I.J. Image, Graphics and Signal Processing, **2026, 4, 65-81**
Published Online on August 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2026.04.04

HCTSpeckle and Lightweight YUV Transformer with ViT Encoder-Sandwich Decoder Network for Underwater Microplastic High Resolution Image Segmentation

Badugu Vimala Victoria*
School of Technology, Woxsen University, Hyderabad, India
Email: vimalavictoriagunupudi@gmail.com
ORCID iD: https://orcid.org/0009-0001-2916-2762
*Corresponding author

Kamil Reza Khondakar
School of Technology, Woxsen University, Hyderabad, India
Email: kamilreza@gmail.com
ORCID iD: https://orcid.org/0000-0001-5136-622X

Ravi Kumar Suggala
Department of Information Technology, Shri Vishnu Engineering College for Women, Bhimavaram, India
Email: ravi.suggala@gmail.com
ORCID iD: https://orcid.org/0000-0003-0962-5261

Received: February 5, 2026; Revised: March 11, 2026; Accepted: May 18, 2026; Published: August 8, 2026

Abstract: Microplastics are tiny particles made of plastic that are a significant source of pollution in the sea, and are dangerous to both the environment and human health. Nevertheless, the existing detection techniques are not robust and general in dynamic underwater settings with varying microplastic shapes and sizes. To overcome these issues, a high-resolution underwater microplastic segmentation framework was implemented that integrates HCTSpeckle and a Vision Transformer Encoder–Sandwich Decoder Network. Initially, underwater sensors record continuous visual images. The captured images are pre-processed with HCTSpeckle, a CNN-Transformer denoising network, for removing speckle noise, and it retains important structural information through the use of hybrid convolution-transformer blocks and double residual interactions. The denoised images were contrasted with the Lightweight YUV Transformer-based Network, in which multistage squeeze-and-excitation fusion is used to improve the visibility of object boundaries. The enhanced images are subjected to a hybrid Vision Transformer- Sandwich Decoder to produce precise underwater microplastic segmentation, where the ViT encoder captures high-level features that are globally correlated without distorting positioning information in space by patch embeddings and self-attention mechanisms. They are decoded using a Sandwich Decoder Network, which learns both local and global dependencies. Also, ranking and region-based pooling priorities fine edges and microplastic structures, whereas the pixel-wise segmentation head precisely categorizes the identified microplastics into fiber, film, pellet, and fragment types. The proposed approach attains the pixel accuracy of 0.97 and specificity of 0.98 with a Dice Coefficient of 0.934, which contains the effective segmentation result of underwater microplastic images. These findings ensure the framework efficiently identifies and categorizes various types of microplastics in diverse underwater sceneries.

Index Terms: Marine Pollution; Microplastic Segmentation; HCTSpeckle Denoising; Lightweight YUV Transformer-Based Network; Vision Transformer; Sandwich Decoder Network.

1. Introduction

Microplastics are tiny and synthetic-based solids or polymeric substances with dimensions between micrometres and a few millimetres in length, either as a result of breaking down larger pieces of plastic debris or made purposefully

during the production cycle. They are common in water bodies, entering oceans, rivers, and lakes via urban runoff, wastewater, and the decomposition of consumer plastics. Microplastics can be consumed by marine species because of their size and pose a bioaccumulation risk, causing an adverse impact on aquatic life and human health. Besides, microplastics may serve as vectors of toxic substances and microorganisms, which contributes to an even greater environmental threat [1]. The allocation, form and concentration of microplastics within different aquatic systems have to be tracked to comprehend their ecological effects and the possibility of mitigating these effects through appropriate strategies and sustainability measures.

Past research has discussed various automated methods to detect and segment microplastics. Two deep learning architectures, such as Mask R-CNN and U-Net, have been applied to identify and segment microplastic particles in microscopic and environmental images and have achieved high segmentation scores [2]. Traditional convolutional neural networks have been implemented to distinguish microplastics and complex backgrounds in urban water images, whereas object detection models such as improved Faster R-CNN have achieved high confidence in microplastic particle detection in seawater. In other studies, machine learning pipelines are used to combine object detection and segmentation to automatically quantify microplastic dimensions and shapes in digital images [3].

Although automated methods of microplastic analysis have improved, current techniques usually have limitations in dealing with changing underwater imaging parameters like light attenuation, scattering, and poor contrast. Most of the existing deep learning models have been trained on microscopic or near-surface images, and do not cope with the fine-scale noise structure and optical distortions of underwater images [4]. Moreover, traditional segmentation networks might fail to resolve the finer details of small and irregular shapes of microplastics at varying scales, resulting in lower segmentation accuracy. The absence of dedicated architectures that concurrently deal with the challenges of noise elimination, contrast enhancementin underwater conditions is a major drawback of credible microplastic segmentation [5]. To overcome these issues, an Underwater Microplastic High-Resolution Image Segmentation Framework Using Lightweight YUV Transformer with ViT Encoder-Sandwich Decoder was proposed.

The primary contributions of the proposed work include:

- Robust High-Resolution Underwater Microplastic Segmentation Using HCTSpeckle with Lightweight YUV Transformer and ViT-Based Decoder Network.
- HCTSpeckle is used as a denoising block to eliminate underwater disturbances without deteriorating the important features, which enhances the image quality in segmentation.
- Lightweight YUV Transformer is applied to enhance contrast of underwater images, and improve the visibility and other object boundaries in low light scenarios, resulting in more robust feature recognition.
- ViT Encoder–Sandwich Decoder is used to extract spatially and context-aware patch features and accurately segment small and irregular microplastics.
- To evaluate the effectiveness of the proposed model, the Pearson coefficient and pixel accuracy are evaluated.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of existing underwater microplastic image segmentation methods. Section 3 details the proposed model, while Section 4 presents the experimental results and performance evaluation. Finally, Section 5 concludes the study and suggests directions for future research.

2. Literature Review

The following section summarizes recent research that leverages machine learning and deep learning techniques for microplastic detection.

Galata et al. [6] introduce two open-source visual recognition models, such as YOLOv7 and Mask R-CNN, for efficient microfiber identification in environmental samples. Experimental findings indicated that YOLOv7 was more accurate with 71.4%, and Mask R-CNN, which had an accuracy of 49.9%.

Han et al. [7] introduced a deep CNN-based approach to the identification and categorization of microplastics in the ocean environment. The enhanced detection schemes that address the issue of low-quality underwater images, overlapping objects and occlusion using two improved versions of YOLOv3. The experimental analysis showed an average mean Average Precision (mAP) of 90%, which was a sign of solid detection performance.

He et al. [8] designed a microplastic detection network called Underwater Image Semantic Segmentation Network (UISS-Net). The semantic feature extraction with the help of auxiliary networks and channel attention mechanisms. Combined cross-entropy and Dice loss and multi-stage up-sampling were used to enhance boundary segmentation. The mIoU of UISS-Net was 95.05%.

