Attention-Based Capsule Network with Vision Transformer for Underwater Hyperspectral Image Classification

Thiyagarajan B, Thenmozhi M
Department of Computer Science and Engineering, Puducherry Technological University, Puducherry, India

***Abstract*—The underwater investigations and research remain challenging due to various underwater distortion factors, scattering, and low wavelength absorption. Hyperspectral imaging helps in obtaining detailed information on each underwater object using the spectral reflectance, using 100 to 300 bands. The Hyperspectral imaging in underwater applications contains the 3D hyperspectral cube, which needs a high level of processing that results in high-accuracy classification. Hence, this study proposes the framework of hybrid deep learning techniques that work on segmentation, feature extraction, and classification processes. The Channel Attention Module (CAM) based U-Net architecture is used for the segmentation process to obtain the spectral spatial characteristics based on the Region of Interest (ROI). The CapsNet Feature extraction helps in obtaining the features of various bands, which helps in the classification of object class-wise using the pose-based relationships. The Vision Transformer (ViT) based classification that depends on the capsule vector token, carries out the multi-class classification by obtaining the global attention among the feature vectors and relationship-based long-range ROI feature data. In this way, the proposed model attains 95.3% accuracy using the maximum IoU of 0.88 and 95.2% of the segmentation process, which helps in achieving 0.99AUC for 8 substrate classes of the underwater HSI dataset.**

***Keywords*—Underwater imaging; hyperspectral; U-Net; Vision Transformer; CapsNet; classification; segmentation**

I. Introduction

Marine exploration, which is 70% part of the Earth's surface, remains slow and error-prone using the data collected from sources like AUVs, underwater sensors and cameras, ROVs, etc., which are used in domains like marine ecology, environmental monitoring, seabed mapping, aquaculture inspection, etc. The visibility in the underwater images are degarded due to the factors [1] like turbidity, low light, scattering, and colour distortion, which makes the visual inspection and interpretation a challenging task where the conventional systems completely rely on the RGB and multispectral sensors [2] that contain less spectral information. The distinguishing process of various marine materials [3] becomes a challenging task where the involvement of the advanced digital imaging techniques [4][5] helps to perceive the structures that have different spectral signatures too [6][7].

Hyperspectral imaging helps in the material identification process using the different spectral bands that capture the information using continuous reflectance, which helps in obtaining the discriminative features of spectral and spatial that help the digital image classification techniques [9]. The reviews provide an overview of the role of hyperspectral imaging in detecting coral species, geological substrates, and the chemical composition of seafloor ecosystems, among other applications, using different spectral bands that enhance both visual inspection and digital imaging techniques [10]. In the RGB imaging techniques, most of the marine materials are inspected as similar structures where the differentiation is effective using the hyperspectral signatures that classify the underwater objects and vegetation types [10].

Underwater, the absorption of red and near infrared wavelengths is rapid, which compresses the effectiveness of its spectral range, leading to continuous reflectance and results in poor signal quality at deeper zones in the sea, and also distorted reflected profiles occur due to the suspended particles that produce the spectral signatures inconsistent. The authors in [2] and [6] enlisted the main factors like depth, salinity, turbidity, chlorophyll concentration, and particulate matter illuminations as the reasons for the spectral nonlinearity that leads to the shift of scenes across the domains. In particular, when the annotated underwater data are insufficient, the introduction of 100-300spectral bands increases the risk of overfitting and computation complexity of the system that ends in an imbalanced dataset. In order to deal with deep neural networks, the algorithm must train with large and diverse information to generalise, where the limited samples are labelled, which makes the data discrepancy in the digital image classification process [7].

The characteristics of the underwater structures are non-rigid shapes, varying textures, and occlusion, which are difficult to visually distinguish from the background substrates that complicate the spectral and spatial classification, considering the intra-class variations and inter-class similarities [4][5]. Convolutional Neural Networks (CNN) are a strong contestant encountered in underwater object recognition and tracking [3], coral segmentation [4], seaweed detection [5], etc., where the local spatial features are given more importance with no information on global attentions. On the other hand, machine learning techniques are majorly dependent on the handcrafted features of hyperspectral image characteristics, where the performance is less when handling larger hyperspectral cubes under high noise underwater conditions [10] [11].

The sonar and RGB-based images lack the fine-grained material discrimination information, which cannot replace the high spectral resolutions of the hyperspectral images [1]. In the

existing literature, the involvement of transformer and attention-based models is used for RGB and traditional HSI dataset that depends on the global relationships [3], where it requires large datasets and strong prior training and finds difficulty with specific spectral distortions like absorption-induced band attenuation [1]. In the underwater HSI, the capturing of images despite the movement of seagrass and light absorption by water that affects infrared bands are encountered by hyperspectral acquisition process where the time variation in the hyperspectral cube and the orientation of tripods are done for movement of grass and to combat the IR bands issues the arrays of LED are used [12] however the method stands as the preliminary insights.

Therefore, the underwater HSI classifications need to develop the spectral distortion aware models, which requires the spectral spatial attention mechanism with a high level of feature learning process. Hence, the proposed work uses the channel attention U-Net for segmentation, CapsNet for Feature learning and Vision Transformer for Classification of underwater HSI.

The rest of the study is organized as follows: Section II: A survey on the various segments of underwater HSI, CapsNet feature extraction, and various transformer models; Section III: A detailed workflow of the proposed framework pipeline process; Section IV: Various results and discussions; and Section V: Conclusion and future work.

II. Related Works

A. Survey on Segmentation of Underwater HSI

In [13], the binary segmentation process is carried out using the Keras-U-Net architecture, where the cross-entropy and Jaccard distance are used to address the class imbalance, along with boundary detection, which helps to identify the object by removing the background information. This model failed to differentiate various ecological elements in the marine environment. Channel attention-based U-Net architecture is used for the image enhancement process of underwater image processing that overcomes the degradation effects due to attenuation and scattering effects, which is trained using the LayerNorm and GeLU activation function, showing improved performance by means of Floating Point Operations Per Second (FLOPS) in [14].

The U-Net architecture is designed in light weight manner [15] with no compensation on accuracy using the additional structures like parameter reduction using MRS (Multi-Residual Structure), accuracy improvement using the MSC (Multi-Scale Skip Connection) block, and weight changes using the CBAM (Convolutional Block Attention Mechanism) blocks that helps in fish segmentation process with good trade-off between accuracy and speed. In [15], there is a limitation to feature extraction layers that reduces the proposed lightweight U-Net architecture to work on fine-grained structures. The high-level feature information is fused, and global attentions are provided that help in the extraction of the target-specific segmentation process of aquatic data, which effectively comprises the boundary information, as described in [16], which forms the Semantic Segmentation Network, helping to improve the accuracy of boundary findings. The work in a smaller skewed distribution of classes shows less performance.

TABLE I. Survey on U-Net Segmentation Architectures

Ref.	Year	Model	Metrics
[13]	2025	Keras-U-Net-based binary segmentation using the NAUTEC—UWI Real Dataset	F-Score=94.59% IoU=89.75% Recall=93.38% Precision=95.83%
[14]	2025	Improved U-Net using Channel Attention, GeLU and LayerNorm and tested with the Underwater Imagenet Dataset	Flops reduction by half compared to CycleGAN, UGAN and UGAN-P UCIQE =0.008 (reduced) NIQE=0.019 (Improved)
[15]	2024	Lightweight U-Net using CBAM, MSC, and MRS	Reduction in Volume by 94.20% Reduction in parameters by 94.39% Reduction in floating-point by 51.2% mIOU=94.44% mPA=97.03% FPS=43.62
[16]	2024	Semantic Segmentation using Feature Fusion and Channel Attention Mechanism for Fish Species validated using SUIM dataset and Deep Fish Dataset	mIoU=95.05% mPA=80.37%
[17]	2024	U-Net Architecture using the Spectral global attention followed by Deformable Transformer for Remote Sensing Spectral Applications	Highes mIoU=57.22% Improvement in Accuracy=56.58%
[18]	2023	Attention U-Net for underwater image enhancement tested on UIEB and EUVP datasets	UICM=6.587 UISM=6.839

To segment the hyperspectral images in [17], the work introduces the global spectral attention to obtain the special characteristics of features in the spectrum and semantic information mining is done with the help of a Deformable Transformer that collectively forms the U-Net architecture, which shows a trade-off between high accuracy and parameter reduction, where the spectral redundancy is high due to the inter-band correlations. Image enhancement using a U-Net attention model is introduced in [18] for underwater image improvisations, where the retention of feature fusion and a weighted objective function used for multi-loss are used to show less loss performance that acts on the real-time world applications, but the work involves the downstream tasks (see Table I).

B. Survey on CapsNet-Based Feature Extraction

In the hyperspectral images, the inter-band selection leads to redundancy band selection that reduces the performance in the underwater HSI is addressed in [19] using the CapsNet architecture, where the relative importance of the spectral characteristics is provided using the deep feature extraction blocks consist of dual stream latent variable encoding, where the model suffers from band selection quality due to not mitigating noise. The model for handling the hyperspectral image requires

tackling the effects of non-linearity, which can be done using the Capsule networks that use the encoder and decoder blocks comprised of L2 normalisation, correlation matrix, distance measure and spatial attention mechanism that uses a small number of parameters results in high accuracy when classified but model performance is limited to high computational complexity in [20].

