EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article

Attention ConvMixer Model and Application for Fish Species Classification

Le Thanh Viet[1], Hoang-Minh-Quang Le[1],Vu Van Yem[1], Thi-Thao Tran[1] and Van-Truong Pham[1,*]

[1]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Dai CoViet, Hanoi, Vietnam

Abstract

Exploring the ocean has always been one of the foremost challenges for humankind, and fish classification is one of the crucial tasks in this endeavor. Manual fish classification methods, although accurate, consume significant time, money, and effort, while computer-based methods such as image processing and traditional machine learning often fall short of achieving high accuracy. Recently, deep convolutional neural networks have demonstrated their capability to ensure both time efficiency and accuracy in this task. However, deep convolutional networks typically have a large number of parameters, requiring substantial training time, and the convolutional operations lack attentional mechanisms. Therefore, in this paper, we propose the AttentionConvMixer neural network with Priority Channel Attention (PCA) and Priority Spatial Attention (PSA). The proposed approach exhibits good performance across all three fish classification datasets without introducing any additional parameters, thus demonstrating the effectiveness of our proposed method.

Received on 12 July 2023; accepted on 27 August 2023; published on 06 September 2023

Keywords: Fish species classification, Attention ConvMixer, Priority Channel Attention, Priority Spatial Attention

doi:10.4108/eetinis.v10i3.3562

1. Introduction

The ocean covers more than 70% of the Earth's surface, but we have only explored a fraction of it. This vast underwater world is home to a diverse array of life, including fish, marine mammals, and plants. However, we still know very little about the majority of these creatures. We always need to learn more about the diversity of marine life. The ocean is home to an estimated 2 million to 30 million species, but only about 250,000 have been identified [1]. Fish play an important role in the marine ecosystem, as both predators and prey. They also help to regulate the flow of nutrients through the ecosystem. By studying fish species, we can learn more about how they interact with each other and with other organisms in the ocean. This knowledge can help us to better understand the health of the marine ecosystem as a whole. Besides, some fish species produce compounds that have potential medicinal properties. For example, the venom of some fish species has been shown to be effective in treating cancer. By studying these compounds, we can develop new drugs and treatments for diseases. With the above urgent tasks, the automatic fish classification problem plays a key role.

Fish classification is of paramount importance in the field of biology and fisheries management. It involves the systematic categorization and organization of fish species based on their shared characteristics and evolutionary relationships. The emergence of fish classification has played a significant role in enhancing our understanding of diverse aquatic ecosystems and has practical implications in various areas. Fish classification provides a systematic framework for organizing and categorizing the vast array of fish species found worldwide [2]. Fish classification helps in assessing and documenting the biodiversity of aquatic environments. By identifying and classifying different fish species, scientists gain insights into the distribution patterns, abundance, and ecological roles of fish populations. This information is crucial for monitoring and managing fisheries, conserving endangered species, and preserving overall ecosystem health. Accurate fish classification is vital for effective conservation and management strategies. It allows scientists to identify threatened or endangered species, prioritize conservation efforts, and develop targeted conservation plans. Understanding the relationships between different fish species also helps in assessing

*Corresponding author. Email: yem.vuvan@hust.edu.vn

the impacts of environmental changes, such as habitat loss, pollution, and climate change, on fish populations and ecosystems. Fish classification is crucial for sustainable fisheries and aquaculture practices. It enables fisheries managers to set appropriate catch limits, implement species-specific regulations, and ensure the conservation of economically important fish species. In aquaculture, classification helps in selecting suitable fish species for cultivation, understanding their nutritional requirements, and developing breeding programs to enhance productivity [3].