Alsuwaylimi [9] developed YOLOv8-Segmentation models based on instance segmentation of underwater microplastics using the TrashCan dataset. YOLOv8n-Seg and YOLOv8s-Seg networks were tested in the problematic underwater conditions, with a precision of 75% and 70%, respectively, showing positive microplastic detection and segmentation.

Govindarajan et al. [10] created the Turbidity Enhanced Microplastic Tracker (TEMPT), an IoT-based microplastic detecting system. The system combined turbidity sensors and a microcontroller for a low-power and long-term monitoring system. The turidity-based algorithm (TETM-Water) allowed the successful performance of detection in noisy conditions. The system had a 91.47% accuracy rate and 5.40% error rate.

Pavithra presented a hybrid model combining the Swin Transformer and ConvMixer within a SwinUNet . This system works effectively even in murky water, uses few-shot learning to train with limited data, and achieves better accuracy. The model achieves the values of MIoU: 84.83.

Pushkala and Subbulakshmi [12] developed the marine litter classification using the panoptic segmentation combined with a Vision Transformer (ViT). Panoptic segmentation allows simultaneous identification of both debris and background by merging semantic and instance segmentation, while image enhancement and noise reduction improve input quality. The result shows the mAP: 0.89.

Table 1. Overview of the Literature Review.

Author	Method	Performances	Dataset	Drawback
Galata et al., [6]	YOLOv7 in rapid detection of microfibres	Accuracy: 49.9%.	MS COCO(Microsoft Common Objects in Context) dataset	Poor segmentation capability for fine microplastic features.
Han et al., [7]	Denoising Diffusion probabilistic model (DDPM) in (DiffWater)	SSIM:0.89	TEST-U90	It may not generalize
He et al., [8]	Underwater Image Semantic Segmentation Network (UISS-Net)	mIoU: 95.05%	SUIM	High computational cost
Alsuwaylimi, [9]	YOLOv8-Seg Instance Segmentation for Real-Time Submerged Debris Detection	Precision: 75%	TrashCan	Limited detection accuracy underwater condition such as low visibility and noise
Govindarajan et al., [10]	Turbidity Enhanced Microplastic Tracker (TEMPT), an IoT-based microplastic detecting system.	Accuracy: 91.47%	TEMPT	The model relies heavily on turbidity
Pavithra [11]	Swin Transformer and ConvMixer	MIoU: 84.83	SUIM	Limit performances when scaling to larger and more diverse datasets
Pushkala and Subbulakshmi [12]	Panoptic segmentation model	mAP: 0.89	AquaTrash dataset	Variation in underwater image quality can significantly affect performance.

2.1. Research Gap

However, despite the major improvements, the current literature has certain research gaps. The visual recognition system was an open-source architecture that was based on object detection models on underwater samples to improve pixel-level segmentation accuracy and fine boundary extraction [6]. The YOLOv3-based method was more effective in detection, but fails in scenarios of overlapping microplastics and fails to deal with fine-grained segmentation [7]. Even though UISS-Net had a high mIoU, complex auxiliary networks make it more computationally expensive and less scalable [8]. YOLOv8-Seg models exhibit moderate accuracy, especially in small and irregularly-shaped microplastic objects [9]. The IoT system relies on turbidity sensors, which do not use image-based spatial division of micro plastics [10]. To tackle these problems, we propose an Efficient Underwater Microplastic Segmentation framework using Lightweight YUV Transformers and a Hybrid ViT Encoder-Sandwich Decoder Network. This combination provides a distinct contribution by improving the boundary sensitivity, reducing computational complexity, and ensuring robust performance across diverse underwater conditions, thereby clearly differentiating it from existing transformer-based segmentation models.

3. Proposed Methodology

Microplastics are widespread in water, soil, air, and food, and are very dangerous to humans and animals. They are tiny and irregular and therefore hard to detect and quantify. Conventional AI applications tend to fail in diverse aquatic environments. Therefore, a deep learning model with a decoder network was implemented for robust underwater microplastic segmentation. Figure 1 illustrates the architecture for underwater microplastic segmentation.

Initially, the sensors are deployed underwater to obtain images beneath the surface of the sea. The sensor provides constant visible surveillance of the underwater conditions; it was specifically made to perform effectively in high-pressure and poor-lighting conditions. Such images underwater, which are frequently harmed by lower contrast, noise, and poor visibility, are initially forwarded to the image pre-processing methods. The input images are pre-processed by using the hybrid CNN-transformer network, namely HCTSpeckle, to eliminate the noise, where essential structural information is preserved while noisy speckles are efficiently suppressed by hybrid convolution transform blocks and double block residual interactions. Additionally, the noise-removed images were processed by the Lightweight YUV Transformer-based Network (LYT-NET) for effective contrast enhancement, where the Multi-stage Squeeze and Excite

Fusion Block emphasises the borders of objects and enhances visibility in challenging underwater environments through dynamically extracting essential characteristics at various scales. The enhanced images were processed by a ViT encoder and a Sandwich Decoder Network. The ViT encoder generated patch-wise feature maps with preserved spatial information using self-attention. The Sandwich Decoder combined contextual embedding, STPF, and TSSA to capture local and global features, while focused pooling improved boundary details and segmentation accuracy. Finally, the segmentation layer generates accurate area outputs that divide the identified microplastics into four different types: fibre, film, pellet, and fragment. This multiple process provides reliable segmentation of microplastic types in marine pollution monitoring.