TABLE II. SURVEY ON CAPSNET ARCHITECTURES RELATED TO THE PROPOSED MODEL

Ref.	Year	Model	Metrics
[19]	2025	CapUBS using DLVE and DFEBs that were tested on the dataset from Indian Pines, InPool and Outpool	OA=98.7% AA=98.71% Kappa=98.04%
[20]	2024	SA-CapsNet for 3D spectral images	Accuracy=98.18% Parameters=369,609 Execution time=99.29s
[21]	2024	ShipGeoNET using Capsule networks, R-CNN and ViT Det	Geometric matching
[22]	2024	CapsNet Attention using Self-Attention mechanism for Hyperspectral images	OA=98.15% AA=98.73% Kappa=98.04%

Geometrical extraction methods highly help in the precise localisation, which is proposed in [21] for identifying and detecting the ships in marine environments, where the model excels in exact matching of the ship size onshore using the Radar images. The Vectors are used for the Attention mechanism in [22] for hyperspectral image classification using the CapsNet-based attention module that provides a high level of entity representation that satisfies most of the classification requirements, where the model fails to provide importance for the band weighting mechanism. The 3DCNN [39] models operate with the local receptive fields, this tends to learn the background-dominated convolutional patterns and lead to the overfitting issue of large labelled datasets (see Table II).

C. Survey on Transformer Models

The SpectralUformer model is proposed for UAV-based HSI technology that uses the hybrid attention block for the collection of spectral features and a hybrid cascaded upsampler for aggregation of features for the detection process, but the explanation on spectral normalisation is limited in [23]. To overcome the huge data requirement of the Vision Transformers, in [24], the researchers used an image block that utilises the potential relationships using the Amended Dual block attention mechanism on self-locally and external models, where the block tokenisation is performed to reduce the parameter requirement and provided a high accuracy range, but the work needs improvement in terms of image enhancement.

The hybrid deep learning model focused on the detection of fish species in the deeper sea and uses the ConvMixer and Swin Transformer for segmentation and classification, which helps in underwater exploration despite its bad lighting conditions [25], where the limitation is seen in the segmentation accuracy. The weak signal detection images of small objects in the marine environments is addressed by Hybrid attention transformer that utilises the global attention features and fine-grained features for object detection that reduces the detector performance with less parameter usage in [26] (see Table III).

TABLE III. SURVEY ON TRANSFORMER ARCHITECTURES RELATED TO THE PROPOSED MODEL

Ref.	Year	Model	Metrics
[23]	2024	SpectralUFormer using hybrid attention modules and an upsampler	Accuracy=93.15%
[24]	2024	ADANSE-based ViT on proprietary and standard datasets	Accuracy=90.9%
[25]	2024	ConvMixer and Swin Transformer for Fish classification	Accuracy=94.6%
[26]	2024	Lightweight Hybrid attention transformer for weak and small object detection	Parameter reduction to 28.5M Performance improvement by 6.3%.
[27]	2023	Lightweight Transformer network for Hyperspectral classification	OA=89.03% AA=93.52% Kappa=87.73%

The computational complexity of ViT increases with data requirement, which is handled in [27] using a lightweight architecture that concentrates on position-based and channel-based approaches for hyperspectral image classification process. The feature information is obtained using the convolution blocks and the self-attention mechanism for obtaining the spectral characteristics, where the system fails for complex scenes, as it requires the multi-feature extraction processes. HSI datasets are limited, and there is a need for the large-scale training data, lack of geometric encoding and noisy spectral tokens of the HSI transformer [38] makes it unsuitable for the classification for underwater objects like coral morphology.

From the survey, it is learnt that the existing segmentation model for the Underwater HIS processing lacks global reasoning. The commonly used CNN model does not provide long-range dependencies and the geometric coding using the small sample underwater conditions are present. This motivates the proposed work to concentrate on CAM-U-Net-based segmentation, CapsNet based feature extraction and classification using ViT hierarchical architecture.

III. PROPOSED FRAMEWORK

Based on the challenging factors of underwater exploration due to light absorption, turbidity, saline conditions, etc., compared to different imaging techniques involved in the capturing of information in underwater, Hyperspectral Imaging (HSI) helps a lot to obtain the material specific reflectance information using 100 to 400 spectral bands in the range of 400-900nm that obtains the spectral signatures of underlying objects. Therefore, the proposed model chose the HSI images for underwater image processing, where every information is captured using the specialised optical sensors, illumination systems, and calibration techniques in spite of the underwater distortions.

In the imaging principle, every pixel uses the wavelength to obtain the reflected intensity. Therefore, HSI contains the factors like (x,y, λ) parameters for processing which helps in the identification of the unique underwater materials. A wide aperture optics is accompanied by the optical sensor that obtains the spectral resolution of 1-5nm with a proper waterproof pressure housing, where the sensor captures the full spectral information that will be contained in a one line of pixels at a time in single slit capture. The factors like artificial illumination and illumination geometry optimisation are mandatory to overcome the absorption and scattering issues in underwater image capturing.

To overcome the spectral distortions, various techniques are handled in the calibration module, where the sensor noise is measured using the dark reference, reflectance normalisation using the white reference, alignment of wavelength channels using the spectral calibration and finally the refraction correct distortion using the geometrical calibration method using the formulation of reflectance conversion. Therefore, the final HSI cube information of the underwater imagery contains the number of scanned lines (H), pixels per line (B), and the number of spectral bands (W) that are acquired from one line of spatial pixels, where the full information per pixel is represented by R^(H×W×B). Fig. 1 is the overall proposed framework, where the functioning of each module is discussed below.

A. Underwater HSI Pre-Processing

To provide the spectrally accurate, consistent radiometrics and noise-free spatial data, the preprocessing steps are utilised to overcome the optical distortions mainly due to wavelength-dependent attenuation, absorption, scattering, etc., which helps the data acquired HSI cube become more stable for further segmentation, feature extraction and classification process.

It is well known that only blue-green bands can penetrate deeper, and quick attenuation of red bands requires radiometric and water column compensation due to the presence of spectral imbalance and suppressed reflectance in the HSI cubes data acquired [28]. In the radiometric correction, dark frame response is neglected from the raw sensor information of the HSI cube. The non-uniformity in the illumination is corrected by the normalisation process using the white reference spectrum and division instability. It is known that light attenuation increases with ingoing depths; hence, the reflectance value can be reduced by exponential multiplication of the absorption coefficient of seawater [8] and the estimated optical path length. To maintain the natural colour appearance of the grey world, correction can be applied where the assumption is made that the image average should be neutral grey and spectral band scaling is done using the attenuation rate that re-improved using the weighting mechanism for that particular band.

In the denoising process [29], it is important to preserve the curvature of the spectral signatures amidst the presence of salt and pepper noise, scattering noise and sensor noise, which is done using the 3D Gaussian filtering process that uses the 3D weighted 3D Gaussian kernels of size 3×3×5 that help in maintaining the spectral edges and material-specific peaks. To overcome the instability during the deep learning process due to the presence of illumination differences, scaling of wavelength responses, and convergence, the normalisation is done using the mean intensity of the band and the standard deviation of the wavelength.

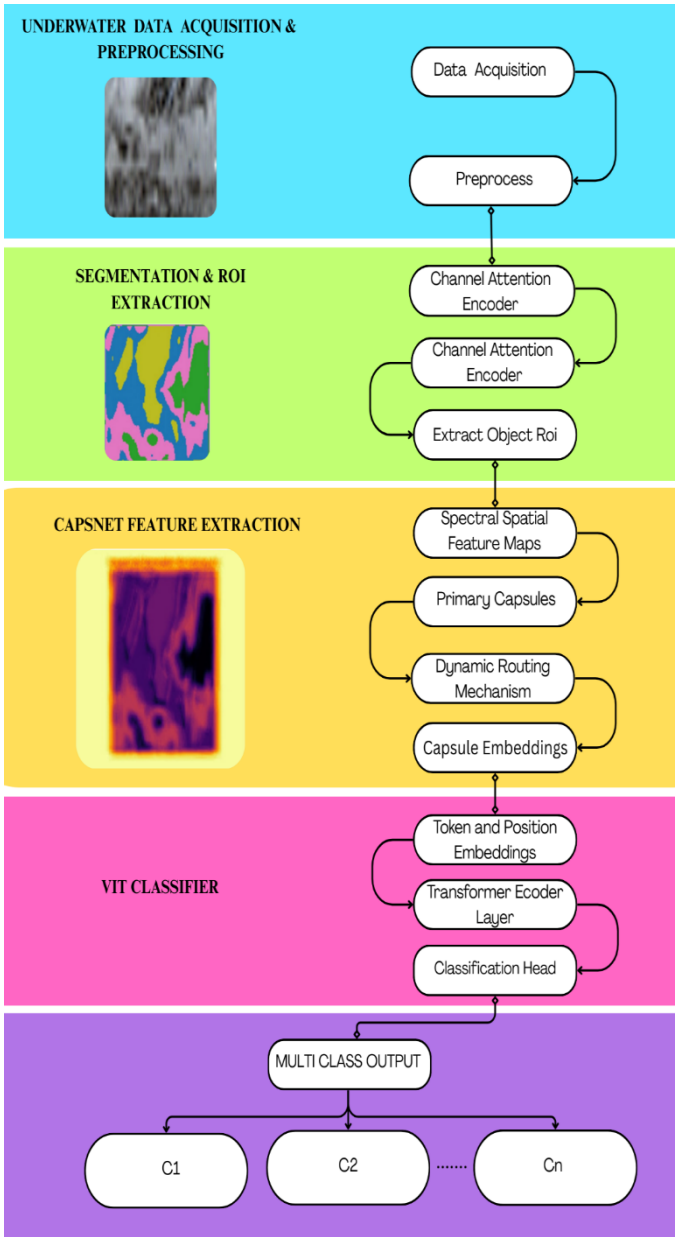

Fig. 1. Proposed framework for underwater HIS.

