Underwater species classification using deep learning technique

Dhana Lakshmi MANIKANDAN[1], Sakthivel Murugan SANTHANAM[2]

[1] National Centre for Coastal Research (NCCR), NIOT Campus, Pallikaranai, Chennai, Tamil Nadu, India

[2] Underwater Acoustic Research Laboratory, Department of Electronics and Communication Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

dhanamanikandan23@gmail.com, sakthivels@ssn.edu.in

Abstract: Automated recognition and classification of aquatic species (fish, shrimp etc.) are very useful for studies dealing with the count of species for population evaluation, fish behaviour analysis, monitoring of the ecosystem and understanding the association between species and the ecosystem. Transformers have shown phenomenal success in computer vision problems. However, it demands extensive data for classification tasks. Existing traditional vision transformers necessitate large datasets for heightened accuracy, perpetuating the belief that transformers are data-hungry. This paper aims to dispel this idea by introducing the Amended Dual Attention oN Self-locale and External (ADANSE) mechanism-based vision transformer for classifying underwater (fish) species. In this approach, input images undergo block-tokenization, followed by the application of the proposed attention mechanism, Amended Dual Self Locale and External attention. The Amended dual self-locale attention layer extracts deep feature representations and the external attention mechanism considers the potential relationship among all image blocks. Then, the outputs from both attention mechanisms are further feeding the Multi-Layer Perceptron (MLP) network for species recognition. A proprietary fish database on complex environments is acquired and a self-collected fish database is constructed. This includes the species of *Penaeus vannamei*, *Hypostomus plecostomus*, *Oreochromis niloticus* and its juvenile. When compared to existing ViT networks, the proposed ADANSE network proved to perform better, attaining an accuracy of 90.9% on proprietary datasets and 92% on standard benchmark datasets, emphasising its robust performance even on small-sized images. This highlights the potential of the ADANSE ViT network to address data dependency concerns and achieve competitive accuracy levels in underwater species classification.

Keywords: Vision transformer, Small-sized Datasets, Fish species, Image classification, Self-locale, Attention mechanism.

1. Introduction

Underwater image classification holds extensive potential, encompassing target recognition, debris detection, aquatic monitoring, and various other applications (Cao et al., 2016). Broadly, image classification involves extracting relevant features from an image and categorising its pixels into distinct classes (Jose et al., 2020). Video and image data of marine species are typically obtained from underwater survey equipment such as Remotely Operated Vehicles (ROV), Side Scan Sonar (SSS), and others. Automated recognition and classification of underwater species, including fish and shrimp, prove invaluable for studies involving species population assessment, fish behaviour analysis, ecosystem monitoring, and comprehending the interplay between species and the ecosystem (Liu et al., 2019). In recent decades, deep learning networks have exhibited promising results in computer vision applications like image classification, object localisation, detection, semantic segmentation, and more. The Convolutional Neural Network (CNN), a widely used machine learning approach for image classification, utilises a "Convolution layer" to detect patterns like edges and shapes by convolving with filters in this layer. While CNN performs well, it falls short in extracting the semantic features of an image, attributed to the limited receptor field size in the convolutional layer, which is equivalent to the filter size. In contrast, the vision transformer has gained rapid popularity and emerged as a focal point in modern machine-learning research. It partitions the input image into fixed-size blocks (16 × 16 patches) and establishes contextual relationships through an attention mechanism. Vision transformers can effectively extract long-range dependencies within the series of image blocks, eliciting high semantic features.

Though the vision transformers have made significant progress in computer vision tasks, there still are many aspects that have to be improved. The main contributions of the paper are:

https://doi.org/10.33436/v34i2y202401

- To dissipate the myth that the transformers are "data-hungry". This paper proposes an Amended Dual Attention oN Self-locale and External (ADANSE) mechanism based vision transformer for automated classification of fish species;
- A proprietary fish database (4 categories) on complex environments is acquired and a self-collected fish database is constructed;
- The proposed network is compared with other existing vision transformer networks and the outcomes reveal that the proposed network achieves competitive trade-offs between accuracy and complexity on different image resolution datasets. i.e.) the proposed ADANSE ViT exhibits an accuracy of 90.9% on proprietary and 92% on standard benchmark datasets even on 32 × 32 sized images.

The rest of the paper is organised as follows. In Section II, the existing techniques on vision transformer-based networks are discussed. In Section III, the datasets considered and the proposed methodology are discussed with special emphasis on the ADANSE mechanisms. Section IV presents the experimental quantitative analysis by comparing the performance of the proposed network with the other state-of-the-art networks. In Section V, the conclusion and future research directions are outlined.

2. Related works

Many researchers have worked on image classification tasks using Vision Transformer (ViT). Initially, Vaswani et al., (2017) proposed a transformer with an associated encoder and decoder implemented through an attention mechanism for machine translation tasks. Later, Dosovitskiy et al., (2020) have advocated vision-based transformer networks for visual recognition tasks. They split the entire image into fixed-size patches and append the position embeddings to them. Then it is provided as input into the encoder containing multi-head attention and Multi-Layer Perceptron (MLP) to obtain the classification token. Chen et al., (2021) have proposed a deformable patch-based transformer for image recognition and object detection tasks. They divide the image into different-sized patches in the view to maintain the semantic information of the image and achieve good accuracy in classification. Guo et al., (2022) have designed an external attention module-based bi-linear transformer for deep feature extraction. They have performed analysis on visual recognition, object localisation, semantic segmentation, point cloud synthesis etc., and have shown better performance compared to that of others. Liu et al., (2021) have presented a Swin transformer that computes the feature representation using Shifted windows. It achieves higher efficiency through the construction of hierarchical feature vectors by integrating the image patches in deeper layers.