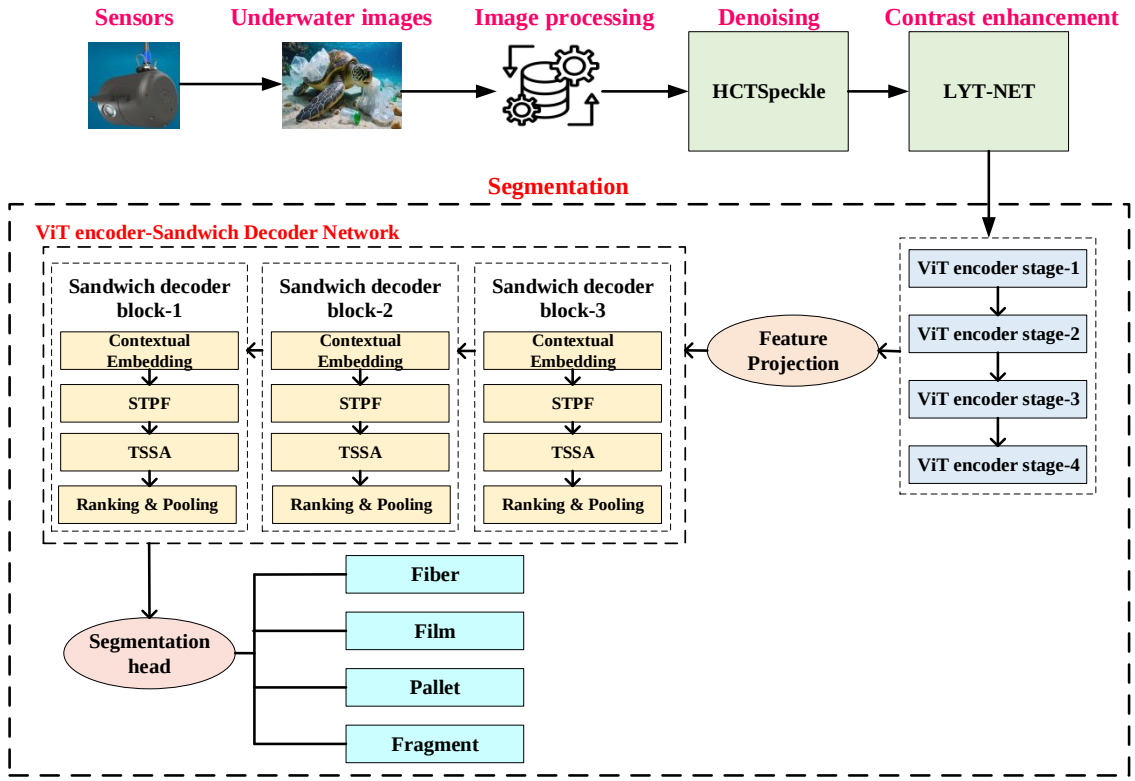

Fig. 1. Architecture for underwater microplastic segmentation.

3.1. Step 1: Preprocessing

Image preprocessing is an important process that enhances the quality of images by eliminating noise and enhancing visual information for image analysis. HCTSpeckle has been used in this work to reduce underwater noise without reducing vital structural details. The Lightweight YUV Transformer-based Network (LYT-NET) is then used to sharpen contrast and object edges in order to increase visibility in adverse underwater environments.

i) HCTSpeckle for denoising

HCTSpeckle is a CNN-Transformer-based speckle noise remover network that aims at eliminating the extreme noise introduced in underwater images. It is effective in speckle noise reduction and preservation of important structural and textual information needed to ensure accurate microplastic segmentation. Through a combination of hybrid convolution-transformer blocks and double residual interactions, the network is able to capture local spatial information as well as global contextual information, such that small microplastic edges are not lost [13]. Such a noise reduction procedure can greatly improve the image quality and give a solid input to further contrast enhancement and segmentation processes. The HCTSpeckle model utilizes L1 pixel loss as its primary loss function in eqn (1).

$$Loss = \frac{1}{N}\sum_{i=1}^{N} ||I_{ref(i)} - J_{out(i)||1} \tag{1}$$

where, J_{out} is the output denoised image, I_{ref} represents the reference noise-free image. N denotes the number of samples.

ii) Lightweight YUV Transformer-based Network for Contrast Enhancement

The Lightweight YUV Transformer-based Network (LYT-NET) is employed in this work to enhance the contrast of underwater microplastic images after denoising [14]. The network works on the YUV colour space in which the luminance (Y) and chrominance (U, V) are handled independently to effectively enhance brightness and colour balance. LYT-NET combines transformer blocks with a lightweight architecture, which allows capturing of multi-scale contextual information under a low computational complexity. Multi-Stage Squeeze-and-Excitation Fusion is an adaptive approach that highlights the object boundaries and removes background noise, enhancing the visibility of fine microplastic structures and later downstream segmentation.

3.2. Step 2: Segmentation

Segmentation refers to partitioning an image into meaningful regions by allocating a class label to each pixel, which allows the accurate localization of objects in a scene. A Vision Transformer (ViT) encoder with Sandwich Decoder Network is used in this work to perform segmentation. This method is able to capture global information and fine spatial information to determine accurate underwater microplastic segmentation.

i) ViT encoder-Sandwich Decoder Network

ViT Encoder and Sandwich Decoder Network is used to carry out precise segmentation of underwater microplastics. Vision Transformer (ViT) encoder splits the enhanced image in patches and feeds them through several layers of self-attention to learn rich representations globally and contextually while retaining positional information[15]. The features obtained are then fed to the Sandwich Decoder Network that combines contextual embedding, Space-Tight Pattern Fusion (STPF) and Two-Step Self-Attention (TSSA) to efficiently integrate local and global spatial data. Also, ranking improvement and area-of-interest pooling are used to highlight edges of fine objects and small microplastic features, which leads to accurate and reliable segmentation results [16]. The input image is divided into fixed-size patches and embedded as eqn (2)

$$E = \left[x_1 E_p; x_2 E_p; \ldots; x_n E_p\right] + E_{pos} \tag{2}$$

where, E_p is the patch embedding matrix, E_{pos} denotes positional embeddings, n shows the total number of patches, x_1, x_2 represents the input images, E_p is patch embedding. STPF integrates both local spatial features and global contextual features in order to improve the accuracy of segmentation. The fusion is performed as eqn (3)

$$F_{stpf} = \alpha F_{local} + (1 - \alpha) F_{global} \tag{3}$$

where, α controls the balance between spatial detail and global context, F_{stpf} denotes the fused feature map, F_{local} represents local spatial features, F_{global} is a global contextual feature. TSSA is an augmentation of feature representation that combines the spatial and channel-wise self-attention in a sequence to represent both structural and semantic dependencies. It is formulated in eqn (4)

$$F_{tssa} = SA_{spatial}\left(F_{stpf}\right) + SA_{channel}(F_{stpf}) \tag{4}$$

where, F_{tssa} denotes the output feature map, $SA_{spatial}$ refers to spatial self-attention, $SA_{channel}$ refers to channel self-attention, F_{stpf} is a fused features map. The final pixel-wise prediction is given by the following expression, which is shown in eqn (5)

$$S = Softmax\ (W * F_{out}) \tag{5}$$

Here, S represent the segmentation map, W is the segmentation head convolution, $Softmax$ represents the activation function, F_{out} is the final features map.

Figure 2 shows the proposed ViT encoder-Sandwich Decoder Network, where refined images are processed in a sequence of stages of the ViT encoder to retrieve hierarchical features. These features are aligned with feature projection and gradually optimized by Sandwich Decoder blocks, which incorporate contextual embedding, STPF, and TSSA to combine multi-scale features successfully. Lastly, a pixel-wise segmentation head is used to categorize microplastics into fiber, film, pellet, and fragment. The following equation shows the standard categorical cross-entropy loss in eqn (6).

$$L_{ce} = -\sum_{c=1}^{M} y_c \log(p_c) \tag{6}$$

Here, M is the number of microplastic classes, y_c denotes the ground-truth label of each pixel, p_c is the predicted probability for the class c at that pixel.