In contrast, the enhancement process, the band-wise contrast stretching is done using the minimum and maximum of hyperspectral data that contains a low dynamic range. In the next step, Channel Attention (CA) based U-net architecture is employed for processing, hence the sliding window extraction is done to obtain the patch generation based on the patch stride and the stride.

B. Underwater HSI CAM-Based U-Net Segmentation

Despite the preprocessing steps undergone for the HSI cube, the spectral imbalance, distortion, noise presence, etc., are still noticed in it; hence, the proposed framework employs the usage of U-Net architecture for the segmentation process that extracts the desired ROI from the spectral information despite the

background information [30]. It is noticed that the U-Net architecture is designed for RGB images; hence, it is difficult to segment the hundreds of channels present in the 3D HSI cubes. Hence, the proposed framework adopts the Channel attention module in Fig. 2 to improve the recalibration of featuring channels, spectral weight improvisation and spectral-spatial fusion. The proposed CAM-U-Net obtains the class adaptive spectral weights that help in the identification of the underwater scenes that exceed the conventional U-net architecture [31], where the suppression of irrelevant wavelengths is also suppressed during the process.

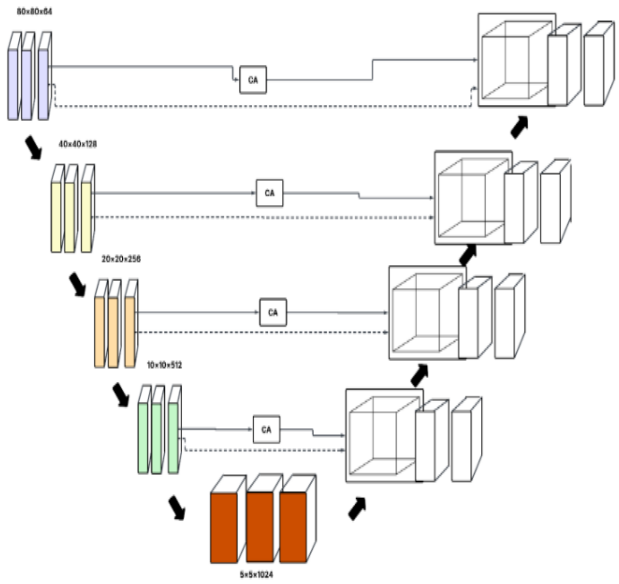

Fig. 2. Channel attention-based U-Net architecture for underwater HIS.

In the encoder of the CAM-U-Net architecture, the spectral spatial feature extraction [32] is done using the convolution blocks and the CAM block. Let us consider F_l as the feature map at the layer l, then the convolution operation using the ReLU activation is given in Eq. (1):

$$F_l = \sigma(W_l * F_{l-1} + b_l) \quad (1)$$

After the convolution is done, the CAM operation is carried out by providing the channel-wise importance weights by learning the feature maps from Eq. (1), where the spectral descriptor for the channel c is given in Eq. (2), which is represented by the Global Average pooling:

$$g_c = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} F_c(x,y) \quad (2)$$

Thus, in Eq. (2), g_c represents the particular channel global spectral activation; therefore, the final spectral representation is given by g [g1, g2……gc]. At the bottleneck layer, the 2-layer MLP is used to obtain the nonlinear channel relationships using the spectral activation and reduction layers ($W1. \in \mathbb{R}^{\frac{c}{r} \times c}$) by Eq. (3):

$$z = \delta(W_1 . g) \quad (3)$$

Using Eq. (3), the channel attention weights are obtained by the sigmoid function (σ) and the expansion layer ($W_{2.} \in \mathbb{R}^{c \times \frac{c}{r}}$), where r is the reduction ratio given in Eq. (4):

$$w = \sigma(W_2 z) \quad (4)$$

Therefore in each channel, the learned importance is obtained by multiplying it with the particular channel feature $F_{c,}$ which is represented in Eq. (5), where the attention weights lie in the range of [0,1].

$$F_c'(x,y) = w_c . F_c(x,y) \quad (5)$$

Thus, the refined feature F' contains the useful spectral bands, where the irrelevant wavelengths are suppressed. The maxpooling operations are done to enlarge the receptive fields of the refined features, which helps in obtaining the deeper features for ROI extraction. The upsampling of the layer l U_l using the transposed convolution operation that is fused with the refined features is given in Eq. (6):

$$S_l = Concat(U_l, F_l') \quad (6)$$

Then the channel attention module is applied to Eq. (7), which produces the fused representation S_l', where the convolution converted to decoder features is given in Eq. (7):

$$O(x,y) = W_{1\times1} * S_o'(x,y) + b \quad (7)$$

Then, the pixel-wise probabilities for Eq. (7), are given in Eq. (8) using the softmax operations:

$$P(x,y,k) = \frac{\exp(O_k(x,y))}{\sum_{j=1}^{k} \exp(O_j(x,y))} \quad (8)$$

Thus, using the final pixel-wise probabilities, the final segmentation map is obtained in Eq. (9):

$$\hat{M}(x,y) = \arg\max_k P(x,y,k) \quad (9)$$

Thus, the proposed framework of CAM-based U-net segmentation provides final segmentation details that contain the CAM amplified informative wavelengths, CAM suppressed noisy channels, edges are recovered using the decoder and skip connections, class-specific channels are obtained, and the U-net helps in obtaining the object size variability.

ROI Extraction: The segmented data need to match the input for the CapsNet feature extraction process; hence, the ROI patch generation is an important task that needs to be carried out to integrate the segmentation and feature extraction process in the proposed framework. Therefore, the bounding box based on the class of interest is obtained using Eq. (10):

$$ROI_{box} = H(x_1 : x_2, y_1 : y_2, :) \quad (10)$$

Then the above desired ROI extractions are resized for the Capsnet input that is given in Eq. (11):

$$X = Resize(ROI_{box}, r \times r \times C) \quad (11)$$

C. Underwater HSI CapsNet-Based Feature Extraction

The targeted underwater objects are obtained from the ROI extraction process using the CAM-based U-net Segmentation process that resolves the issues of mixed pixel contamination, boundary ambiguity, and spatial clutter in underwater scenes

based on the pixel-wise segmentation mask. It is important to extract the features [33] from the ROI that needs to contain high-level features that need to capture the irregular shapes, non-rigid structure, point variations, and depth-based spectral variations that are the characteristics of the underwater objects (see Fig. 3).

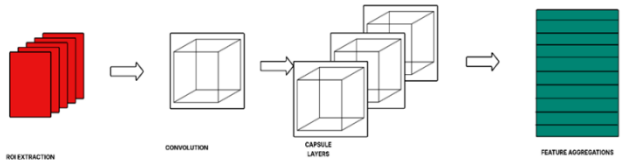

Fig. 3. ROI extraction for CapsNet for underwater HIS.

The learning of low-level features like texture, edges and local spectral gradients is obtained from two convolutional layers before the formation of capsules, which is given in Eq. (12) and Eq. (13):

$$F_1 = \sigma(W_1 * X + b_1) \tag{12}$$

$$F_2 = \sigma(W_2 * F_1 + b_2) \tag{13}$$

Thus, the F_2 contains the capsule formation that contains information about underwater object features. The capsule output [34] for each that is reshaped using the convolution feature maps is given in Eq. (14):

$$s_i = \sum_p W_{ip} . F_2(p) \tag{14}$$

The capsule vector is computed using the squash function that helps in providing the nonlinear vector length given in Eq. (15):

$$\hat{u}_{j|i} = w_{ij} u_i \tag{15}$$

The dynamic routing mechanisms help in mapping the parent node to the child node with the help of the coupling coefficients given in Eq. (16), where the nonlinear capsule outputs are produced in the vector form using the squash function that is represented in Eq. (17):

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \tag{16}$$

$$v_j = squash(s_j) \tag{17}$$

The transformational relationships help in capturing the coral reefs' curvature, texture, contour shapes, orientation variations based on the water movement, etc., which are represented in the Fcaps shown in Eq. (18), that is, actually the penultimate capsule output, where the softmax is not used, provides the deep features information.

$$F_{caps} = [v_1, v_2, \ldots . v_U] \tag{18}$$

D. Underwater HSI Vision Transformer-Based Classification

The HSI classification for underwater is done using the Vision Transformer that provides the long-range relationships to the spectral-spatial relationship, where the global attentions are provided to the capsule vectors obtained from Eq. (19):

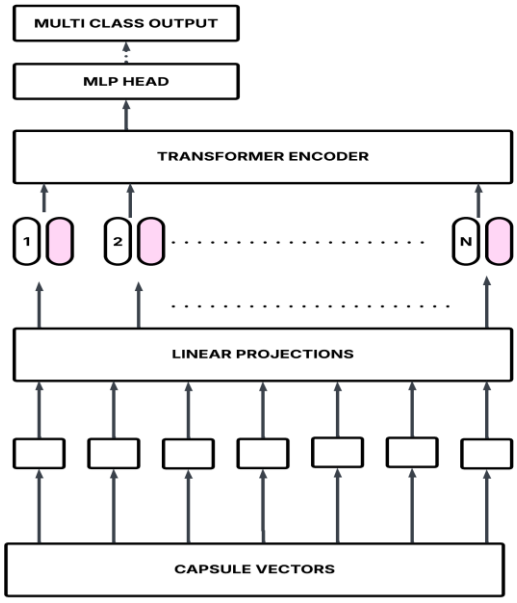

Fig. 4. ViT-based classification for underwater HIS.

$$u_i = squash(s_i) \tag{19}$$

The predictions are shared using the primary capsules for the higher level capsules using the formulation in Eq. (16).