Generally, the ViT models demand huge data to perform image classification tasks. To address this, Lee et al., (2021) have put forward a vision transformer model for a small-sized dataset. They have proposed shifted patch tokenisation to overcome the locality inductive bias and have acquired the features from scratch on small-sized data. They spatially divide the whole image into several shifted directions and concatenate the same with the source image to traverse through the encoder and generate the classification token. Wu et al., (2021) have introduced convolution into ViT called Convolutional Vision Transformer (CvT). Through modification of convolutional token embedding and leverage of convolution, appropriate characteristics of CNNs have been projected onto ViT (i.e., shift, scale and distortion invariance) while conserving the advantages of Transformers (i.e., dynamic attention, global context, and better generalisation). Touvron et al., (2021) have augmented the CNNs with attention-based global maps to attain non-local reasoning. Here, they have replaced the pooling layer with an aggregation layer. This acts as a distinct transformer block that loads the behaviour of the patches in the classification decision. They have introduced the aggregation layer with basic patch-based CNN that is parameterised by two parameters namely width and depth. Hassani et al., (2021) have developed a Compact Convolution Transformer (CCT) by replacing patch tokenisation with the convolution process. Thus, their model eradicates the usage of class tokens and positional embeddings through a novel sequence pooling strategy and shows good precision even on small-sized datasets.

Wang et al., (2022) have provided a basic alternative attention mechanism through the incorporation of shift operation in ViT. Muthuraman & Santhanam, (2022) have proposed a hybrid restoration network-weighted filter for pixel regularisation thereby achieving complete edge detail, consistent brightness and good contrast on underwater images. Li & Chen, (2021) have designed UDA-Net (Underwater Densely Attention - Network) with a feature-extraction attention layer. During training, UDA-Net blends a variety of information and extracts the channel attention maps to obtain the weighted interest points. Lakshmi et al., (2021) have developed a modified underwater light attenuation prior (MULAP) model based on image contrast and sharpening filter to enhance the degraded underwater image. Huang et al., (2022) have introduced an Adaptive Group Attention (AGA) module network to select the complementary channels based on the attention parameters. It is used on the swin transformer as an end-to-end underwater image enhancement network. Peng et al., (2023) have designed an U-shaped transformer with a Channel-wise Multi-Scale Feature Fusion Transformer (CMSFFT) for Underwater Image Enhancement (UIE) tasks. Li et al., (2022) have presented high-precision Underwater Object Detection (UOD) based on self-supervised deblurring and enhanced transformer modules. Due to the limitations of different perspective images, the network works on perspective transformation in the view to enrich the image features within the network. Qu et al., (2022) have proposed a Multi-Color Convolutional and Attentional Stacking Network (MCCA-Net) that fuses image features from the attention module and convolution layer. Since the variants of ViT are highly dependent on large-sized data, they lack performance on small-sized data. To the best of our knowledge, the previously existing techniques lack inductive bias and are highly “data-hungry”. To address this issue, this paper proposes an ADANSE mechanism for vision transformers towards achieving automated classification of fish species, even on small-sized datasets.

3. Proposed methodology

3.1. Data Pre-processing

The data is acquired through an equipment called “Sofar Trident UW drone”. It is composed of a built-in camera with 6 Light Emitting Diodes (LEDs), 25 m Tether and a Joystick Controller with an Android display. This display helps the moderator to navigate the drone in the required direction and capture the underwater scenes. The data is acquired from different caged aquaculture tanks with varying atmospheric lights. The location information of the collected data is described in Table 1.

Table 1. Location Description of Proprietary dataset

Date	**Survey Area**			**Depth (m)**
	Location	**Position**	**Notation used**	
December 9, 2020	Paraprofessional Institute of Aquaculture Technology, Muttukadu	12°48'49.50" N 80°14'34.91" E	L1	6-7
August 2, 2021				
August 14, 2021	Fishery Department Office, Chengalpattu	12°43'31.08" N 79°57'4.18" E	L2	7

The captured video is recorded in MP4 format with 720-pixel resolution at 30 frames per second. As a pre-processing step, the video is converted into frames of images and duplicate frames are filtered using hashing technique (Zendel & Zinner, 2021) to avoid overfitting of the network. Then, image labelling is done manually and the images are categorized into their respective species classes. Finally, the proprietary fish database is compiled to dissipate the myth that the transformers are "data-hungry". A sample of the proprietary dataset considered for building the network is shown in Figure 1. Due to the properties of water and its impurities, the acquired data exhibits haze, colour deviations, non-uniform illumination, blurred details and low-contrast. Thus, the acquired data has undergone the image enhancement process (Muthuraman & Santhanam, 2022) to obtain the visibility improved images.

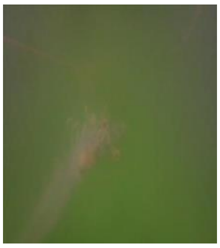
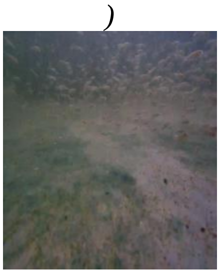
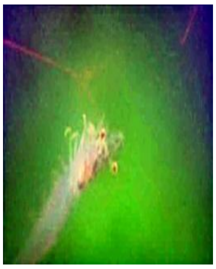
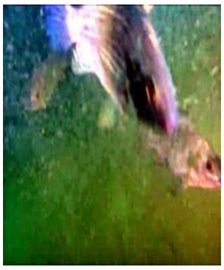
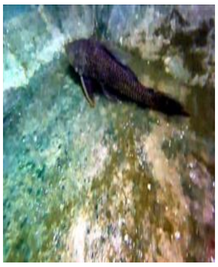
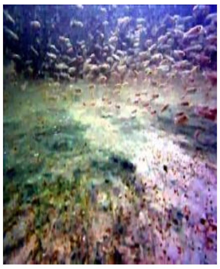
Image index	**I1**	**I2**	**I3**	**I4**
Common name	White-leg Shrimp	GIFT Nile Tilapia	Common Pleco	Fry Nile Tilapia
Scientific name	*Penaeus vannamei*	*Oreochromis nilotics*	*Hypostomus plecostomus*	*Oreochromis nilotics (juvenile)*
				