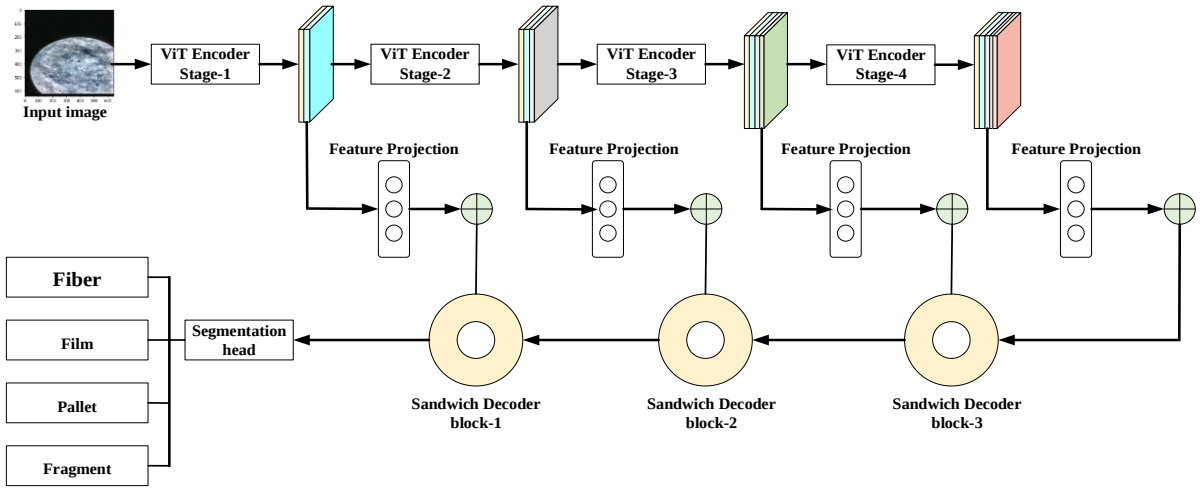

Fig. 2. Architecture for the ViT encoder-Sandwich Decoder Network.

Table 2. Key Layer –wise Architecture.

Layer Type	Output Shape	Parameters
Input Image	640 × 640 × 3	-
HCTSpeckle (Conv + Transformer)	640 × 640 × 64	Kernel 3×3
LYT-NET (Enhancement)	640 × 640 × 64	SE Ratio: 16
Patch Embedding (ViT)	1600 × 768	Patch: 16×16
ViT Encoder (×12)	1600 × 768	Heads: 8
Feature Reshape	40 × 40 × 768	-
STPF Module	80 × 80 × 384	α = 0.5
TSSA Module	80 × 80 × 384	Spatial + Channel Attention
Upsampling	640 × 640 × 64	Scale: ×2
Segmentation Head (Conv)	640 × 640 × 4	Kernel: 1×1
Output (Softmax)	640 × 640 × 4	Classes: 4

Table 2 presents the layer description of the model. It includes the input image, HCTSpeckle (Conv + Transformer), LYT-NET (Enhancement), Patch Embedding (ViT), ViT Encoder (×12), Feature Reshape, STPF Module, TSSA Module, Upsampling, Segmentation Head (Conv), and Output (Softmax).

4. Result and Discussion

The proposed model is developed for accurate microplastic segmentation underwater. The low-quality images obtained underwater by sensors are denoised by HCTSpeckle and enhanced with LYT-NET. The improved images are fed into a ViT encoder-Sandwich decoder network for segmenting the microplastic as fibre, film, pellet, and fragment. The proposed platform is developed using Python 3.8 and is powered by an Intel Core i5 CPU, Nvidia GeForce GTX 1650 GPU, and 16GB of RAM.

Table 3. Simulation Parameter for ViT-ESDN.

Parameter	Methods	Values
Input Image Size	ViT-ESDN	640 ×640
Batch Size		64
Number of Epochs		50
Optimizer		AdamW
Initial Learning Rate		1×10^{-3}
Final Learning Rate Factor		1×10^{-2}
Weight Decay		5×10^{-4}
Momentum		0.9
Early Stopping Patience		20 epochs
Mixed Precision Training		Enabled (AMP)
Pretrained Weights		Enabled
Number of Data Loader Workers		8
Batch size	Vision Transformer Encoder–Sandwich Decoder Network	16
Dropout rate		0.1
Optimizer		Adam
Activation		GELU
Patch size	Lightweight YUV Transformer-based Network	16 × 16
Optimizer		Adam

The simulation parameters for the proposed model are listed in Table 3. The parameters include Batch Size, Number of Epochs, Initial Learning Rate, and Momentum.

4.1. Dataset Description

The Microplastic_100 dataset includes annotated underwater images of micro plastics which are gathered to train segmentation and detection models. It consists of varied microplastic figures and backgrounds to demonstrate actual variability underwater. Here, it helps to train and assess the deep learning segmentation pipeline to classify fiber, film, pellet, and fragment microplastics correctly[17]. The images are manually annotated into four classes with consistency checks performed to minimize labelling errors. The dataset's variability was emphasised by including diverse microplastic shapes, sizes, and orientations, as well as challenging underwater conditions such as low lighting, noise, and complex backgrounds. Additionally, preprocessing steps such as resizing, normalization and augmentation were explicitly described to improve the model.

Table 4. Sample Distribution for Model Training and Testing.

Total samples	Training	Testing
2000	1800	200

Table 4 presents the distribution of samples used for training and testing the model. The dataset was augmented 5 times, with a total of 400 samples, and attained 2000 samples. Of the 2000 samples (80%), 1800 were allocated for training, and 200 (20%) were allocated for testing the model.