Fish classification methods can be categorized into manual methods and computer-assisted methods. Manual methods often require a significant amount of time, effort, and human resources. On the other hand, computer-assisted fish classification methods using image processing techniques may not always achieve high accuracy. Manual methods involve experts or trained individuals visually inspecting the fish and identifying their species based on various distinguishing features. This approach can be time-consuming, especially when dealing with a large number of fish samples or complex species identification. Computer-assisted methods, on the other hand, leverage image processing algorithms and traditional machine learning techniques to automate the fish classification process. It is important to consider the trade-off between accuracy and efficiency when choosing a fish classification method. Manual methods can provide higher accuracy but at the cost of increased time and labor. Computer-assisted methods offer faster processing but may sacrifice some accuracy. In recent times, deep learning - a computer-assisted method, has demonstrated its capabilities by ensuring high accuracy and fast processing times. Almost deep learning methods use convolutional neural networks to automatically learn features from fish images. CNNs are able to learn more complex features than traditional methods, and they are more robust to changes in the environment. CNNs have been shown to achieve high accuracy on a variety of fish datasets. One of the challenges of this task is the variability of fish appearance. Fish can vary in size, shape, color, and texture. They can also be found in a variety of different environments, which can affect their appearance. Another challenge is the difficulty of obtaining large, high-quality fish image datasets. Fish are often difficult to photograph, and it can be time-consuming to collect a large enough dataset to train a deep-learning model. Despite these challenges, fish classification is a promising field with a wide range of potential applications. As deep learning methods continue to improve, and as more fish image datasets become available, the accuracy of fish classification is likely to increase. Some commonly CNNs networks such as AlexNet[4], which was the first introduced deep neural network for this purpose. Additionally, more advanced networks like VGG16[5] and ResNet[6] are frequently employed due to their deeper architecture, allowing for improved performance. In addition to these models, there are other neural networks that utilize different types of convolutional operations. For example, Efficient-Net[7] and MnasNet[8] employ standard convolutions along with depthwise convolution and pointwise convolution. This allows for network expansion in terms of width and depth without significantly increasing the number of parameters. Furthermore, although Transformers[9] were originally utilized in natural language processing (NLP), they have also shown promising results when applied to image classification tasks. However, due to their large number of parameters and high complexity, as well as the requirement for substantial amounts of data to achieve high performance, Transformers are not necessarily the top choice for fish classification tasks. The recent introduction of the MLP-Mixer architecture has also achieved highly competitive results in image classification tasks. However, it should be noted that MLP-Mixer[10] and variants like AxialAtt-MLP-Mixer[20] tends to have a large number of parameters. Ultimately, convolution remains the primary choice in vision-related tasks. Convolutional operations involve sliding filters over image regions, allowing the network to extract meaningful features. However, convolutional operations alone cannot inherently learn to prioritize relevant information. To address this limitation, the support of attention mechanisms is necessary.

Attention mechanisms enable the network to focus on the most important regions or features within an image. They allow the model to learn where to allocate its attention and enhance its ability to capture relevant patterns and relationships in the data. By incorporating attention mechanisms alongside convolutional operations, the model can effectively learn to attend to the most salient features and improve its classification performance.

Based on the challenges mentioned earlier in fish classification using deep learning, this paper proposes the following key contributions:

1. We proposed Priority Channel Attention (PCA) for improving channel selection of convolution.
2. We proposed Priority Spatial Attention (PSA) for the purpose of creating interest in important feature areas.
3. With no-increasing parameters, PSA and PCA is integrated to ConvMixer make better classification performance on three datasets: Fish-Gres, Croatian Fish, BD Indigenous Fish.

2. Related works

2.1. ConvMixer

ConvMixer [11] is a vision model that operates directly on patches as input, separates the mixing of spatial and channel dimensions, and maintains equal size and resolution throughout the network. It is similar in spirit to the ViT and the MLP-Mixer, but it uses only standard convolutions to achieve the mixing steps. This makes it much simpler to implement and train, while still achieving state-of-the-art results. ConvMixer consists of a stack of ConvMixer blocks, each of which consists of two steps: First, a patch mixing step, which mixes the patches in the spatial dimension using a standard convolution. Second, a channel mixing step, which mixes the channels in the channel dimension using a standard convolution. The patch mixing step helps to preserve the spatial information in the image, while the channel mixing step helps to learn more abstract representations of the image. ConvMixer has been shown to outperform other SOTA vision models on image classification tasks. For example, on the ImageNet dataset, ConvMixer achieves a top-1 accuracy of 87.7%, which is comparable to the performance of the ViT and the MLP-Mixer. One of the advantages of ConvMixer is that it is much simpler to implement and train than other SOTA vision models. This is because it uses only standard convolutions, which are a well-understood and efficient operation. Another advantage of ConvMixer is that it is more robust to changes in the input size. This is because it maintains equal size and resolution throughout the network, which makes it less sensitive to changes in the input size. Overall, ConvMixer is a simple yet effective vision model that achieves state-of-the-art results on image classification tasks. It is a promising approach for future vision models.