	Degraded raw images of species			
Sample Data				
	Visibility improved images			
No. of Images	178	177	170	176

Figure 1. Proprietary Dataset Considered

3.2. Proposed ADANSE Vision transformer

A standard transformer (Fu, 2022) comprises of Multi-Head Attention (MHA) layer followed by Layer Norm (LN) and Multi-Layer Perceptron (MLP) block with residual connections. Initially, the input feature is processed as Query Q, Key K and Value V with the help of MLP. Later, the encoder is processed according to Eqn. (1) and Eqn. (2).

$$MHA\left(Q,K,V\right)=concatenate\left(h_1,\ldots..h_n\right)W^O \tag{1}$$

$$h_i = Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{2}$$

Motivated by the scaling of the transformer in Natural Language Processing (NLP), the ADANSE Vision transformer (ViT) is proposed. To process 2D images, the proposed ADANSE ViT considers the input as a 1D sequence of token embeddings. The input image $I \in R^{H\times W\times C}$ is reshaped into a series of flattened 2D blocks $I_{pe} \in R^{N\times\left(P_a^2 .C\right)}$. Here (H,W) is the height and width, i.e., the resolution of the source image. Further, C refers to the number of channels (RGB image, $C=3$), $\left(P_{a,}P_{a,}\right)$ is the resolution of each block and $N = HW / P_a^2$ is the number of resultant blocks. The higher the number of image blocks, the higher the image resolution and thus, the higher the memory consumption. It uses static latent vector size L throughout the layers. The flattened blocks are mapped to n dimensions with a linear projection as expressed in Eqn. (3) through Eqn. (7).

$$\sqrt{d_k} \rightarrow Y_0 = \left[I_{class}; I_{P_a}^{1} E;\ I_{P_a}^{2} E\ ; \ldots ; I_{P_a}^{N} E \right] + E_{pos},$$

$$E \in R^{(P_a^{\,2}.C)\times L},\ E_{pos} \in R^{(N+1)\times L} \tag{3}$$

$$Y_m^{'} = ALSA\left(LN\left(Y_{m-1}\right)\right) + Y_{m-1}, \qquad m = 1\ldots n \tag{4}$$

$$Y_m = MLP\left(LN\left(Y_m^{'}\right)\right) + Y_m^{'}\ , \qquad m = 1\ldots n \tag{5}$$

$$Y_m^{'} = EA\left(LN\left(Y_{m+1}\right)\right) + Y_{m+1}, \qquad m = 1\ldots n \tag{6}$$

$$Z = LN(Y_n^0) \tag{7}$$

The resultant of the projections is the patch embedding that retains the positional information. Each vector is independent and the sequence of embedded vectors is given as an input to the encoder. The existing self-attention layer lacks locality inductive bias, i.e., image pixel values are locally correlated, and their correlation maps are translation-invariant. This makes the standard ViT demand more data for efficient classification. In order to dissipate the myth that the transformers are "data-hungry", the proposed ADANSE ViT encoder consists of three modules: Amended Locale Self Attention (ALSA), External Attention (EA) and LN followed by classification head MLP block to the respective attention modules. ALSA and EA mechanisms can extract more essential semantic features of the images than existing CNNs and Vision transformer networks. The module makes use of Query Q, Key K and Value V . Initially, the cosine similarity between Q and K is calculated using the dot product, and their resultant is divided by the square root of the key dimension. The cosine similarity between two vectors is calculated as in Eqn. (8).

$$Cosine\left(x_i, y_i\right) = \frac{x\ .y}{x * y} \tag{8}$$

In Eqn. (8), $x\,.y$ is the dot product and $x * y$ is the cross-product of two feature vectors. Further, x and y denote the length of the two vectors. The proportion between the dot product and key dimension helps avoid small gradients in the probability function. In order to produce the attention weights, the softmax probability function is applied to the dot product as in Eqn. (9).

$$Locality\,Self\ \ Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

Here, Q, K and V are taken from the same input. Hence, the dot product may result in huge self-token gatherings rather than inter-token relations. To avoid this, the dot product of diagonal edge features is masked. This masking intensifies the attention rank between different tokens and constructs a sharper attention rank. Each edge masking focuses on the $-\beta$ on diagonal components of E . Thus, it compels the attention module to pay more attention to local features in an inter-relation manner. The proposed edge diagonal masking is defined in Eqn. (10).

$$E_{l,m}^{M}\left(x\right) = \begin{cases} E_{l,m}\left(x\right) + y & \left(l \neq m\right) \\ -\beta & \left(l = m\right) \end{cases} \tag{10}$$

In Eqn. (10), $E_{l,m}^{M}\left(x\right)$ denotes the masked similarity matrix of each component. The scaling ratio factor is a constant in the existing attention module, whereas, in the proposed module, it is a learnable temperature parameter that can modify the softmax probability function. Based on Eqn. (8), the ALSA with edge diagonal masking and learnable temperature modification is applied according to Eqn. (11).