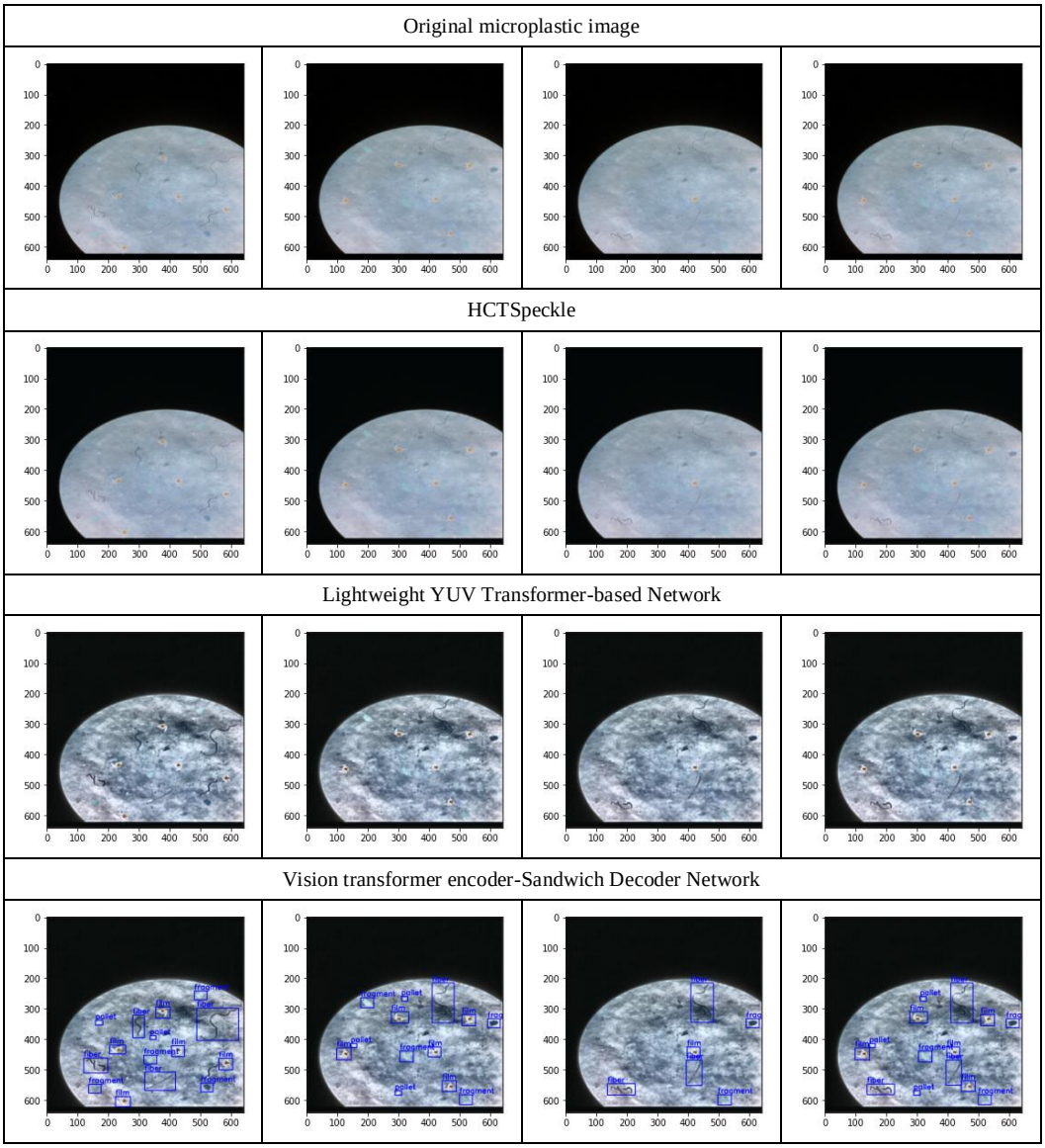

Fig. 3. Sample images of processed microplastic.

Figure 3 depicts the stages of underwater processing of microplastic images. The original images in the first row are corrupted by noise and low contrast. The second row shows the denoised results using HCTSpeckle. The third row shows contrast-enhanced results of the Lightweight YUV Transformer-based Network. The last row illustrates accurate segmentation obtained through the ViT encoder-Sandwich Decoder Network.

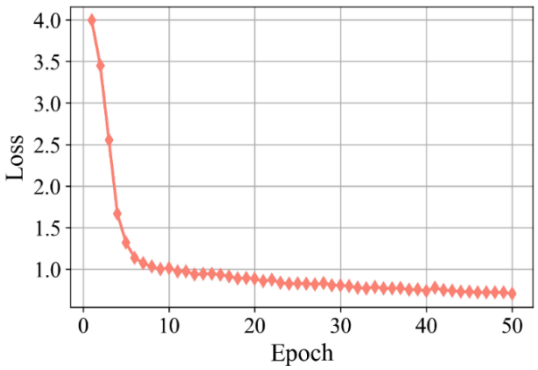

Fig. 4. Training Loss Curve by varying epochs.

The training loss curve in Figure 4 indicates that the loss sharply decreases in the early stages of training. It indicates that underwater microplastic features are learned successfully after pre-processing using HCTSpeckle and LYT-NET. Consistent refinement of the ViT encoder–Sandwich Decoder Network is shown by the gradual stabilization of the loss. This validates the stability of convergence and performance using the segmentation of the proposed model.

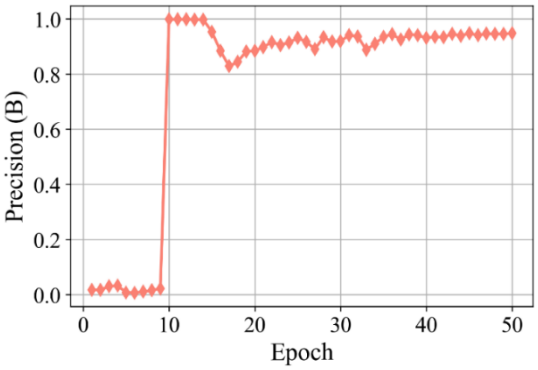

Fig. 5. Training Precision Curve through epochs.

Figure 5 shows the training precision curve. The quick development in the early epochs shows efficient learning of features by the improvement of input images. The high-precision level in the later epochs demonstrates the consistency of learning behaviour and accurate segmentation of the proposed model.

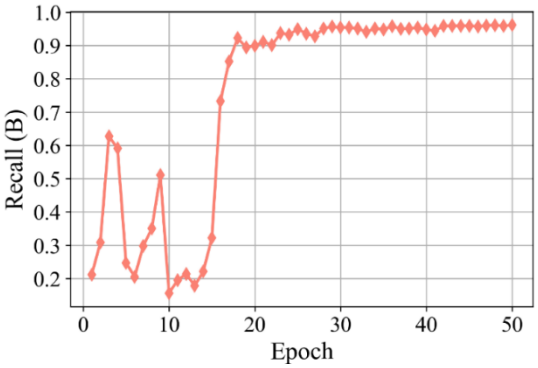

Fig. 6. Training Recall Curve for the proposed model.

Figure 6 indicates the recall performance of the proposed underwater microplastic segmentation model with different training epochs. The preliminary fluctuations are associated with the difficulty of the characteristics of underwater images, and then the quality of the features becomes significantly enhanced. The high recall that was

maintained in the later epochs shows that there was successful learning and precise identification of regions of microplastic.

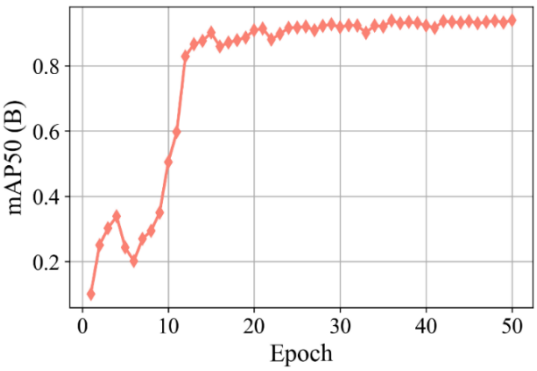

Fig. 7. Varying Epochs with mAP50 Curve.