2.2. Attention Mechanism

Attention mechanisms in deep learning models for computer vision play a crucial role in selectively focusing on relevant image regions or features while performing various tasks. These mechanisms are inspired by the human visual system, which allocates attention to specific areas of an image for processing. In the context of computer vision, attention mechanisms are commonly used in convolutional neural networks (CNNs) to enhance the model's ability to recognize and understand visual content. Here are two popular attention mechanisms used in deep learning models for computer vision:

1. Spatial Attention: Spatial attention mechanisms help the model focus on relevant spatial locations in an image. One commonly used spatial attention mechanism is called Spatial Transformer Networks (STN)[12]. In addition, there are methods that utilize spatial attention mechanisms such as the Convolution Block Attention Module (CBAM)[13] or self-attention.

2. Channel Attention: Channel attention mechanisms aim to capture interdependencies between different channels or feature maps in a CNN. One widely used channel attention mechanism is called Squeeze-and-Excitation (SE)[14]. SE modules learn to adaptively recalibrate channel-wise feature responses based on their importance. It consists of a squeeze operation that globally averages the feature maps to obtain channel-wise statistics and an excitation operation that models the channel dependencies through a small neural network. The recalibrated feature maps enhance the important channels while suppressing the less relevant ones, leading to improved discrimination and feature representation.

Both spatial and channel attention mechanisms can be combined and integrated into deep learning models, providing a more refined and focused representation of the visual content. These attention mechanisms help models to better localize objects, handle occlusions, and capture relevant context in images, leading to improved performance in tasks such as object detection, image classification, image segmentation, and visual question answering.

3. Methodology

3.1. The Proposed Model

Our proposed method, named AttentionConvMixer, is depicted in the accompanying Fig. 1. It builds upon the core structure of ConvMixer [11] while incorporating two additional mechanisms: Priority Spatial Attention(PSA) and Priority Channel Attention (PCA). Specifically, we replaced the Depthwise convolutional layer in ConvMixer with Priority Channel Attention (PCA), and we replaced the Pointwise convolutional layer with Priority Spatial Attention (PSA). The key point of this replacement is to enable the model to capture the oscillations that occur after the convolutional layers. We will elaborate on these mechanisms in the following sections.

Priority Channel Attention. Depthwise Convolution is a type of convolution where we apply a convolutional filter to each input channel. In regular 2D convolutions, which are performed across multiple input channels, the filters are deep as the input and allow us to

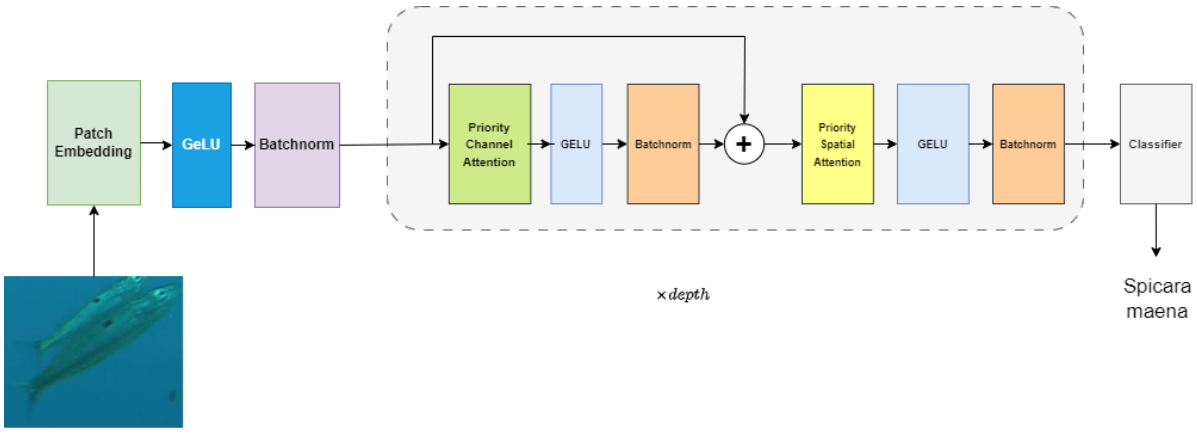

Figure 1. The proposed Attention ConvMixer

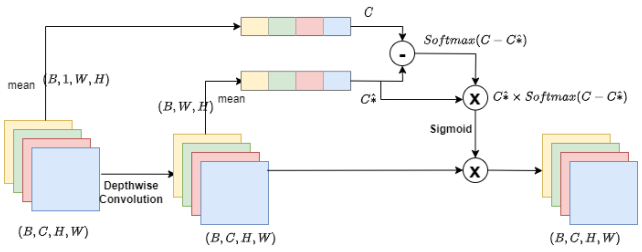

Figure 2. The proposed Priority Channel Attention (PCA)