The input of the ViT [35] requires a fixed level of dimension tokens; hence, the capsule vectors in straight act as the tokens, which is followed by the linear transformations [36] that are given in Eq. (20), which contains the learnable projection matrix Wp, and bias b.

$$L_i = W_p v_i + b_p \tag{20}$$

Now the tokenisation is given by T [L1, L2….LU]. Now the positional encoding is added to help the final classification process, which is done using the additions of the CLS token to the patches that are represented in Eq. (21):

$$Z_i^{(0)} = CLS + L_i + PE_i, \ \omega\eta\varepsilon\rho\varepsilon \ \iota = 1,2,,,,,,Y. \tag{21}$$

Now the self-attention is applied in the transformer block, where the neural network contains the query, key, and Value functions that extract the global and long-range dependency of the data. The value function extracts the actual object information present in ROI based features, the Query collects the information where the data contains the object-based information, and the Key collects the information that the query is trying to identify. The Query and Key parameters help to identify the classification pattern highlighted in the Value function using the attention formula given in Eq. (22):

$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{D}}\right) . V_i \tag{22}$$

The above attention values are obtained from the multi-heads of the transformer, to which the residual and layer normalisation are carried out. At the end using the CLS token that represents the entire patch is used for classification using the feed-forward neural network. Thus, the ViT helps in

handling the class ambiguity, irregular object shapes, which are common in underwater environments, and feature loss due to scattering.

IV. Results and Discussion

A. Implementation

The proposed framework is tested on the dataset Taxonomically annotated underwater hyperspectral and coral images of coral reef transects [37] that contains the hyperspectral images of underwater using the wavelength (400-750nm) using the 100-300spectral bands under the natural water illumination. The dataset contains the hyperspectral cubes and is annotated with detailed labelling of coral leaf substrates, corals, algae, sponges, sand, and rubble. Table IV provides the attributes of the datasets.

TABLE IV. HIS Dataset Attribute

Parameters	Value
No of Hyperspectral cubes	*120*
RGB images	*240*
Hyperspectral Cube Size	*512×512×(100-300) bands*
Spectral range	*400-750nm*
Mean annotated classes	*8*
Depths of transects	*5-30m*
Spatial resolution	*1-3mm/pixel*
ROI patches	*3730*

B. Data Preprocessing

The dataset contains the HSI cubes of size 512×512×no of bands, undergoes data preprocessing to reduce the noise and spectral distortion, which helps in the stable learning of the deep learning techniques involved in the proposed framework. At the initial stage, to reduce the spectral bias among the wavelengths, radiometric correction is done using the sensor-level dark current frames and white reference calibration for compensation purposes. Exponential absorption model is used for water column attenuation, and the normalisation is carried out using the z-sore among the cross-band illumination differences. The 3D Gaussian filters are used to suppress the shot noise and scattering distortions without disturbing the spectral gradients using the 3×3×5 spatial window filter size. The redundant information of spectral bands is reduced by the use of a 1×1 convolutional filter that compresses the channels to 16, which is followed by contrast stretching for improving the boundary visibility.

C. Segmentation

The encoder stage contains the augmented process of the U-net architecture and the CAM module using the four downsampling blocks, using 3×3 convolution per block, followed by the global average pooling and two MLPs, where the reweighting of the spectral channels is carried out using the CAM block. The connection between the encoder and decoder is maintained using the skip connections that help to maintain the spatial details, which are distorted due to the turbidity present in the underwater environment. The transpose of convolution upsampling takes place in the decoder, and the pixel-wise probabilities are obtained from the softmax operation using a 1×1 convolution operation. Now the mask of resolution 256×256 that is segmented for the classes of Pocillopora, Porites, Acropora, Macroalgae, sponge, and sand.

D. Feature Learning

The segmented masks are resized to a 64×64 bounding box that forms the ROI patches for the CapsNet feature extraction process. The two convolution layers provide the 9×9 kernel filters for extracting the low-level spectral and spatial features that are reshaped to 8-dimensional capsules for primary capsules that extract the high-level feature extraction that is concatenated into 10-digit capsule layers by a dynamic routing mechanism using the coupling coefficients. The features contain the capsule vector encoded geometry, orientation, and spectral reflectance patterns of similar coral taxa.

E. Classification

The capsule vectors act as the tokens that help in the tokenisation process, which are linearly projected for transformations followed by positional encoding. The 8 transformer layers with 6 Multi-Head Self-Attention and GeLU-activated feed-forward neural network are used for classification. The prediction of coral substrate classes is identified using the two-layer MLP using the CLS from the final layer.

This architecture combination ensures that the spatial relationships preserved by the capsules are effectively interpreted through the transformer's global context. By utilizing multi-head attention, the model selectively focuses on the intricate coral textures and geometric patterns that distinguish various substrate types. The feed-forward layers further refine these features, allowing for high-dimensional representations that remain resilient against underwater noise and lighting variations. As the data passes through the final transformer stage, the CLS token aggregates the most salient information required for a definitive categorization. Consequently, the two-layer MLP serves as a precise decision-making head, mapping these complex embeddings into distinct categorical probabilities with high accuracy. This integrated approach significantly enhances the model's ability to differentiate between the subtle morphological variations inherent in various coral species. The trainable parameter counts are estimated to be 15M, with the inference time of 30ms/ROI and Peak GPU memory of 3.5GB.

F. Implications

The visualization of Class Activation Maps (CAM) across the 10-band spectral data provides profound insight into the model's localized focus during the segmentation phase. By leveraging a U-Net architecture integrated with Channel Attention mechanisms, the network effectively filters out underwater backscatter while prioritizing the most discriminative morphological features of the coral substrates. Each feature map serves as a diagnostic window, revealing how the multi-band input—extending from visible light to specific spectral wavelengths—enables the system to bypass the inherent limitations of traditional RGB imagery. This high-dimensional feature extraction ensures that the boundaries of complex, overlapping coral colonies are delineated with high fidelity, significantly reducing false positives in heterogeneous seafloor

environments. Integrating 10-band spectral data with Channel Attention enables the U-Net to isolate critical morphological features while suppressing underwater noise, as demonstrated in the CAM visualizations (Fig. 4), for autonomous benthic mapping in complex seafloor environments.

Fig. 5 is the illustration of segmentation visualisation using the CAM-based U-net architecture for 10 bands, where the different band level shows different feature representation that helps in obtaining the informative bands. The higher bands 7 to 10 show good performance with more stable activations, which help in identifying the class-specific spectral characteristics using the channel attention recalibrations. The Gaussian-like activation regions are noted in the feature bands 3 to 6 that show more reliable reflectance structure for coral and benthic surfaces. In bands 1 and 2, the discriminative performance of the proposed framework has been obtained due to the strong attenuation for shorter wavelengths. Fig. 6 helps in obtaining the ROI extraction, where the mask was generated for band 3, which shows the high activation region, spatially coherent clusters, and segmentation contour overlay. Fig. 7 is the mean activation intensity of the segmentation process, which shows bands 7 to 9 have dominating spatial discriminations, which helps in obtaining the prioritised spectral channels. Table V presents the simulation parameters.

The bands 1, 3, and 10 have low values, indicating less involvement in the segmentation process. This selective prioritization not only streamlines the computational load during feature extraction but also enhances the signal-to-noise ratio by de-emphasizing the blue-dominated backscatter typically found in bands 1 and 3.

Consequently, the segmentation process becomes significantly more robust against the variable turbidity and lighting conditions common in deep-sea environments. Such channel-level optimization is pivotal for ensuring that the subsequent classification stages receive a refined, highly representative input, ultimately leading to superior identification of subtle coral substrate variations. This data-driven channel selection validates the use of multispectral imaging as a superior alternative to conventional three-band systems for complex biological mapping.