$$ALSA(x) = softmax\left(E^{M}(x)/\rho\right)xE_{v} \tag{11}$$

In Eqn. (11), E_v is the linear projection value, and ρ is the learnable temperature. The amended dual self-locale attention layer extracts and appraises the deep feature representation of diagonal features at every position. This is done by calculating the weighted sum of features using pair-wise affinities across all positions within a single image. Then, the resultant from ALSA is fed into the Layer Norm for normalisation and then passed onto the External attention mechanism module. The external attention mechanism helps in considering the potential correlation between all blocks of images. EA makes use of dual cascaded linear layers and a normalisation layer without the multi-head mechanism. It computes the attention between the local diagonal features and the memory units. $(\beta)_{x,y}$ is the similarity between the x-th pixel and y-th row of learnable parameter ρ of the input. This behaves as a memory for the entire training data. P is the attention map extracted from the prior knowledge of the trained dataset and it is normalised in the same manner as that done in the ALSA mechanism. Later, the update is done to the input local features from ρ in P as in Eqn. (12) and Eqn. (13).

$$P = (\beta)_{x,y} = Norm\left(F.\rho^{T}\right) \tag{12}$$

$$Feat_{out} = P.\rho \tag{13}$$

It makes use of dual memory units M_x and M_y as K and V to enhance the network's ability. This amends the EA mechanism as shown in Eqn. (14) and Eqn. (15).

$$P = (\beta)_{x,y} = Norm\left(F.\rho_{x}^{T}\right) \tag{14}$$

$$Feat_{out} = P.\rho_{y} \tag{15}$$

A Layer Norm is applied before and after each of the attention mechanisms (ALSA, EA). Then, the output of both the attention mechanisms after applying Layer Norm are headed to the tri-layer Multi-Layer Perceptron (MLP) network, which is composed of two fully connected layers with a GELU (Gaussian Error Linear Unit) non-linearity function. Since, batch normalisation is dependent on batch size, it cannot be applied for small-sized datasets. This can be overcome by the usage of Layer Norm as it is independent of the batch size. Through incorporation of Layer Norm, all neurons in the specific level will have an identical distribution across all the features for a given source. For instance, if the input has Q features, then it is a Q-dimensional vector. If there are F elements in a batch set, the normalisation is carried along the length of the Q-dimensional vector and not across the batch size F. It normalises the output vector obtained from layer $k-1$. The Layer Norm is applied as presented in Eqn. (16) and Eqn. (17).

$$x_i = \frac{x_i - \mu_l}{\sqrt{\sigma_l^2}} \tag{16}$$

$$y_i = LN(x_i) = \gamma.x_i + \beta \tag{17}$$

In Eqn. (16), μ_l and σ_l^2 are the mean and variance of the source input. With the help of these parameters, the neuron on each layer is normalised in an independent manner. Thus, LN helps in normalising each of the inputs in the set independently across all the edge diagonal features and other global features. The methodology involved in the proposed ADANSE network is shown in Figure 2. GELU is an activation function that weights the source by their percentile rather than gates inputs by their sign as in ReLUs (Rectified Linear Unit) $x1_{x>0}$. The GeLU function is expressed as in Eqn. (18).

$$GELU(x) = x.P(X \le x) = x\varphi(x) = x.\frac{1}{2}\left[1 + \mathrm{erf}\left(x/\sqrt{2}\right)\right] \tag{18}$$

GeLU is different from ReLU in that it does not have any limit range like upper bound or lower bound. For example, the ReLU function outputs zero in the negative input limit whereas the GeLU is much smoother in this region. It is differentiable in all limits and permits to have gradients even in the negative limit. Further, the ReLU function activates some of the neurons to be zero if their condition is not satisfied. This makes the GeLU more beneficial than the ReLU. A MLP layer and a residual connection after each of the attention module layers are used for the species recognition task. It is also found that the existing self-attention layer on the vision transformer lacks the locality inductive bias and this is overcome by the proposed ADANSE ViT network.