freely mix the channels to generate each element in the output. This is the main difference between 2D convolution and Depthwise Convolution. Therefore, Depthwise Convolution adjusts features on a per-channel basis. In each filter of the Depthwise Convolution operation, updating the gradient after each iteration helps the filter parameters converge towards optimal values. Each filter contributes to generating distinct features, and these features can be either important or not reflected in the average value of the entire feature. After the Depthwise Convolution layer, the system reorganizes the attended features.

Algorithm 1 The algorithm of Priority Channel Attention

$c \leftarrow reduce(x,'BCHW->B11C', mode='mean')$
$x \leftarrow DepthwiseConvolution(x)$
$c' \leftarrow reduce(x,'BCHW->B11C', mode='mean')$
$RaiseCh \leftarrow Softmax(c'-c)$
$CAScore \leftarrow \sigma[c' \times (1 + RaiseCh)]$
$x' \leftarrow x' \cdot CAScore$

In the current study, we propose the Priority Channel Attention for improving channel selection of convolution. The Priority Channel Attention (PCA), shown in Fig.2 operates based on identifying channels that have been more attended to after the Depthwise Convolution layer. Firstly, the feature $x^{(B,C,H,W)}$ is averaged along the channel dimension, resulting in $c^{(B,1,1,C)}$. Then, a depthwise convolution is performed similarly to the previous step, averaging along the channel dimension and returning c' with shape $(B, 1, 1, C)$. To calculate the channel attention gain and normalize it probabilistically, the difference $c' - c$ is subtracted, and the softmax function is applied to $c' - c$. Additionally, to maintain stability and avoid excessive fluctuations during training, the channel attention coefficients are computed using the following formula:

$$F'_c = \sigma[c' \times (1 + softmax(c' - c)] \quad (1)$$

$$x = x \cdot F'_c \quad (2)$$

Priority Spatial Attention (PSA). Pointwise Convolution is a type of convolution that uses a 1×1 kernel, which is applied to each individual point. This kernel has the same depth as the number of channels in the input image. It can be used in conjunction with Depthwise Convolution to create an efficient convolutional layer called Depthwise Separable Convolution. Therefore, Pointwise Convolution helps to organize the features in the spatial domain by operating on each pixel with a kernel of the same depth. To create interest in important

Algorithm 2 The algorithm of Priority Spatial Attention

$s \leftarrow reduce(x,'BCHW->BHW', mode='mean')$
$x \leftarrow PointwiseConvolution(x)$
$s' \leftarrow reduce(x,'BCHW->BHW', mode='mean')$
$RaiseSp \leftarrow Softmax2d(s'-s)$
$SPScore \leftarrow \sigma[s' \times (1 + RaiseSp)]$
$x' \leftarrow x' \cdot SPscore$

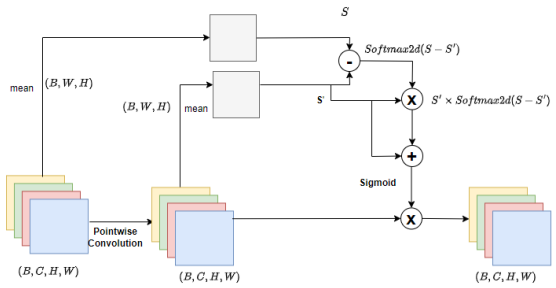

Figure 3. The proposed Priority Spatial Attention (PSA)

feature areas, we propose the Priority Spatial Attention (PSA) shown in Fig.3. PSA relies on identifying the increment of attention in specific pixel regions. Firstly, the feature $x^{(B,C,H,W)}$ is averaged across all channels and returns $s^{(B,H,W)}$. Then, a pointwise convolution (PW) is performed similarly to the previous step, averaging across all channels and returning $s'^{(B,H,W)}$. To calculate the spatial attention gain and normalize it probabilistically, the difference $s' - s$ is subtracted, and the softmax function is applied to $s' - s$. Additionally, to maintain stability and avoid excessive fluctuations during training, the spatial attention coefficients are computed using the following formula:

$$F'_s = \sigma[s' \times (1 + softmax2d(s' - s)] \tag{3}$$

$$x = x \cdot F'_s \tag{4}$$

3.2. Evaluation Metrics

In general classification problems, including fish classification, there are several metrics to evaluate the performance of a model, and each metric represents its own evaluation criterion. The choice of evaluation metric depends on various factors, but the most important consideration is to address the data imbalance issue. When dealing with imbalanced datasets, accuracy alone may not provide an accurate assessment of the model's performance. Instead, it is often recommended to consider metrics such as precision, recall, and F1 score. These metrics take into account the true positive (TP), false positive (FP), and false negative (FN) rates, which are particularly useful in scenarios where one class is significantly underrepresented compared to others. Precision represents the ability of the model to correctly classify positive instances, while recall measures the model's ability to correctly identify all positive instances. F1 score is the harmonic mean of precision and recall, providing a balanced evaluation metric that considers both false positives and false negatives. The fomula of these metrics below here:

- Accuracy : Accuracy is the ratio of the number of correct predictions to the total number of predictions. Interpretation: A high accuracy indicates that the model is performing well overall.

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalPredictions} \tag{5}$$

- F1 Score: The F1 score is a measure of the accuracy and completeness of a model's predictions. It is calculated as the harmonic mean of precision and recall.A high F1 score indicates that the model is performing well in terms of both accuracy and completeness.

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

- Recall : Recall is the ratio of the number of correctly predicted positives to the total number of actual positives. A high recall indicates that the model is good at identifying all of the positive cases.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{7}$$

- Precision : Precision is the ratio of the number of correctly predicted positives to the total number of predicted positives. A high precision indicates that the model is good at avoiding false positives.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{8}$$

4. Experiments

4.1. Datasets

Fish Gres[15]. The Fish-gres dataset is designed specifically for fish species classification and comprises a total of 8 fish species. The number of images available for each species ranges from 240 to 577, as it depends on the random samples collected from traditional markets located in Gresik, East Java, Indonesia. It is worth noting that all the fish species included in this dataset can be readily found in the traditional markets where the dataset was curated. The original acquisition image resolution is 4160x3120 pixels, which is then resized to 390x520 pixels. The fish species present in the dataset are as follows: Chanos Chanos, Johnius Trachycephalus, Nibea Albiflora, Rastrelliger Faughni, Upeneus Moluccensis, Eleutheronema Tetradactylum, Oreochromis Mossambicus, and Oreochromis Niloticus.

Table 1. Fish species and the number of fishes in each category

Croatian Fish		Fish Gres		BDIndigenous2019	
Class	Number of Samples	Class	Number of Samples	Class	Number of Samples
Chromis chromis	106	Chanos Chanos	500	Byen	500
Coris julis female	57	Eleutheronema Tetradactylum	240	Foli	300
Coris julis male	57	Johnius Trachycephalus	240	Koi	380
Diplodus annularis	94	Nibea Albiflora	252	Sing	400
Diplodus vulgaris	111	Oreochromis Mossambicus	331	Sol	120
Oblada melanura	57	Oreochromis Niloticus	564	Sorputi	200
Sarpa salpa	56	Rastrelliger Faughni	544	Taki	390
Serranus scriba	51	Upeneus Moluccensis	577	Tengra	320
Spicara maena	49	-	-	-	-
OSpondyliosoma cantharus	105	-	-	-	-
Symphodus melanocercus	34	-	-	-	-
Symphodus tinca	17	-	-	-	-

Croatian Fish[16]. The Croatian Fish Dataset serves as a valuable resource for fine-grained visual classification (FGVC) tasks, specifically focused on identifying fish species in their natural habitats. This dataset comprises a collection of 794 images, showcasing 12 distinct fish species that were captured in the Adriatic Sea in Croatia. All the images in this dataset portray fishes in their authentic live environments, recorded using high-definition cameras. Marine researchers increasingly employ remote and diver-based videography techniques to study the spatial and temporal variations of habitats and species, including fish assemblages. To handle the substantial amount of data generated by these research methods, computer vision tools are essential for automated processing of the extensive video footage, which often features high fish diversity and density.