TABLE V. SIMULATION PARAMETERS

Parameters	Values
CAM-U-Net Architecture	
Input size	**256×256×16**
Encoder	4
Filter per level	64,128,256,512
Convolution	3×3, stride 1, padding 1
Activation	Relu
Pooling	Maxpooling 2×2
Channel attention reduction ratio	R=8
Decoder upsampling	Transposed convolution
ROI EXTRACTION	
Mask threshold	ARGMAX
Bounding box method	Contour based
Resize	Bilinear interpolation
Roi size	64×64×16
CAPSNET FEATURE EXTRACTOR	
Conv layer 1 and 2	256 filter, 9×9 kernel
Primary capsules	32 capsules × 8d vectors
Class capsules	10d ×16d vectors
Routing iterations	3
Capsule dim output	16d per class capsule
VIT CLASSIFIER	
Token dimension	**128**
No f token from Capsnet	32
Projection layer	16D→128D
Transformer layers	6
Attention heads	8
MLP DIM	256
CLS	Enabled
TRAINING SETUP	
Framework	PYTORCH
Optimiser	ADAM
Learning rate	1E-4
Batch size	16
Dropout	0.1
Hardware	NVIDIA RTX GPU
Regularization	1E-5

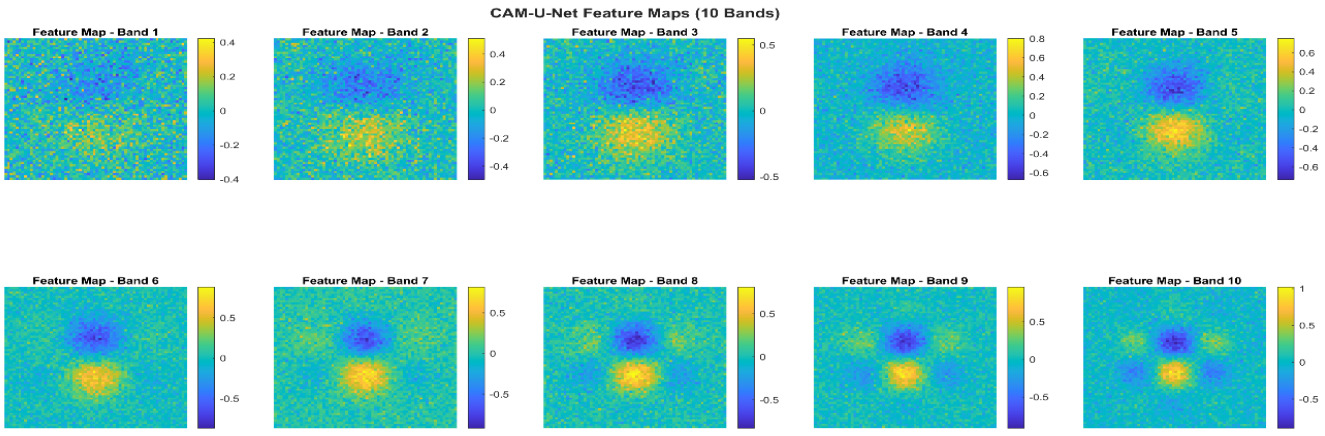

Fig. 5. CAM-based U-Net segmentation visualisation feature map for 10 bands.

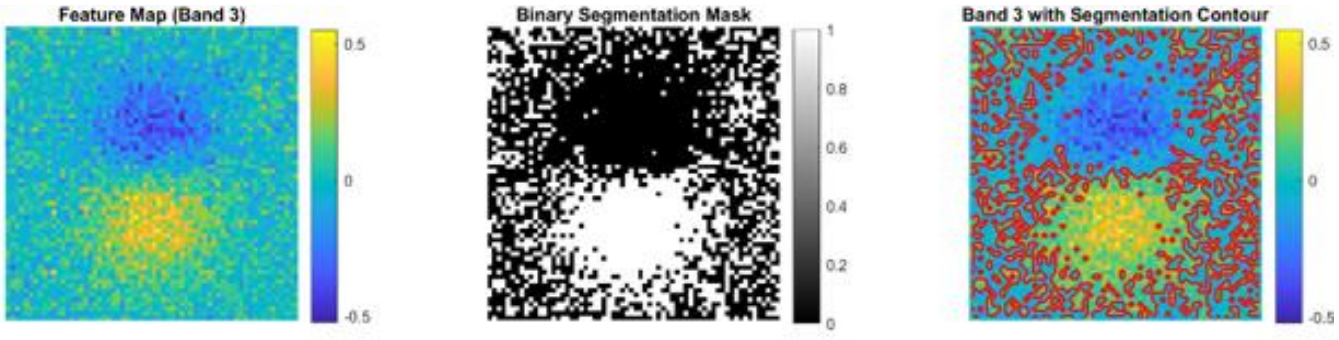

Fig. 6. CAM-based U-Net segmentation visualisation feature map and mask.

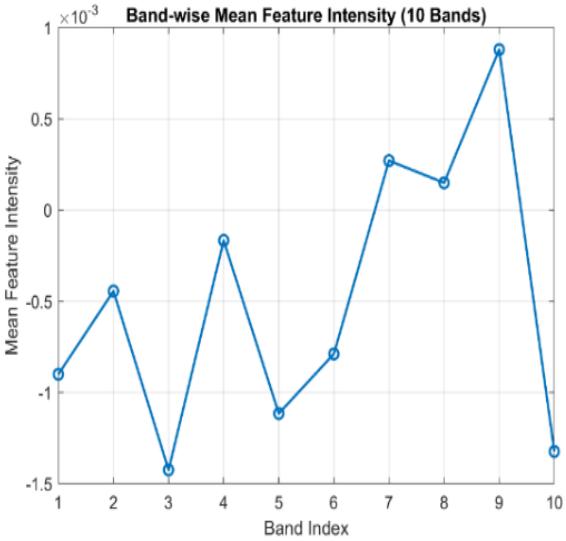

Fig. 7. Bandwise mean feature intensity for 10 bands using CAM-based U-Net segmentation visualisation.

Table VI is the CAM-based U-Net architecture for segmentation of substrate classes using the Intersection Over Union (IoU) to measure the overlapping, precision used for identifying the predicted positive pixels, recall helps to know the capturing most of the object, F scores helps in false negatives and positives and pixel segmentation helps in correctness that are manipulated using True positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

The highest IoU is seen in the sand that has different texture patterns and unique spectral reflectance, where the coral class Acropora preserves the good boundary conditions with the IoU of 0.82. In the overall estimation, the strong segmentation performance is noted by the IoU value between 0.72 and 0.88 among the substrate classes. It is important to obtain a false positive rate that is minimal, which is estimated using the precision calculation, showing a range from 0.80 to 0.93. The recall helps in obtaining the features of class regions, where 0.91 is at maximum and helps in calculating the trade-off F-score that shows consistent segmentation performance across different classes of spectral bands. Thus, the ROI patches have sharper boundaries, region consistency, and confident segmentation output, which is indicating highly discriminative and uniform characteristics.

TABLE VI. CAPSNET FEATURE EXTRACTION METRICS

Class	*Capsule Mean Length*	*Capsule Variance*	*Spread*
Acropora coral	0.862	0.041	0.203
Pocillopora coral	0.835	0.049	0.221
Porites coral	0.812	0.052	0.238
Soft coral	0.778	0.067	0.265
Macroalgae	0.902	0.036	0.184
Turf algae	0.791	0.059	0.249
Sponge	0.804	0.055	0.232
Sand	0.918	0.029	0.162
Acropora coral	0.862	0.041	0.203

Table VI is the CapsNet feature extraction performance on the ROI extracted data from the segmentation modules, where the factors like capsule mean length, capsule variance and spread help in the statistical analysis of capsule outputs. To know the presence of a feature, the capsule length can be used from the analysis of the sand, and the macroalgae show the highest identity of spectral-spatial features. To measure the stability of the capsule across various classes of ROI extraction, capsule variance is used, where the higher variance of soft coral has a higher 0.067 because of the greater intra-class variability based on the textures and spectral mixing. The capsule activation among the intra-class dispersion is noted using the spread function, where the soft coral shows a high of 0.265.

Table VI provides a comprehensive statistical analysis of the performance of CapsNet in feature extraction from region-of-interest (ROI) data derived from the segmentation modules. A comprehensive understanding of how well the capsule network reflects spectral–spatial features across several aquatic classes is provided by the analysis, which focuses on capsule mean length, capsule variance, and spread. By displaying the strength, stability, and dispersion of capsule activations, these measures provide deeper interpretation beyond classification accuracy.

Within each class, the feature existence confidence is directly shown by the capsule mean length. Stronger and more reliable activation of class-specific spectral-spatial characteristics is implied by higher values. Because of its

relatively homogeneous texture and low spectral mixing, sand has the highest capsule mean length of 0.918 in the data. In underwater photography, macroalgae have a high mean length of 0.902, which reflects their unique spectral response and unusual spatial patterns.

The statistical analysis in Table VI indicates that CapsNet proficiently captures substantial and interpretable spectral–spatial data, especially for classes exhibiting consistent visual traits. In order to improve global contextual knowledge, complementary mechanisms like attention and transformer-based modules are required, as evidenced by the observed variance and spread in complex biological classes. These results confirm that capsule-based feature modelling is a solid basis for classifying underwater images.