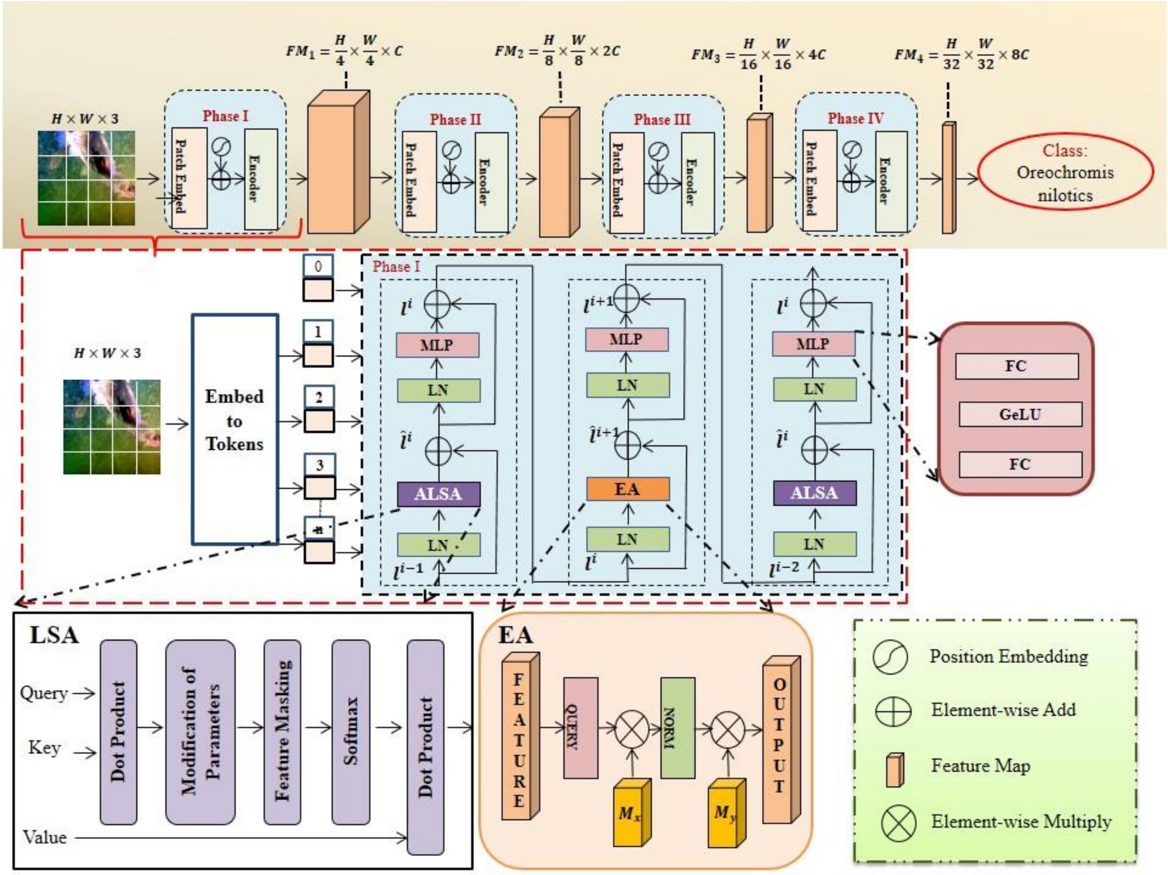


Figure 2. Methodology involved in proposed ADANSE ViT Network

4. Experiments

4.1. Training regime

The model configuration of ADANSE ViT is as follows: In the case of experimentation with ADANSE ViT, the number of heads is set to 4 and the size of the blocks of the embedding layer is fixed to 6. In addition, the regularisation technique called stochastic depth (Fu, 2022) is applied such that it randomly drops a set of layers. It is similar to Dropout but it takes control over chunks of layers rather than the distinct nodes that exist inside the layer. It is used before the residual blocks of the transformer network. Furthermore, AdamW (Steiner et al., 2021) is used as the optimiser to update the weights during the training. It is different from the Adam optimiser (Li et al., 2023). AdamW produces better training loss and therefore the network generalises much better than the models trained with the Adam. Here, the regularisation parameter called weight decay (Dong et al., 2022) is set to 0.0001 as a small penalty to the loss function. Further, the batch size is set to 32, and the learning rate is set to 0.001. This configuration has been determined experimentally. To improve the classification accuracy and maintain the stability of the model, the epochs and the learning rate are adjusted during the training process. The model was executed on a server with SUPERMICRO make/model, equipped with an Intel Xeon processor and ample RAM (128 GB DDR4) in remote access. The storage configuration includes a combination of HDDs and SSDs for different storage needs. In terms of GPUs, the system features four NVIDIA GeForce RTX 2080 cards, each with 11 GB of GDDR6 VRAM, providing significant computational power for GPU-accelerated tasks. The deep learning model was trained using the above hardware configuration for a total duration of 50 epochs. Each epoch took approximately 1.29 s to complete, resulting in a total training time of 64.5 s with an inference time of 0.16 s. The model architecture comprised of 8 transformer layers with image patch size of 6×6 and 6 M parameters. The deep learning model was executed within a Conda environment on an Ubuntu terminal, utilising the Keras framework. The proposed ADANSE network is a lightweight model as it incurs low complexity and has been trained on a CPU (AMD 5900X) and still yields high performance. Moreover, the ADANSE network characterises a fewer number of parameters comparatively with the existing vision transformers. Even if the researchers do not have access to high-end hardware systems, one can democratise the vision transformer and make it more available to anyone who needs to exploit it.

4.2. Ablation study

In order to validate the proposed modules in the entire network model, experiments are conducted on both standard benchmark datasets and proprietary datasets. The benchmark Dataset considered in this work is the WildFish (Zhuang et al., 2018) Dataset. It consists of 1000 fish categories with 54,459 unconstrainted images. The ablation experiments of ADANSE ViT are carried out on both datasets. Since the resolution of images will have an impact on classification accuracy, it is also experimented on different resolution images (32 × 32, 64 × 64, 96 × 96 and 128 × 128) to validate the proposed network on downsampled and upsampled images. It is observed that the incorporation of the proposed attention module facilitates the learned attention maps to focus on both foreground and background species in an efficient manner. The Layer Norm technique produces a significant improvement in EA attention and also makes better improvements on ALSA. Table 2 illustrates the pseudocode on the proposed ADANSE ViT model.