BDIndigenous Fish 2019 [17]. The BDIndigenous Fish 2019 dataset comprises a collection of 2610 images representing 8 distinct categories of indigenous fishes found in Bangladesh. The dataset, created by knowaminul, is publicly accessible on GitHub. The 8 categories of fish included in the dataset are Byen, Foli, Koi, Sing, Sol, Sorputi, Taki, and Tengra. This dataset serves as a valuable resource for tasks such as fish species classification and recognition, allowing researchers and developers to train and evaluate models in this domain.

4.2. Implementation Details

We trained our proposed Attention ConvMixer network using the ADAM[18] optimizer with an initial learning rate of 0.001. The learning rate was decreased by a factor of 10 every 50 epochs until reaching 0.00001, which was then kept constant for the remaining 100 epochs. The Cross Entropy Loss function was used as the loss metric. Each dataset was split into 80% for training and 20% for testing. The training time on a workstation with an NVIDIA Tesla T4 16GB GPU ranged from 5 to 20 minutes.

4.3. Results

Our method is compared to several state-of-the-art (SOTA) methods in image classification. The asterisk (*) denotes that the compared method uses pretrained weights. All values in the table are the average values of all test samples used for each evaluated metric. The highest results are shown in bold, and the second highest results are shown in italics.

Table 2. Comparison performance of the proposed method with other networks for fish classification

Methods	Accuracy			Input size
	Croatian Fish	Fish Gres	BDIndigenousFish 2019	
VGG16[5]	0.8144	0.8983*	0.9722	3 × 112 × 112
VGG19[5]	0.7847	0.8751	*0.9846**	3 ×112 × 112
EfficientNet-b4[7]	0.8115	0.844	0.9628	3 × 112 × 112
InceptionNet-v3 [19]	0.8776*	*0.9322**	0.9764*	3 × 299 × 299
ResNet-50[6]	**0.9054**	0.9275	0.9784	3 × 112 × 112
Our Proposed	*0.8995*	**0.9345**	**0.9877**	3 × 112 × 112

From Table 2, it can be observed that despite not using pre-trained weights, the proposed model still achieves the highest effectiveness compared to other models. Notably, among these models are ones that utilize pre-trained weights. This confirms that the proposed model has the ability to learn and extract features very well even without a good initial parameter set.

Table 3. Comparison of total parameters, training time of proposed method with other networks for fish classification

Methods	Training Time 100 epochs (sec.)			Parameters
	Croatian Fish	Fish Gres	BDIndigenousFish 2019	
VGG16 [5]	450	1300	900	134M
VGG19[5]	500	1600	1100	139M
EfficientNet-b4[7]	800	1300	1000	17.5M
InceptionNet-v3[19]	700	2150	1700	25.1M
ResNet-50 [6]	450	1000	700	23.5M
Our Proposed	100	400	300	0.35M