Figure 7 shows the mean average precision of the proposed underwater microplastic segmentation model in terms of the training epochs. The high rate of escalation following the initial epochs indicates efficient feature optimization in training. The high and sustained mAP50 of later epochs shows proper localization and consistent segmentation of microplastic areas.

4.2. Comparison Analysis for Denoising

The proposed HCTSpeckle model is compared to Underwater Denoising Network (UDnNet) [18], DiffWater [19], and deep CNN-based Colour Balancing and Denoising Technique (CNN-CBDT) [20] to assess its denoising efficiency. Performance metrics for this evaluation encompass BER, MAE R^2. The graph below visualizes these elements.

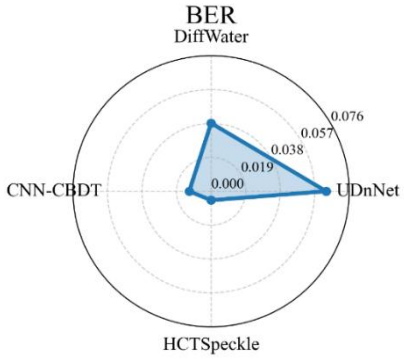

Fig. 8. BER Comparison of HCTSpeckle to other algorithms.

Figure 8 shows the BER performance of the proposed HCTSpeckle method in comparison with the current enhancement methods. The proposed HCTSpeckle has a BER of 0.005, and the improvement margin is between 0.007 and 0.063 compared to other techniques. This improvement shows that it has an effective noise suppression capability, providing higher-quality underwater images for accurate microplastic segmentation.

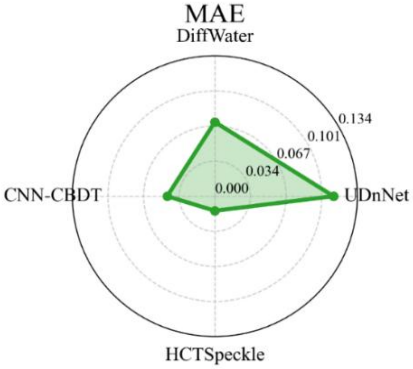

Fig. 9. Analysis of MAE of HCTSpeckle Compared to Existing Algorithms.

Figure 9 shows the performance of the proposed HCTSpeckle technique in terms of MAE as compared to the current enhancement methods. The proposed HCTSpeckle has a low MAE value of 0.014 with an improvement margin of between 0.031 and 0.096 compared to other methods. This shows that it has a higher capacity to minimize reconstruction errors and fine details of underwater microplastic images.

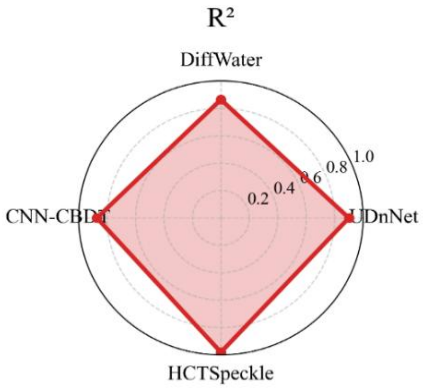

Fig. 10. Comparison of HCTSpeckle R^2 among existing Algorithms.

The performance of the proposed HCTSpeckleR^2 method in comparison to the existing enhancement techniques is depicted in Figure 10. The proposed HCTSpeckle shows a high value of R ²that is 0.981 as compared to other methods whose values fall between 0.864 and 0.900. This enhancement demonstrates a high correlation and better reconstruction fidelity realized by HCTSpeckle in the underwater image enhancement.

4.3. Comparison Analysis for Contrast Enhancement

The performance of the proposed LYT-NET model was compared to established algorithms, including Minimum color Loss and Local adaptive contrast Enhancement (MLLE) [21], Optimal Contrast and Attenuation Difference (OCAD) [22], and Weighted Wavelet visual Perception Fusion (WWPF) [23], with a graphical summary of the comparison metrics presented below.

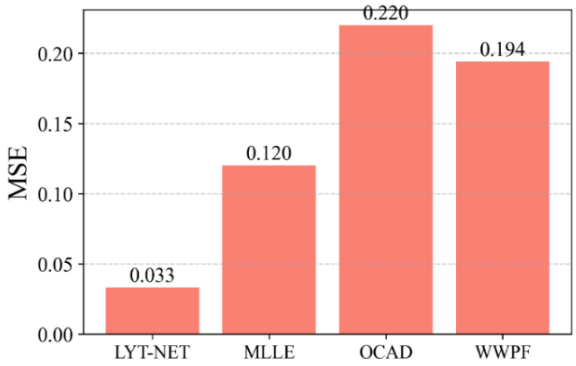

Fig. 11. Comparison of LYT-NET MSE among existing Algorithms.

Figure 11 shows the MSE performance of the proposed LYT-NET against the available enhancement techniques. The proposed LYT-NET has an MSE value of 0.033, which means that it has less reconstruction error when dealing with underwater images. This outcome corresponds to stable enhancement behavior and promotes a higher image quality towards further microplastic segmentation in underwater conditions.

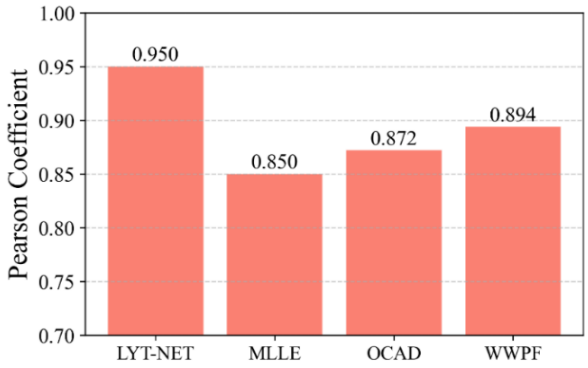

Fig. 12. Pearson Coefficient compared to the current algorithm and LYT-NET.

The Pearson Correlation Coefficient performance of the proposed LYT-NET compared to existing enhancement techniques is shown in Figure 12. The proposed LYT-NET gives a correlation of 0.950, which shows that there is a high linear consistency between the enhanced and reference images. Such an action assists in the credible structural conservation, which leads to the reliable and correct downstream microplastic segmentation.