Table 2. Pseudocode on proposed ADANSE ViT network

```
Pseudocode: Proposed ADANSE ViT network

Input: Visibility improved source species Image, $I(x,y)$ of size $m \times n$
Output: Classified output index, $Classified_{species}$
begin
{
stage: Block-Embedding
        sub-stage: Reshape the source image into series of flattened 2D blocks
                $I_{pe} \in R^{N \times (P_a^2 . C)} \leftarrow Reshape(I \in R^{H \times W \times C});$
        // $(H,W)$ → Height and Width of the source image; $C$ → Number of channels
          ($RGB$ image, $C = 3$)
        // $(P_{a,} P_{a,})$ → Resolution of each block
        end sub-stage
      sub-stage: Mapping of flattened blocks to a linear projection
      $N \leftarrow HW / P_a^2$
```

```
        // $N \rightarrow$ The Resultant number of blocks
        $Y_0 \leftarrow \left[ I_{class}; I^1_{P_a}E;\ I^2_{P_a}E\ ;\ldots.; I^N_{P_a}E \right] + E_{pos},\quad E \in R^{P_a{}^2 .C)\times L}, E_{pos} \in R^{(N+1)\times L}$
        // $L \rightarrow$ Static latent vector size
        end sub-stage
end-stage
stage: Compute the Amended Locale Self Attention (ALSA) map
        $Locality\,Self\;\;Attention(Q,K,V) \leftarrow softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
        // $Q \rightarrow$ Query; $K \rightarrow$ Key; $V \rightarrow$ Value; $\sqrt{d_k} \rightarrow$ Square root of the key dimension
        sub-stage: Calculate the edge diagonal mask
        $E^M_{l,m}(x) \leftarrow E_{l,m}(x) + y\ ;\quad (l \neq m)$
        // $E^M_{l,m}(x) \rightarrow$ Edge diagonal mask with similarity matrix of each component
        end sub-stage
        sub-stage: Estimate ALSA with edge diagonal masking & learnable temperature modification
        $ALSA(x) \leftarrow softmax\left(E^M(x)/\rho\right)xE_v$
        // $E_v \rightarrow$ learnable linear projection
        // $\rho \rightarrow$ learnable temperature parameter
        end sub-stage
end-stage
stage: Perform Layer Normalization
        $\mu_l \leftarrow \frac{1}{Q}\sum_{i=1}^{Q} x_i$
        $\sigma_l^2 \leftarrow \frac{1}{Q}\sum_{i=1}^{Q}(x_i - \mu_l)^2$
        $x_i \leftarrow \frac{x_i - \mu_l}{\sqrt{\sigma_l^2}}$
        $LN(x_i) \leftarrow \gamma . x_i + \beta$
        // $\mu_l$ and $\sigma_l^2 \rightarrow$ mean and variance of the features
end-stage
stage: Apply GeLU activation function in MLP
        $GELU(x) \leftarrow x.\frac{1}{2}\left[1 + \mathrm{erf}\left(x/\sqrt{2}\right)\right]$
end-stage
stage: Determine the External Attention Map
        $P \leftarrow Norm\left(F.\rho_x{}^T\right)$
        $P \rightarrow (\beta)_{x,y}$
        $Feat_{out} \leftarrow P.\rho_y$
        // $P \rightarrow$ Attention map extracted; $\rho_x$ and $\rho_y \rightarrow$ Memory units; $Feat_{out} \rightarrow$ Output features
end-stage
        /** Similar to ALSA stage, Layer normalization and GeLU activation function is again carried out after the EA layer and again computation of ALSA module is performed **/
stage: Apply GeLU activation function of ALSA II module
        $GELU\left(x_2\right) \leftarrow x.\frac{1}{2}\left[1+\text{erf}\left(x/\sqrt{2}\right)\right]$
        Classified output index, $Classified_{species} \leftarrow GELU\left(x_2\right)$
end-stage
}
end
```

4.3. Results and discussions

In order to classify the species images, the evaluation is carried out with the proposed ADANSE ViT network and other existing ViTs as shown in Table 3.

Table 3. Performance comparison in terms of accuracy (%) of different ViT models on both Proprietary and Wildfish Datasets

Image size	**32 × 32**		**64 × 64**		**96 × 96**		**128 × 128**	
Existing ViT	**Proprietary (%)**	**Wild Fish (%)**	**Proprietary (%)**	**Wild Fish (%)**	**Proprietary (%)**	**Wild Fish (%)**	**Proprietary (%)**	**Wild Fish (%)**
CCT (Hassani et al., 2021)	69.1	54.2	70.56	57.1	71.56	57.1	71.7	60.4
EA (Guo et al., 2022)	73.4	67.5	76.82	68.7	76.99	68.7	74.1	70.1
ViT (Dosovitskiy et al., 2020)	80.2	72.4	82.73	76.8	83.21	76.8	85.1	76.7
Swin transformer (Liu & Chen, 2021)	82.9	81.4	83.42	82.1	86.19	82.1	86.6	84.8
ViT for Small Dataset (Lee et al., 2021)	84.6	83.7	86.29	83.9	86.84	83.9	87.0	82.4
ViT without attention (Wang et al., 2022)	24.1	40.4	29.18	46.1	31.47	58.7	36.8	59.6
Proposed ADANSE ViT	**90.9**	**92.8**	**91.86**	**93.1**	**91.99**	**93.8**	**92.3**	**93.9**

It includes CCT (Compact Convolution Transformer) (Hassani et al., 2021), EAT (External Attention Transformer) (Guo et al., 2022), Vision Transformer (ViT) (Dosovitskiy et al., 2020), Swin transformer (Liu & Chen, 2021), ViT for small dataset (Lee et al., 2021) and ViT without attention (Wang et al., 2022) on both proprietary data and benchmark data WildFish (Zhuang et al., 2018) characterising different image resolutions .

The notable observations include: ViT without attention (Wang et al., 2022) consistently performing poorly compared to other models. The proposed ADANSE ViT consistently demonstrates the highest accuracy percentages across all image sizes and datasets, showcasing its effectiveness in underwater species classification. On the impact of image size, as image size increases, there is an improvement in accuracy for most models (Si et al., 2023), indicating the significance of larger input dimensions (Zeng et al., 2023). It is also observed that the proposed ADANSE ViT has shown an accuracy of more than 90%, this score is the highest when compared with that of the other existing vision transformer network models for both datasets. It is also revealed that the proposed network achieves competitive trade-offs between accuracy and complexity. It is claimed that the contribution of each ALSA and EA produces a synergy. So, the performance of ADANSE ViT progresses and the computational cost also decreases. For example, the proposed attention modules on proprietary data and WildFish data have improved the performance by +6.3% and +9.1%, respectively, compared to the respective module(s) used by ViT for small Dataset (Lee et al., 2021) on 32 × 32 resolution images.