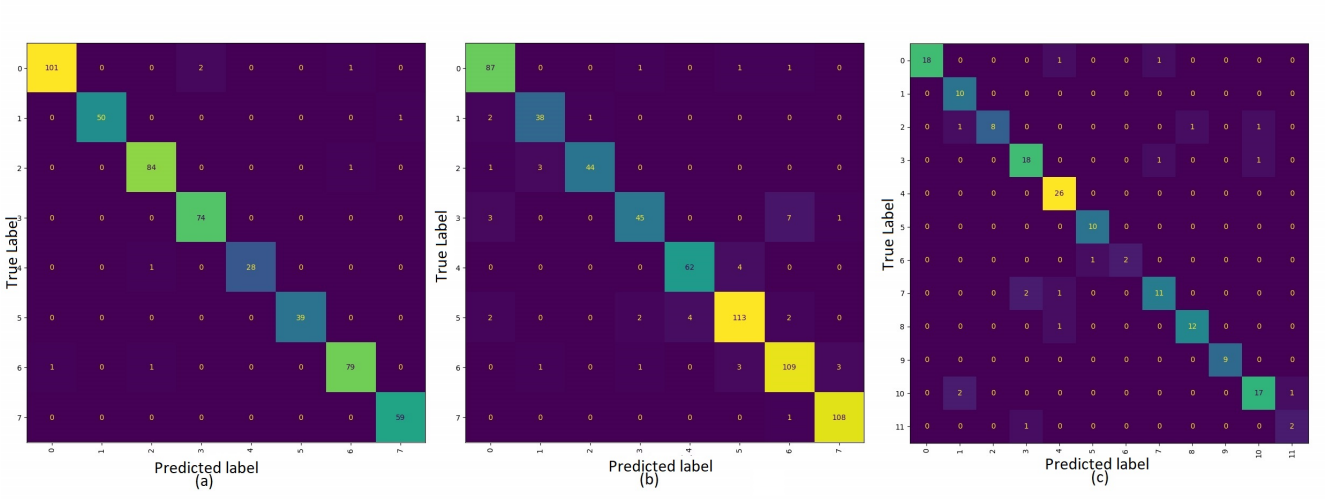

Figure 4. The image alongside depicts the Confusion matrix generated by our proposed method on three datasets. The numbers in the image correspond to the respective classes as follows:
(a) **BDIndigenous Fish 2019**: Byen: 0, Foli:1, Koi: 2, Sing: 3, Sol: 4, Sorputi: 5, Taki:6, Tengra: 7
(b) **Fish Gres**: Chanos Chanos: 0, Eleutheronema Tetradactylum:1, Johnius Trachycephalus:2, Nibea Albiflora: 3, Oreochromis Mossambicus: 4, Oreochromis Niloticus: 5, Rastrelliger Faughni: 6, Upeneus Moluccensis: 7
(c) **Croatian Fish**: Chromis chromis: 0, Coris julis female: 1, Coris julis male: 2, Diplodus annularis: 3, Diplodus vulgaris: 4, Oblada melanura: 5, Sarpa salpa: 6, Serranus scriba: 7, Spicara maena: 8, Spondyliosoma cantharus: 9, Symphodus melanocercus: 10, Symphodus tinca: 11

It is evident from Table 3 that the proposed model has a very small parameter count (around 0.35M parameters). In contrast, other methods have a significantly larger number of parameters, ranging in the tens of millions. As a result, both the training and inference times of the model are reduced by a significant factor while still achieving the highest results among all the mentioned methods.

Fig.4 illustrates the measurement of accuracy using the proposed confusion matrix. For mildly imbalanced datasets like BDIndigenous Fish or FishGres, our method shows excellent results. Even for heavily imbalanced datasets like Croatian Fish, our approach achieves strong performance, despite the fact that the class with the largest training count is approximately 6-7 times greater than the class with the smallest training count.

4.4. Ablation Study

The proposed methods consistently outperform Convmixer. When comparing with other attention mechanisms such as SE or CBAM, it can be observed that the proposed methods achieve competitive results with CBAM and outperform SE, shown in Table 4. Despite not introducing any additional parameters during training, the results are still improved compared to not using any attention mechanism. This demonstrates the strong passive adjustment capabilities of PCA or PSA. For the ConvMixer model with $depth = 12$ and $dim = 128$, the parameter count increases by only around 20,000 - 30,000 params when using CBAM or SE. However, for larger and deeper models, the parameter count increase can reach tens of millions of parameters. Considering the competitive results of PCA or PSA, they can be considered for application in larger and deeper models without significantly increasing the parameter count. However, there is a slight increase in complexity caused by the mechanism that we propose.

We evaluate the Recall, Precision, and F1 scores on the imbalanced Croatian Fish dataset, as shown in Table 5. The results demonstrate that Attention ConvMixer achieves a higher F1 score compared to ConvMixer, and the Recall and Precision scores in the proposed method are also more balanced than ConvMixer. This indicates that the proposed method performs well on a small and imbalanced dataset.

5. Conclusion

In this paper, we have presented and elucidated two proposals: Priority Channel Attention and Priority Spatial Attention. Through testing on ConvMixer, we have demonstrated the excellent effectiveness of our

Table 4. The roles of PCA, PSA on ConvMixer. The highest results are shown in bold, and the second highest results are shown in italics

Methods	Accuracy			Parameters
	Croatian Fish	Fish Gres	BDIndigenousFish 2019	
ConvMixer	0.907	0.984	0.881	352407
ConvMixer + PCA	0.932	0.985	0.893	352407
ConvMixer + PSA	0.927	*0.989*	0.887	352407
ConvMixer + PCA + PSA	**0.934**	0.987	*0.899*	352407
ConvMixer + SE[14]	0.926	0.989	0.887	378615
ConvMixer + CBAM[13]	*0.932*	**0.992**	**0.911**	378119