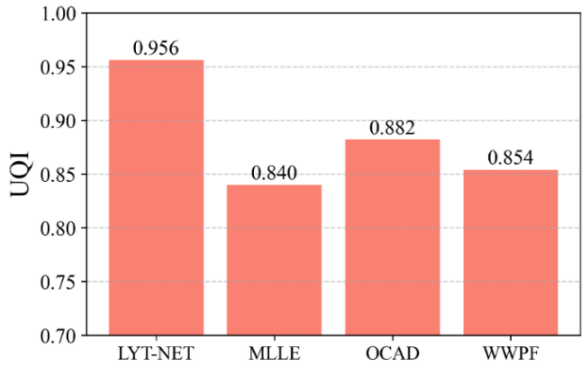

Fig. 13. UQI Comparison of LYT-NET against Established Algorithms.

Figure 13 shows the UQI performance of the proposed LYT-NET with the existing enhancement methods. The proposed LYT-NET method achieves a UQI of 0.956, which is an enhanced structural similarity and luminance consistency in improved underwater images. This preservation of quality facilitates better representation of features that help in proper and consistent microplastic segmentation in subsequent stages.

4.4. Comparison Analysis for Segmentation

The proposed ViT-ESDN model is compared to a number of current techniques in order to evaluate its effectiveness. It includes Mask-RCNN [6], Deep-CNN [7], YOLOv8 [9] and UISS-Net [8]. The evaluation metrics employed in this analysis encompass Pixel accuracy, Specificity, Pixel Error, USR, SSIM, Dice Coefficient, and Intersection over Union.

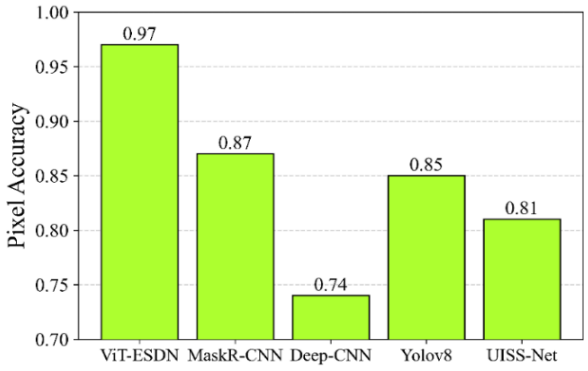

Fig. 14. Analysis of Pixel accuracy of ViT-ESDN Compared to Existing Algorithms.

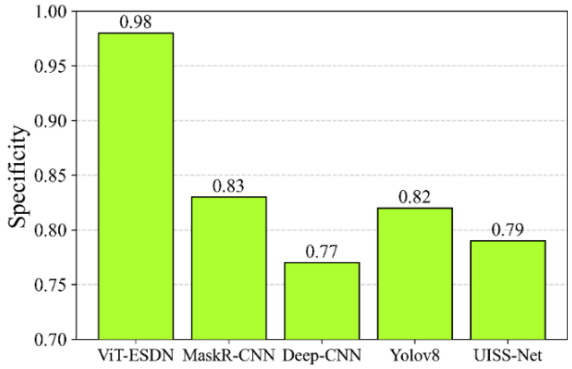

Fig. 15. Specificity compared to the current algorithm and ViT-ESDN.

The pixel accuracy and specificity performance of the proposed ViT-ESDN model compared to the current segmentation methods are shown in Figures 14 and 15. The proposed model attains a pixel accuracy of 0.97 and specificity of 0.98, indicating high confidence in pixel-level classification and effective background discrimination to identify underwater microplastic fragments.

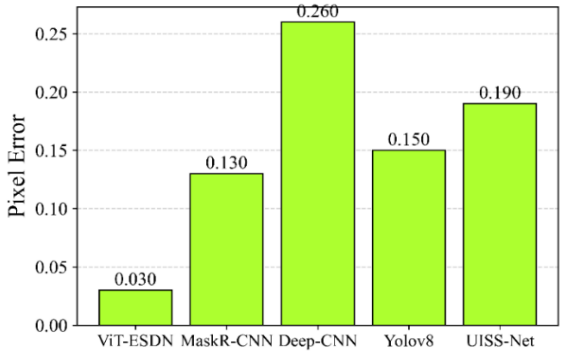

Fig. 16. Pixel Error comparison for ViT-ESDN with existing models.

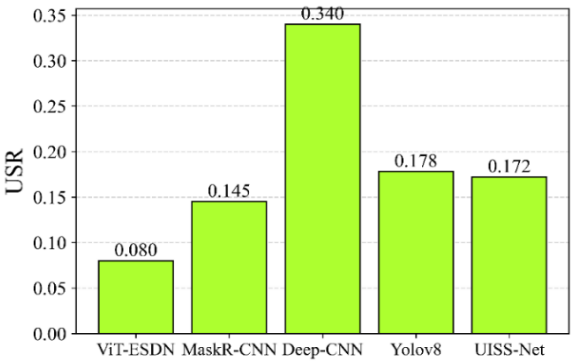

Fig. 17. ViT-ESDN comparison of USR with existing algorithms.

Figures 16 and 17 show the Pixel Error and the USR performance of the proposed ViT-ESDN segmentation model compared to the existing methods. It is observed that the proposed ViT-ESDN has a pixel error of 0.030 and a USR of 0.080. These findings suggest stable segmentation accuracy and low uncertainty in the underwater microplastic analysis.

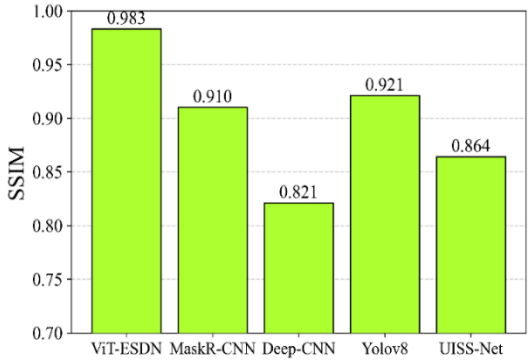

Fig. 18. SSIM compared to the current algorithm and ViT-ESDN.

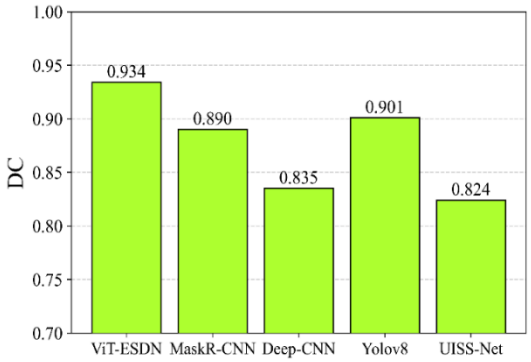

Fig. 19. DC compared to the current algorithm and ViT-ESDN.