On overall comparative analysis, the proposed ADAN-SE ViT surpasses existing ViT models, achieving accuracy percentages ranging from 90.9% to 93.9% across all image sizes and datasets. ViT without attention (Wang et al., 2022) consistently lags behind other models, highlighting the importance of attention mechanisms in ViT models. From Table 3, it is also revealed that the ViT shift module has shown inferior throughput when compared to that of others. On average, the proposed ADANSE ViT has outperformed the other ViTs with the superior improvement of +27.33% and +36.94% for both datasets on 32 × 32 resolution images. In Table 3, it is observed that an increase in image resolution improves classification accuracy. For 32 × 32 images, the proposed ADANSE network achieves an accuracy of 90.9% for proprietary data and 92.8% for the standard benchmark WildFish dataset. Similarly for 64 × 64, 96 × 96 and 128 × 128 resolution images, it attains an accuracy of more than 91% for both datasets. In the case of the proprietary dataset (~701 images), the ADANSE ViT is trained and evaluated on the divided (train: validation: test) ratio of 60: 10: 30. Thus, the training set and a validation set comprises of 490 (70%) images and the testing set is composed of 211 (30%) images. Finally, the ADANSE ViT has shown the best results for species identification in a generalised manner. In summary, the proposed ADANSE ViT demonstrates superior performance, showcasing its potential for precise underwater species classification across varying image sizes and datasets.

5. Conclusion

The paper introduces the ADANSE ViT network for the classification of fish species. The network comprises an amended dual self-locale and external attention layer referred to as ALSA and EA, designed to extract deep feature representations. The ALSA module captures edge diagonal features across all locations, adjusting learning parameters for enhanced extraction. The EA mechanism treats memory units as dictionaries, fostering the learning of distinct features and uncovering correlations between image blocks. Subsequently, both attention mechanisms feed the Multi-Layer Perceptron (MLP) network for species recognition. Moreover, the ADANSE network has a comparatively smaller number of parameters. Even if the researchers do not have access to high-end hardware systems, one can democratise the vision transformer and make it more available to anyone who needs to exploit it. Experiments on different marine species datasets (own proprietary dataset and Wild Fish datasets) acquired from scratch at varying atmospheric light demonstrate the robust and effective performance of the proposed network. The experiments and comparisons prove that the proposed method outperforms the state-of-the-art methods with an overall accuracy improvement of +27.33% and +36.94% for both datasets on 32 × 32 images. On average, the proposed ADANSE network achieves an accuracy of 92% for both datasets on different resolution images. Future work may extend the application to the localisation of aquatic species using underwater video and other modalities.

Author contribution statement

Data processing and analysis and the first draft of the manuscript were written by Dr. Dhana Lakshmi Manikandan with the guidance of Dr. Sakthivel Murugan Santhanam. All authors read and approved the final manuscript.

Acknowledgement

This work is a part of a project funded by the Department of Science and Technology (DST) under SSTP. Grant No: DST/SSTP/Tamilnadu/102/2017-18. The authors wish to thank Dr. N. Felix, Director and Dr. Mir Ishfag Nazir, Assistant Professor, Paraprofessional Institute of Aqua-culture, Muttukadu for their support in data collection. We also thank SSN College of Engineering for allowing us to use the High-Performance Computing (HPC) Laboratory during our work.

Data availability and access

The data that support the findings of this study are available from UWARL, Sri Sivasubramaniya Nadar College of Engineering, Chennai but restrictions apply to the availability of these data, and so they are not publicly available.

REFERENCES

Cao, J., Zhang, K., Luo, M., Yin, C. & Lai, X. (2016) Extreme learning machine and adaptive sparse representation for image classification. *Neural networks*. 81, 91-102. doi: 10.1016/j.neunet.2016.06.001.

Chen, Z., Zhu, Y., Zhao, C., Hu, G., Zeng, W., Wang, J. & Tang, M. (2021) Dpt: Deformable patch-based transformer for visual recognition. In *MM '21: Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event China, October 20 - 24, 2021*. Association for Computing Machinery, New York, NY, United States. pp. 2899-2907.

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., ... & Guo, B. (2022) Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA.* pp. 12124-12134.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020) An image is worth 16x16 words: Transformers for Image Recognition at Scale. To be published in *Computer Vision and Pattern Recognition*. [Preprint] https://doi.org/10.48550/arXiv.2010.11929.

Fu, Z. (2022) Vision Transformer: ViT and its Derivatives. To be published in *Computer Vision and Pattern Recognition*. [Preprint] https://doi.org/10.48550/arXiv.2205.11239.

Guo, M. H., Liu, Z. N., Mu, T. J. & Hu, S. M. (2022) Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45(5), 5436-5447.

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J. & Shi, H. (2021) Escaping the big data paradigm with compact transformers. To be published in *Computer Vision and Pattern Recognition*. [Preprint] https://doi.org/10.48550/arXiv.2104.05704.

Huang, Z., Li, J., Hua, Z. & Fan, L. (2022) Underwater Image Enhancement Via Adaptive Group Attention-Based Multiscale Cascade Transformer. *IEEE Transactions on Instrumentation and Measurement*. 71, 1-18. doi:10.1109/TIM.2022.3189630.

Jose, J. A. & Kumar, C. S. (2020) Genus and Species-Level Classification of Wrasse Fishes Using Multidomain Features and Extreme Learning Machine Classifier. *International Journal of Pattern Recognition and Artificial Intelligence*. 34(11), 2050028. doi:10.1142/S0218001420500287.

Lakshmi, M. D. & Murugan, S. S. (2021) Modified restoration technique for improved visual perception of shallow underwater imagery. *Current Science*. 121(1), 103-108. doi:10.18520/cs/v121/i1/103-108.

Lee, S. H., Lee, S. & Song, B. C. (2021) Vision transformer for small-size datasets. To be published in *Computer Vision and Pattern Recognition*. [Preprint] https://doi.org/10.48550/arXiv.2112.13492.

