkinghub.elsevier.com/retrieve/pii/S0168169920330830.

[19] Z. Ju, Y. Xue, Fish species recognition using an improved alexnet model, Optik 223 (2020) 165499, https://doi.org/10.1016/j.ijleo.2020.165499, https://linkinghub.elsevier.com/retrieve/pii/S0030402620313358.

[20] H.T. Rauf, M.I.U. Lali, S. Zahoor, S.Z.H. Shah, A.U. Rehman, S.A.C. Bukhari, Visual features based automated identification of fish species using deep convolutional neural networks, Comput. Electron. Agric. 167 (2019) 105075, https://doi.org/10.1016/j.compag.2019.105075, https://www.sciencedirect.com/science/article/pii/S0168169919313523.

[21] N.E.M. Khalifa, M.H.N. Taha, A.E. Hassanien, Aquarium family fish species identification system using deep neural networks, in: A.E. Hassanien, M.F. Tolba, K. Shaalan, A.T. Azar (Eds.), Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018, Springer International Publishing, Cham, 2019, pp. 347–356.

[22] P. Hridayami, Fish species recognition using vgg16 deep convolutional neural network, J. Comput. Sci. Eng. 13 (3) (2019) 7.

[23] S. Villon, D. Mouillot, M. Chaumont, E.S. Darling, G. Subsol, T. Claverie, S. Villéger, A deep learning method for accurate and fast identification of coral reef fishes in underwater images, Ecol. Inform. 48 (2018) 238–244, https://doi.org/10.1016/j.ecoinf.2018.09.007, https://linkinghub.elsevier.com/retrieve/pii/S1574954118300694.

[24] A. Salman, A. Jalal, F. Shafait, A. Mian, M. Shortis, J. Seager, E. Harvey, Fish species classification in unconstrained underwater environments based on deep learning, Limnol. Oceanogr., Methods 14 (9) (2016) 570–585, https://doi.org/10.1002/lom3.10113, https://onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10113.

[25] N. Gómez-Vargas, A. Alonso-Fernández, R. Blanquero, L.T. Antelo, Re-identification of fish individuals of undulate skate via deep learning within a few-shot context, Ecol. Inform. 75 (2023) 102036.

[26] R.B. Ariede, C.G. Lemos, F.M. Batista, R.R. Oliveira, J.F. Agudelo, C.H. Borges, R.L. Iope, F.L. Almeida, J.R. Brega, D.T. Hashimoto, Computer vision system using deep learning to predict rib and loin yield in the fish colossoma macropomum, Anim. Genet. (2023).

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, https://doi.org/10.48550/arXiv.2010.11929, 2020.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv:1706.03762, Dec. 2017.

[29] H. Maloy, Echobert: a transformer-based approach for behavior detection in echograms, IEEE Access 8 (2020) 218372–218385, https://doi.org/10.1109/ACCESS.2020.3042337, https://ieeexplore.ieee.org/document/9281296/.

[30] X. Li, F. Li, J. Yu, G. An, A high-precision underwater object detection based on joint self-supervised deblurring and improved spatial transformer network, arXiv:2203.04822, Mar. 2022.

[31] J. Kay, M. Merrifield, The fishnet open images database: a dataset for fish detection and fine-grained categorization in fisheries, CoRR, arXiv:2106.09178, 2021.

[32] S.Z.H. Shah, H.T. Rauf, M. IkramUllah, M.S. Khalid, M. Farooq, M. Fatima, S.A.C. Bukhari, Fish-Pak: fish species dataset from Pakistan for visual features based classification, Data Brief 27 (2019) 104565, https://doi.org/10.1016/j.dib.2019.104565, https://linkinghub.elsevier.com/retrieve/pii/S2352340919309205.

[33] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, S.-M. Hu, Visual attention network, arXiv preprint, arXiv:2202.09741, 2022.

[34] A. Trockman, J.Z. Kolter, Patches are all you need?, arXiv preprint, arXiv:2201.09792, 2022.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[36] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, arXiv:1409.1556, 2014.

[38] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359, https://doi.org/10.1007/s11263-019-01228-7, arXiv:1610.02391.