Alongside class-specific statistical results, the capsule-based study underscores CapsNet's capacity to maintain the spatial hierarchies intrinsic to underwater objects. Capsule representations capture both the existence and instantiation properties of features, in contrast to traditional CNN feature maps. This is especially useful in underwater environments where illumination, scale, and orientation vary greatly. Because of this feature, CapsNet can continue to provide discriminative representations even in the face of extreme optical aberrations. Compact capsule distributions with longer mean lengths indicate better separability in the latent space from the standpoint of categorization, which directly promotes trustworthy decision boundaries. The statistical compactness found in substrate-dominant groups like macroalgae and sand suggests that CapsNet can accurately and minimally confuse low-texture regions. On the other hand, the network's capacity to encode part-whole interactions, including branching, folding, or layered structures, is advantageous for biologically rich classes.

Furthermore, the segmentation-driven ROI extraction approach is indirectly validated by the capsule statistics. The segmented regions maintain important object-level information while decreasing irrelevant background noise, as demonstrated by stable capsule activations across several classes. The whole framework's resilience is strengthened by the synergy between segmentation and capsule-based feature extraction.

Global contextual awareness is required to further resolve ambiguities among visually overlapping classes, even while CapsNet excels at local structure preservation. Consequently, a solid empirical foundation for integrating attention mechanisms and transformer-based modelling in subsequent trials is provided by the capsule feature behaviour seen in Table VI.

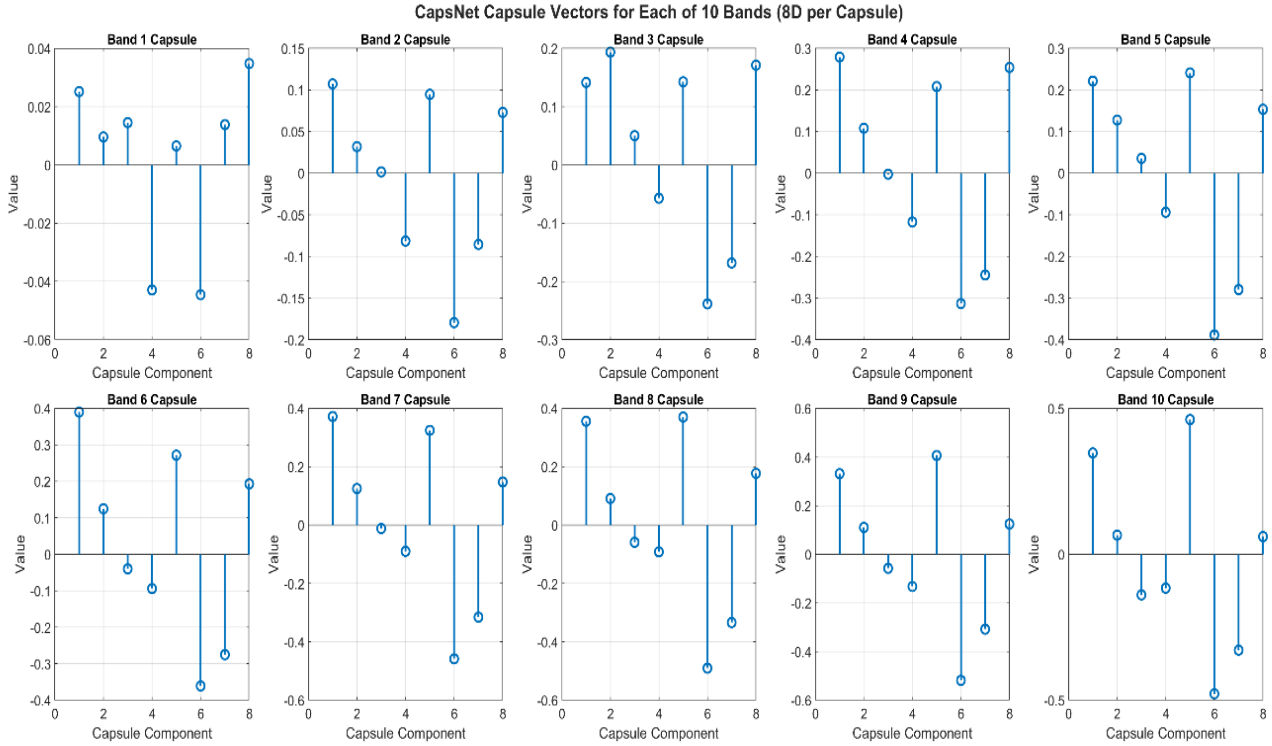

Fig. 8. Capsule vector for 10 bands based on the ROI extracted patches from the segmentation process.

Fig. 8 is the depiction of the CapsNet feature extraction projections using the ROI patches extracted from the segmentation process, which shows the effective multi-band feature abstraction with both positive and negative vectors indicating the spatial relationships based on the pose and magnitude-based features. The band 10 outputs are obtained from the 8-dimensional feature representation helps in exhibiting the spectral spatial characteristics among the bands. In the sensitive underwater imaging, the bands 3,6,7,8, and 10 help in the identification of corals, algae, sand, etc., classes present in the HSI cubes based on the higher vector components that capture the band-specific spatial spectral characteristics. Less dominance is noticed in the spectral bands of 4, 5 and 9 that contain the structured activations with informative spectral information. The shorter wavelengths showed low performance at bands 1 and 2, mainly due to the attenuation factor under the

sea, which showed weak capsule activations. These findings show that the capsule network creates discriminative pose-aware representations by efficiently utilizing informative spectral bands, which directly enhances class separability. In complex underwater sceneries, the predominance of mid-to-long wavelength bands allows for robust recording of object borders and texture variations. Stronger capsule activations in these bands are therefore associated with improved physiologically varied class recognition. The significance of choosing reliable spectral regions underwater is further highlighted by the diminished impact of shorter wavelengths. All things considered, the band-wise capsule behaviour shows that CapsNet is appropriate for spectral-spatial feature learning in difficult underwater hyperspectral situations.

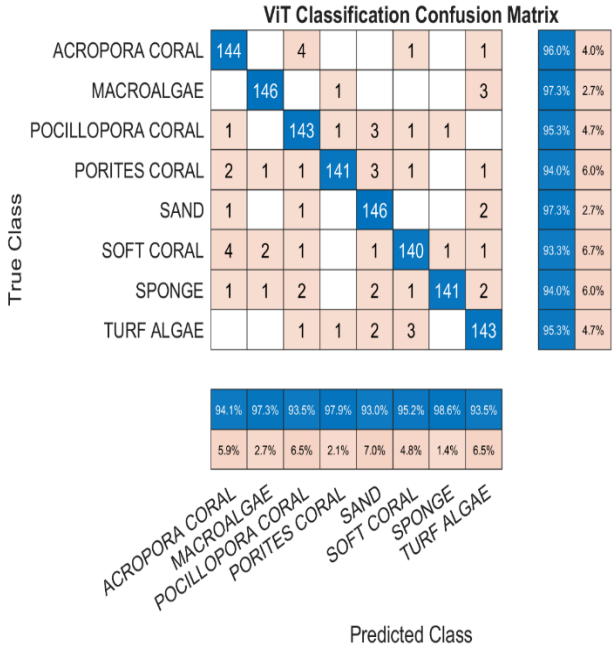

Fig. 9. Confusion matrix from ViT classification.

Fig. 9 is the confusion matrix obtained for the 8 substrate classes of the dataset of underwater HSI, which are preprocessed and segmented using CAM CAM-based U-net architecture and CapsNet-based feature extraction. The confusion matrix produces the cues for true class and predicted class for 8 substrates. The proposed framework highly classifies the sand class with 97% efficiency because of its fine-grained texture patterns that are effectively segmented and feature extracted.

The overall range is between 93 and 98% for all the classes of substrates in the underwater HSI datasets. This shows the preservation of spectral spatial characteristics by the proposed segmentation, features collection by CapsNet feature extraction, and classification, which obtains global attention based on range dependencies among the classes. The misclassification is noted in the coral taxonomy due to the similar structural patterns and spectral overlaps. Based on the confusion matrix, the proposed work plots the accuracy and error curves for 100 epochs, where the accuracy stood on average at 95% with an error percentage of 5%. The proposed method shows consistent learning activity till 60 epochs and saturates after that, as depicted in Fig. 10.

Fig. 11 is the depiction of AUC curves for 8 substrate classes that help to show the proposed framework's performance robustness for true negative and false positive rates. All the classes shown above 0.99 had the highest AUC values of macroalgae due to the distinctive spectral signatures for various hyperspectral bands. Table VII discusses the various ViT classification hybrid models, where the proposed combination framework shows higher performance.

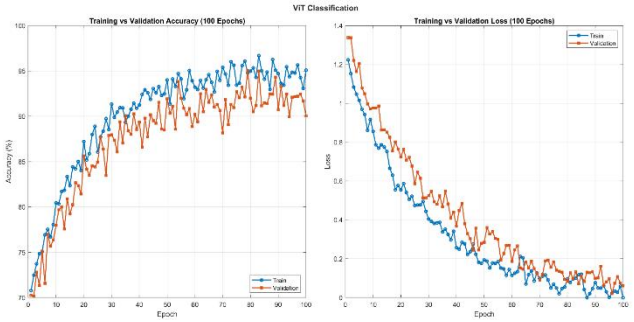

Fig. 10. ViT classification of underwater HSI based on CapsNet feature extraction and CAM-based U-Net segmentation.