Li, L., Shi, G. & Jiang, T. (2023) Fish detection method based on improved YOLOv5. *Aquaculture International*. 31(2), 1-18. doi:10.1007/s10499-023-01095-7.

Li, Y. & Chen, R. (2021) UDA-Net: Densely attention network for underwater image enhancement. *IET Image Processing*. 15(3), 774-785. doi:10.1049/ipr2.12061.

Li, X., Li, F., Yu, J., & An, G. (2022) A high-precision underwater object detection based on joint self-supervised deblurring and improved spatial transformer network. To be published in *Computer Vision and Pattern Recognition* [Preprint] https://doi.org/10.48550/arXiv.2203.04822.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV*), *Montreal, QC, Canada, October 10-17, 2021*. IEEE. pp. 9992-10002. doi: 10.1109/ICCV48922.2021.00986.

Liu, X., Jia, Z., Hou, X., Fu, M., Ma, L. & Sun, Q. (2019) Real-time Marine Animal Images Classification by Embedded System Based on Mobilenet and Transfer Learning. In *OCEANS 2019 - Marseille, Marseille, France, June 17-20, 2019*. IEEE. pp. 1-5. doi: 10.1109/OCEANSE.2019.8867190.

Muthuraman, D. L., & Santhanam, S. M. (2022) Visibility improvement of underwater turbid image using hybrid restoration network with weighted filter. *Multidimensional Systems and Signal Processing*. 33(2), 1-26. doi:10.1007/s11045-021-00795-8.

Peng, L., Zhu, C. & Bian, L. (2023) U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*. 32, 3066-3079. doi:10.1109/TIP.2023.3276332.

Qu, P., Li, T., Li, G., Tian, Z., Xie, X., Zhao, W., ... & Zhang, W. (2022) MCCA-Net: Multi-color convolution and attention stacked network for Underwater image classification. *Cognitive Robotics*. 2, 211-221. doi:10.1016/j.cogr.2022.08.002.

Si, G., Xiao, Y., Wei, B., Bullock, L. B., Wang, Y. & Wang, X. (2023) Token-Selective Vision Transformer for fine-grained image recognition of marine organisms. *Frontiers in Marine Science*. 10, 1-11. doi:10.3389/fmars.2023.1174347.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. & Beyer, L. (2021) How to train your ViT? Data, Augmentation, and Regularisation in Vision Transformers. To be published in *Computer Vision and Pattern Recognition*, *Transactions on Machine Learning Research* [Preprint] https://doi.org/10.48550/arXiv.2106.10270.

Touvron, H., Cord, M., El-Nouby, A., Bojanowski, P., Joulin, A., Synnaeve, G. & Jégou, H. (2021) Augmenting convolutional networks with attention-based aggregation. To be published in *Computer Vision and Pattern Recognition* [Preprint] https://doi.org/10.48550/arXiv.2112.13692.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017) Attention is All you Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Wang, G., Zhao, Y., Tang, C., Luo, C. & Zeng, W. (2022) When Shift Operation Meets Vision Transformer: An Extremely Simple Alternative to Attention Mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 36(2), 2423-2430. doi:10.1609/aaai.v36i2.20142.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L. & Zhang, L. (2021) CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada*. pp. 22-31.

Zendel, O. & Zinner, C. (2021) NAPHash: Efficient Image Hash to Reduce Dataset Redundancy. In *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), December 9-10, 2021*. IEEE. pp. 1-6. doi:10.1109/ICECET52533.2021.

Zeng, Y., Yang, X., Pan, L., Zhu, W., Wang, D., Zhao, Z., ... & Zhou, C. (2023) Fish school feeding behaviour quantification using acoustic signal and improved Swin Transformer. *Computers and Electronics in Agriculture*. 204, 107580. doi:10.1016/j.compag.2022.107580.

Zhuang, P., Wang, Y. & Qiao, Y. (2018) Wildfish: A large benchmark for fish recognition in the wild. In *MM '18:Proceedings of the 26th ACM International Conference on Multimedia*, *October 22 - 26, 2018, Seoul, Republic of Korea.* Association for Computing Machinery, New York, NY, United States. pp. 1301-1309.

Dhana Lakshmi MANIKANDAN received her B.E. degree in Computer Science and Engineering from Prathyusha Engineering College, Tiruvallur, India and her M.E. degree in Computer Science and Engineering from Sri Sivasubramaniya Nadar College of Engineering (SSNCE), Chennai, Tamil Nadu, India, in 2018. She earned her Ph.D. degree from Anna University for her research in Underwater Image Processing with Deep Learning. She achieved University rank in her post-graduation degree from Anna University, India. She had served as a Junior Research Fellow for the Department of Science and Technology (DST) funded project at SSNCE. Currently, She holds the position of Project Scientist II at the National Centre for Coastal Research (NCCR), Chennai. Her areas of interest include Image Processing and Deep Learning. Additionally, She has published her research works in International and National Journals and Conferences.

Sakthivel Murugan SANTHANAM obtained his B.E. degree from Madras University and M.Tech, degree from Pondicherry University. He received his Ph.D. degree from Anna University for his research work on Underwater Signal Processing. He currently serves as an Associate Professor in the Department of ECE, SSN since June 2001. His research area of interest is in underwater – acoustic communication, signal processing, acoustic wireless sensor networks, Green Energy Harvesting, and Deep Learning. He established an exclusive research lab for underwater, namely “Underwater Acoustic Research Lab” in 2014 in the department of ECE, SSN College of Engineering (https://sites.google.com/prod/view/uwarlssn). He is a Life Member of the Indian Society for technical education, Ocean Society of India, and the Acoustic Society of America.