Table 5. Precison, recall and F1 score for 12 fish species on Croatian Fish dataset

Class	ConvMixer			ConvMixer + PSA + PCA		
	F1	Precisison	Recall	F1	Precisison	Recall
Chromis chromis	0.9744	0.9500	1	0.9474	0.9000	1
Coris julis female	0.8571	0.9000	0.8182	0.8696	1	0.7692
Coris julis male	0.8182	0.8182	0.8182	0.8421	0.7273	1
Diplodus annularis	0.7778	0.7000	0.8750	0.8780	0.9000	0.8571
Diplodus vulgaris	0.9286	1	0.8667	0.9445	1	0.8966
Oblada melanura	0.8571	0.9000	0.8182	0.9524	1	0.9091
Sarpa salpa	0.6667	0.6667	0.6667	0.8000	0.6667	1
Serranus scriba	0.8462	0.7857	0.9167	0.8148	0.7857	0.8462
Spicara maena	0.9600	0.9231	1	0.9231	0.9231	0.9231
OSpondyliosoma cantharus	1	1	1	1	1	1
ymphodus melanocercus	0.8780	0.9000	0.8671	0.8718	0.8600	0.8947
Symphodus tinca	0.5714	0.6667	0.5000	0.6667	0.6667	0.6667

method, even with very few parameters and without the need for pretraining. With promising results on three datasets: Fish Gres, Croatian Fish, and BD Indigenous Fish 2019, with only 0.35M parameters, Attention ConvMixer emerges as a viable choice for fish classification tasks. In the future, we aim to extend these attention mechanisms not only to fish classification but also to various other computer vision tasks related to ocean exploration.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2021.34

References

[1] National Geographic."Ocean."National Geographic, n.d., https://education.nationalgeographic.org

[2] K.V. Ramachandran "The Importance of Fish Taxonomy" (2007).

[3] Peng Zhang, Qingyuan Liu, Yuanming Wang, Kefeng Li, Leilei Qin, Ruifeng Liang, Jiaying Li, "Does drifting passage need to be linked to fish habitat assessment? Assessing environmental flow for multiple fish species with different spawning patterns with a framework integrating habitat connectivity", Journal of Hydrology, Volume 612, Part C, 2022, 128247, ISSN 0022-1694,https://doi.org/10.1016/j.jhydrol.2022.128247.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet classification with deep convolutional neural networks." ,In NIPS, (2012)

[5] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. The 3rd International Conference on Learning Representations (ICLR2015). https://arxiv.org/abs/1409.1556

[6] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90. (2015)

[7] Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[8] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. "MnasNet: Platform-Aware Neural Architecture Search for Mobile." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2820-2828.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, "Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" Proceedings of the 9th International Conference on Learning Representations, (2021)

[10] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X.,Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J. and Others "MLP-Mixer: An all-mlp architecture for vision." Advances In Neural Information Processing Systems. 34 (2021)

[11] Trockman, A., and Kolter, J. Z. (2022). "Patches Are All You Need? " Transactions on Machine Learning Research, 33(1), 28. https://arxiv.org/abs/2201.09792

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu "Spatial Transformer Networks" NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 December 2015 Pages 2017–2025

[13] Woo, S., Park, J., Lee, JY., Kweon, I.S. (2018). CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11211. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_1

[14] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[15] Prasetyo E, Suciati N, Fatichah C. Fish-gres Dataset for Fish Species Classification. Mendeley Data. 2020.

[16] Jäger, J., Simon, M., Denzler, J., Wolff, V., Fricke-Neuderth, K. and Kruschel, C., 2015. Croatian fish dataset: Fine-grained classification of fish species in their natural habitat. Swansea: Bmvc, 2.

[17] Md. Aminul Islam; Md. Rasel Howlader; Umme Habiba; Rahat Hossain Faisal; Md. Mostafijur Rahman "Indigenous Fish Classification of Bangladesh using Hybrid Features with SVM Classifier" IC4ME2 (2019)

[18] Adam: A method for stochastic optimization. D. Kingma, and J. Ba. arXiv preprint arXiv:1412.6980 (2014). (2014)

[19] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[20] Hong-Phuc Lai, Thi-Thao Tran, and Van-Truong Pham. "Axial attention mlp-mixer: A new architecture for image segmentation." *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*. IEEE, 2022.