The training curves' constant convergence behaviour demonstrates how stable the suggested learning framework is in challenging underwater environments. While the saturation after 60 epochs shows that the model achieves an ideal representation without overfitting, the progressive decrease in error rate during early epochs reveals efficient feature learning from segmented ROI patches. This characteristic is especially crucial for undersea hyperspectral data, as significant spectral redundancy and a lack of labelled samples can otherwise impair generalization.

The robustness of the suggested framework in differentiating between visually and spectrally similar categories is further demonstrated by the good AUC performance across all substrate classes. Reliable underwater scene interpretation depends on the model's capacity to maintain a favourable balance between true positive and false positive rates, which is confirmed by the nearly flawless AUC values. Macroalgae achieves the highest AUC due to its distinctive spectral reflectance and spatial continuity, enabling clearer decision boundaries within the feature space.

Common morphological traits and overlapping spectral responses are the main causes of misclassification within coral taxonomy, particularly between branching and large coral taxa. However, by maintaining spatial hierarchies while concurrently modelling long-range relationships, the network is able to reduce these ambiguities through the merging of capsule-based representations with transformer-driven global context. In difficult situations where local texture alone is insufficient, this dual representation capability enhances differentiation.

The channel attention mechanism suppresses redundant or noise-prone channels and highlights useful spectral regions to further improve feature selection. This selective amplification stabilizes categorization results across epochs and enhances learning efficiency. The Vision Transformer, CapsNet, and channel attention modules work together to acquire complementary information at various representation levels.

Overall, the performance trends show that the suggested hybrid design successfully strikes a compromise between global contextual understanding and local structural awareness. The framework's appropriateness for underwater hyperspectral image classification is validated by its good classification accuracy, robust convergence behaviour, and consistently high AUC values. These findings demonstrate that the suggested approach can function as a trustworthy basis for sophisticated mapping of marine habitats and automated substrate analysis.

Additionally, the integrated design shows tolerance to illumination variability and spectrum disturbance that are frequently found in underwater conditions. More accurate feature discrimination across diverse substrates is made possible by the combined learning of spectral context and spatial structure. For complicated benthic classes, where minute characteristics affect classification results, this is especially advantageous. Strong generalization ability across several ROI samples is shown by the observed performance consistency. As a result, the suggested architecture provides an efficient and well-rounded solution for underwater hyperspectral classification problems.

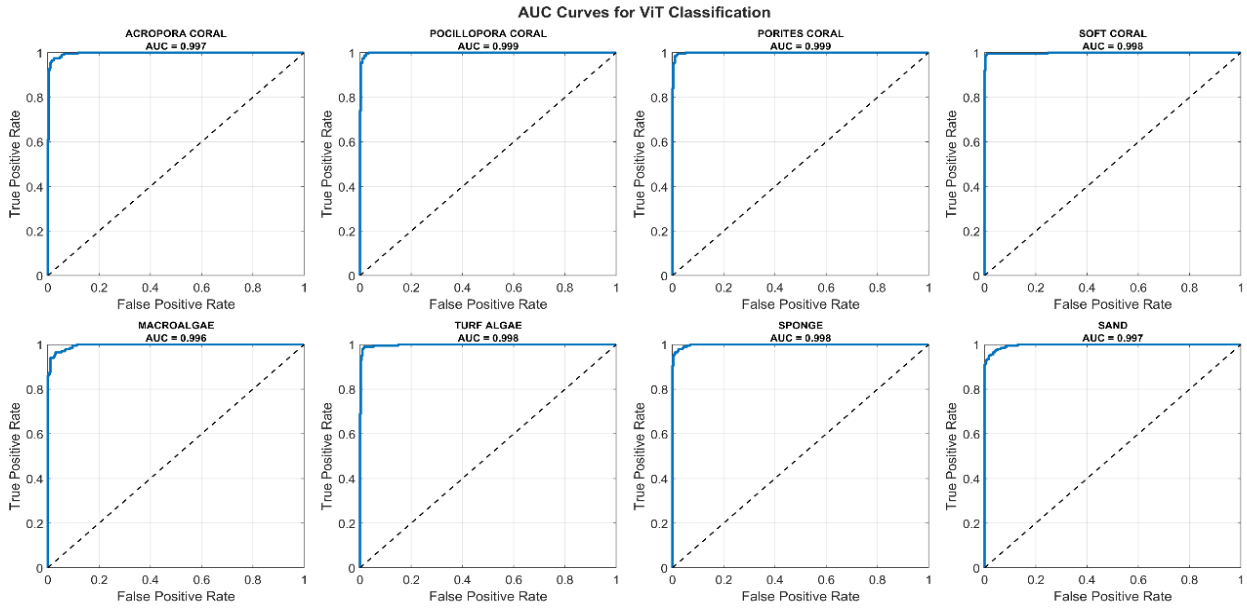

Fig. 11. AUC curves for 8 classes of underwater HSI using ViT classification.

TABLE VII. COMPARATIVE STUDY BASED ON THE PROPOSED MODELS

Ref	Year	Models	Accuracy
[23]	2024	SpectralUFormer using hybrid attention modules and an upsampler	93.15%
[24]	2024	ADANSE-based ViT on proprietary and standard datasets	90.9%
[25]	2024	ConvMixer and Swin Transformer for Fish classification	94.6%
[27]	2023	Lightweight Transformer network for Hyperspectral classification	89.03%
Proposed Model (CAM-based U-Net+Capsnet+ViT)			**95.3%**

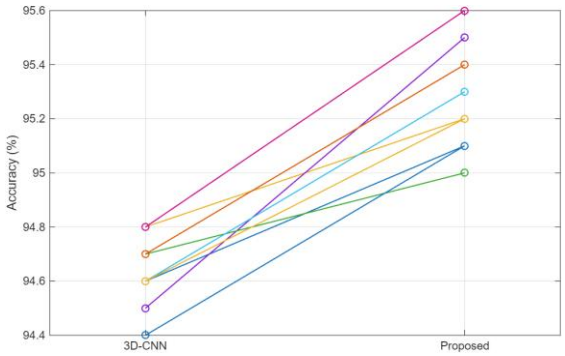

Fig. 12. Paired runs test for 3DCNN with the proposed model.

The statistical analysis is carried out using the paired runs with the 3D CNN [39] and the proposed framework against the measure of accuracy. It is seen that the baseline accuracy range is approximately 94.5% with the proposed accuracy range of 95.5% that shown the improvement per run to 0.4% to 0.9% (see Fig. 12).

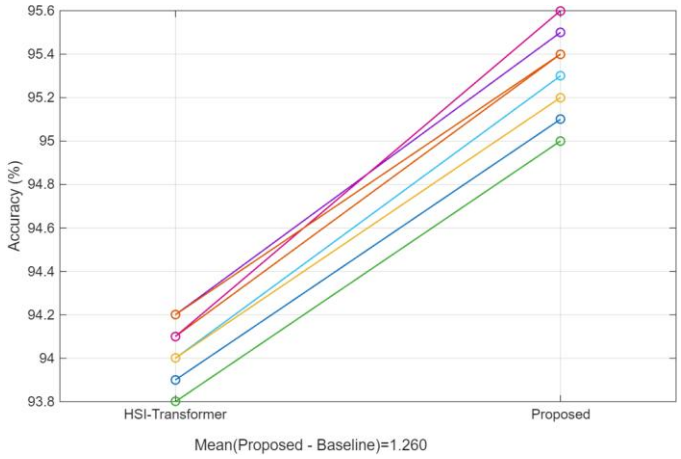

Fig. 13. Paired runs test for the HSI transformer with the proposed model.

Fig. 13 depicts the paired run test for the HIS transformer and the proposed framework. The proposed work shows stable gain using the channel attention-based segmentation, capsule features that provide pose-aware and the context based global transformer with a baseline accuracy range of 94.5% accuracy approximately and a mean improvement of +1.26%.

V. Conclusion

The advanced hybrid deep learning techniques that utilise the CAM-based U-et architecture for segmentation, CapsNet-based Feature Extraction, and ViT classification for Underwater Hyperspectral Image classification are framed and obtained the accuracy of 95.3% with AUC 0.99. The proposed framework obtains the region of interest from the hyperspectral images from 10 bands input by channel attention, which helps the U-net architecture to spatially and spectrally characterise multi-band hyperspectral images. The CapsNet features extract the pose and spatial relationships using the capsules and dynamic routing mechanism from the ROI extraction of segmented bands. The VitClassification uses the capsule vectors for identifying the global relationships using the query, key, and value in the transformer heads that provide range relationships among the features to identify the classes of substrates. The proposed framework achieved pixel accuracy of 95.2% with a maximum IoU of 0.88. In future works, as follows: a) The scalability of the proposed work applies to all hyperspectral image-based applications. b) Band-wise classification enhancement using the advanced generative models to improve the performance of the overall systems, and c) The proposed framework suffers from computational complexity; hence, the lightweight models need to be incorporated.

Conflict of Interest

The authors declare that there is no conflict of interest related to the research.

References

[1] Mittal, Sparsh, Srishti Srivastava, and J. Phani Jayanth. "A survey of deep learning techniques for underwater image classification." IEEE Transactions on Neural Networks and Learning Systems 34, no. 10 (2022): 6968-6982.

[2] Naveen, Palanichamy. "Advancements in underwater imaging through machine learning: Techniques, challenges, and applications." Multimedia Tools and Applications 84, no. 22 (2025): 24839-24858.

[3] Elmezain, Mahmoud, Lyes Saad Saoud, Atif Sultan, Mohamed Heshmat, Lakmal Seneviratne, and Irfan Hussain. "Advancing underwater vision: a survey of deep learning models for underwater object recognition and tracking." IEEE Access (2025).

[4] Li, Ming, Hanqi Zhang, Armin Gruen, and Deren Li. "A survey on underwater coral image segmentation based on deep learning." Geo-spatial Information Science 28, no. 2 (2025): 472-496.

[5] Prabhakar, G., S. Selvaperumal, and B. Baranitharan. "YOLOv5-based Enhanced Underwater Seaweed Detection Using Open-Source Datasets." Recent Patents on Engineering 19, no. 6 (2025): E300524230521.

[6] Aubard, Martin, Ana Madureira, Luís Teixeira, and José Pinto. "Sonar-Based Deep Learning in Underwater Robotics: Overview, Robustness, and Challenges." IEEE Journal of Oceanic Engineering (2025).

[7] Nawaz, Uzma, Mufti Anees-ur-Rahaman, and Zubair Saeed. "A Survey of Deep Learning Approaches for the Monitoring and Classification of Seagrass." Ocean Science Journal 60, no. 2 (2025): 19.

[8] Zhu, Rongxin, Lei Sheng, Kaitao Wu, Azzedine Boukerche, Libo Long, and Qiuling Yang. "Toward Efficient Underwater Visual Perception through Image Enhancement, Compression, and Understanding." ACM Computing Surveys (2025).

[9] Cheng, Ming-Fang, Arvind Mukundan, Riya Karmakar, Muhamed Adil Edavana Valappil, Jumana Jouhar, and Hsiang-Chen Wang. "Modern Trends and Recent Applications of Hyperspectral Imaging: A Review." Technologies 13, no. 5 (2025): 170.

[10] Wang, Zhixin, Peng Xu, Bohan Liu, Yankun Cao, Zhi Liu, and Zhaojun Liu. "Hyperspectral imaging for underwater object detection." Sensor Review 41, no. 2 (2021): 176-191.

[11] Montes-Herrera, Juan C., Emiliano Cimoli, Vonda Cummings, Nicole Hill, Arko Lucieer, and Vanessa Lucieer. "Underwater hyperspectral imaging (UHI): A review of systems and applications for proximal seafloor ecosystem studies." Remote sensing 13, no. 17 (2021): 3451.

[12] Longhi, Valeria, Nives Grasso, Paolo Felice Maschio, Fabio Menna, Arianna Pansini, Andrea Maria Lingua, Filiberto Chiabrando, Erica Nocerino, Giulia Ceccherelli, and Sabrina Rossi. "Underwater hyperspectral imagery for Posidonia oceanica mapping: challenges and preliminary results from the POSEIDON Project." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48 (2025): 177-184.

[13] George, Geomol, and Anusuya S. "Pretrained U-Net: in-depth analysis of binary image segmentation in underwater marine environment." Machine Learning for Computational Science and Engineering 1, no. 1 (2025): 4.

[14] Zhu, Sisi, Zaiming Geng, Yingjuan Xie, Zhuo Zhang, Hexiong Yan, Xuan Zhou, Hao Jin, and Xinnan Fan. "New Underwater Image Enhancement Algorithm Based on Improved U-Net." Water 17, no. 6 (2025): 808.

[15] Zhang, Zhenkai, Wanghua Li, and Boon-Chong Seet. "A lightweight underwater fish image semantic segmentation model based on U-Net." IET Image Processing 18, no. 12 (2024): 3143-3155.

[16] He, ZhiQian, LiJie Cao, JiaLu Luo, XiaoQing Xu, JiaYi Tang, JianHao Xu, GengYan Xu, and ZiWen Chen. "UISS-Net: Underwater Image Semantic Segmentation Network for improving boundary segmentation accuracy of underwater images." Aquaculture International 32, no. 5 (2024): 5625-5638.

[17] Zhang, Tianjian, Zhaohui Xue, and Hongjun Su. "Deformable transformer and spectral U-Net for large-scale hyperspectral image semantic segmentation." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17 (2024): 20227-20244.

[18] Tang, Pengfei, Liangliang Li, Yuan Xue, Ming Lv, Zhenhong Jia, and Hongbing Ma. "Real-world underwater image enhancement based on attention U-Net." Journal of Marine Science and Engineering 11, no. 3 (2023): 662.

[19] Li, Qi, Junwen Wang, Xingyuan Zu, Dan Chen, Ke Zhang, and Jinghua Li. "CapUBS: Capsule Network-Based Band Selection for Underwater Hyperspectral Imagery." IEEE Transactions on Geoscience and Remote Sensing (2025).

[20] Xiaoxia, Zhang, and Zhang Xia. "Attention-based deep convolutional capsule network for hyperspectral image classification." IEEE Access 12 (2024): 56815-56823.

[21] Yasir, Muhammad, Shanwei Liu, Xu Mingming, Jianhua Wan, Saied Pirasteh, and Kinh Bac Dang. "ShipGeoNet: SAR image-based geometric feature extraction of ships using convolutional neural networks." IEEE Transactions on Geoscience and Remote Sensing 62 (2024): 1-13.

[22] Wang, Nian, Aitao Yang, Zhigao Cui, Yao Ding, Yuanliang Xue, and Yanzhao Su. "Capsule attention network for hyperspectral image classification." Remote Sensing 16, no. 21 (2024): 4001.

[23] Yu, Zijian, Tingyu Xie, Qibing Zhu, Peiyu Dai, Xing Mao, Ni Ren, Xin Zhao, and Xinnian Guo. "Aquatic plants detection in crab ponds using UAV hyperspectral imagery combined with transformer-based semantic segmentation model." Computers and Electronics in Agriculture 227 (2024): 109656.

[24] Manikandan, Dhana Lakshmi, and Sakthivel Murugan Santhanam. "Underwater species classification using deep learning technique." Rev. Română De Informatică și Autom 34 (2024): 7-20.

[25] Pavithra, S. "An efficient approach to detect and segment underwater images using Swin Transformer." Results in Engineering 23 (2024): 102460.

[26] Chen, Gangqi, Zhaoyong Mao, Kai Wang, and Junge Shen. "HTDet: A hybrid transformer-based approach for underwater small object detection." Remote Sensing 15, no. 4 (2023): 1076.

[27] Zhang, Xuming, Yuanchao Su, Lianru Gao, Lorenzo Bruzzone, Xingfa Gu, and Qingjiu Tian. "A lightweight transformer network for hyperspectral image classification." IEEE Transactions on Geoscience and Remote Sensing 61 (2023): 1-17.

[28] Prasenan, Pooja, and C. D. Suriyakala. "A Study of Underwater Image Pre-processing and Techniques." In Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021, pp. 313-333. Singapore: Springer Singapore, 2022.

[29] Vijayalakshmi, M., and A. Sasithradevi. "A comprehensive review on deep learning architecture for pre-processing of underwater images." SN Computer Science 5, no. 5 (2024): 472.

[30] Huang, Guoheng, Junwen Zhu, Jiajian Li, Zhuowei Wang, Lianglun Cheng, Lizhi Liu, Haojiang Li, and Jian Zhou. "Channel-attention U-Net: Channel attention mechanism for semantic segmentation of oesophagus and esophageal cancer." IEEE Access 8 (2020): 122798-122810.

[31] Zhao, Peng, Jindi Zhang, Weijia Fang, and Shuiguang Deng. "SCAU-net: spatial-channel attention U-net for gland segmentation." Frontiers in Bioengineering and Biotechnology 8 (2020): 670.

[32] Guo, Changlu, Márton Szemenyei, Yangtao Hu, Wenle Wang, Wei Zhou, and Yugen Yi. "Channel attention residual u-net for retinal vessel segmentation." In ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1185-1189. IEEE, 2021.

[33] Jacob, I. Jeena. "Performance evaluation of caps-net based multitask learning architecture for text classification." Journal of Artificial Intelligence 2, no. 01 (2020): 1-10.

[34] SENGUL, Sumeyra Busra, and Ilker Ali OZKAN. "New Architecture in Deep Learning: Capsule Networks (CapsNet)." International Research in Engineering Sciences III (2022): 245-262.

[35] Han, Kai, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45, no. 1 (2022): 87-110.

[36] Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. "Transformers in vision: A survey." ACM Computing Surveys (CSUR) 54, no. 10s (2022): 1-41.

[37] Chennu, Arjun; Rashid, Ahmad Rafiuddin; den Haan, Joost; de Beer, Dirk (2020): Taxonomically annotated underwater hyperspectral and color images of coral reef transects from Curaçao [dataset]. PANGAEA, https://doi.org/10.1594/PANGAEA.911300

[38] Hong, Danfeng, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. "SpectralFormer: Rethinking hyperspectral image classification with transformers." IEEE Transactions on Geoscience and Remote Sensing 60 (2021): 1-15.

[39] Roy, S. K., G. Krishna, S. R. Dubey, and B. B. Chaudhuri. HybridSN: Exploring 3D–2D CNN Feature Hierarchy for Hyperspectral Image Classification. IEEE Geoscience and Remote Sensing Letters, 17(2), 277–281, 2020.