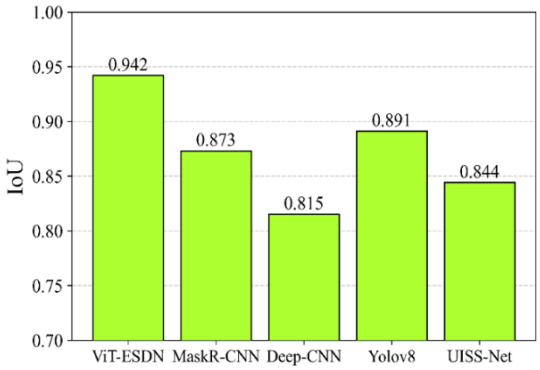

Fig. 20. ViT-ESDN comparison of IoU with existing algorithms.

Figures 18, 19 and 20 demonstrate the performance of the proposed ViT-ESDN segmentation model in terms of SSIM, Dice Coefficient, and IoU in comparison with the existing models. The proposed ViT-ESDN achieves SSIM of 0.983, Dice Coefficient of 0.934 and IoU of 0.942. These findings suggest that there is a structural similarity, overlap accuracy, and region agreement, which supports the reliability of microplastic segmentation in underwater images.

Table 5. Performance Comparison of Loss Functions for Underwater Microplastic Segmentation of the cross-entropy loss function with other existing loss functions.

Loss function	Pixel accuracy	SSIM
Cross entropy	0.97	0.983
MSE	0.882	0.895
Charbonnier	0.884	0.893

Table 5 provides the impact of different loss functions on the performance of segmentation. The pixel accuracy and SSIM of the cross-entropy loss were 0.97 and 0.983, respectively, which indicated the stable pixel-level learning. In comparison, MSE and Charbonnier losses were less precise and with similar performance, which shows that the cross-entropy is suitable for the segmentation task.

Table 6. Cross-validation performances of the proposed model.

No. of Fold	Pixel Accuracy	Specificity	Pixel Error
1	0.95	0.97	0.050
2	0.96	0.94	0.040
3	0.93	0.95	0.035
4	0.97	0.98	0.030
5	0.94	0.94	0.040

Table 6 illustrates the cross-validation results. Fold 4 achieved a pixel accuracy of 0.97 and a specificity of 0.98, with a low pixel error of 0.030, indicating reliable and consistent segmentation performance.

Table 7. Statistical test for the model.

ANOVA			
Source	SS	df	MS
Between Models	185.42	4	46.35
Within Models	102.75	20	5.13
Total	288.17	24	
Wilcoxon Signed-Rank Test			
Comparison	W-value	p-value	
ViT-ESDN vs Deep CNN	2	0.003	
ViT-ESDN vs UISS-Net	4	0.005	
ViT-ESDN vs Mask R-CNN	6	0.007	
ViT-ESDN vs YOLOv8-seg	8	0.009	

Table 7 presents the statistical validation of the model. The ANOVA result shows that the between-group variance (SS=185.42, MS=46.35) is much higher than the within-group variance(SS=102.75, MS=5.13), indicating significant differences among the models. Furthermore, the Wilcoxon test result reveals that all comparisons between the proposed ViT-ESDN and other models have p-values less than 0.01.

Table 8. Performance Comparison with Confidence Intervals.

Model	Pixel Accuracy	95% CI (±)	SSIM	95% CI (±)	IoU	95% CI (±)
Deep-CNN	0.74	0.0186	0.821	0.0155	0.815	0.0174
UISS-Net	0.81	0.0155	0.864	0.0124	0.844	0.0143
YOLOv8	0.85	0.0136	0.921	0.0111	0.891	0.0124
Mask R-CNN	0.87	0.0124	0.910	0.0105	0.873	0.0118
ViT-ESDN(proposed)	0.97	0.0093	0.983	0.0074	0.942	0.0087

Table 8 compares model performance using pixel accuracy, SSIM, and IoU, along with the 95% confidence intervals. The proposed method achieves pixel accuracy (0.97), SSIM (0.983), and IoU (0.942), outperforming Deep-CNN, UISS-Net, YOLOv8, and Mask R-CNN.

4.5. Discussion

The model shows performance degradation in challenging scenarios such as extremely low visibility, heavy speckle noise, complex backgrounds, and overlapping or very small microplastic particles, where boundary delineation and class discrimination become difficult. Generalization capability, data augmentation, and multi-scale feature learning improve robustness; performance may vary on unseen datasets with different environmental conditions or sensor characteristics. These enhancements provide a more balanced and critical evaluation of the proposed approach.

5. Conclusion

Microplastics are microscopic and intractable pollutants that are causing significant dangers to both marine ecosystems and human health because of their ubiquitous presence and their inability to decompose naturally. However, there is a lack of reliable underwater microplastic segmentation even with the growing research interest due to the typical issues of underwater images, including low contrast, speckle noise, colour distortion, low visibility, and small, irregularly-shaped particles. Current image-based approaches have significant disadvantages in maintaining fine structural information, dealing with noise and changes in illumination, and the accuracy of segmentation and the reliability of microplastic classification is limited. To address these issues, Advanced Underwater Microplastic Segmentation through ViT Encoder-Sandwich Decoder Architecture is proposed. First, underwater sensors were implemented to capture images continuously under the sea surface. Noise and contrast-degraded images were initially preprocessed using an HCTSpeckle hybrid CNN-Transformer network, which was effective in reducing speckle noise and maintaining the necessary structural and textural details with the help of hybrid convolution-transform blocks and double interaction residue. Then, the denoised images were refined with the Lightweight YUV Transformer-based Network, where the multistage squeeze-and-excitation fusion was employed to highlight the appearance of objects and

enhance the visualization by dynamically extracting salient features at various scales in the YUV color space. The improved images were further sent to the Vision Transformer encoder to produce rich feature maps while preserving spatial and positional details in patch embeddings and multi-head self-attention. These feature maps are then decoded through the Sandwich Decoder Network which combines contextual embedding, Space-Tight Pattern Fusion and Two-Step Self-Attention to effectively encode both local and global contextual dependencies. Besides, the ranking of improvement and region-centred pooling allowed us to focus precisely on fine edges and small microplastic structures that can greatly enhance the accuracy of segmentation. Lastly, the segmentation head generated correct pixel-wise results, classifying the detected microplastics into fibre, film, pellet, and fragment. The performance analysis, with a specificity of 0.98%, a Pixel Error of 0.030%, and a USR of 0.080%, indicates that the proposed technique segments microplastics more effectively than existing methods. However, the proposed approach may be less effective in highly turbid water, in the presence of severe occlusion or overlapping microplastic particles, and it requires high-resolution images with considerable computational capabilities. Future research could aim at increasing robustness in low-visibility conditions, classify other types of microplastics or pollutants, and enabling lightweight edge or embedded processing.

All the Declarations and Statements

Author Contributions Statement

Badugu Vimala Victoria – Conceptualization, Methodology, Formal analysis, Data curation, Funding acquisition.
Kamil Reza Khondakar – Project administration; Resources; Software, Validation; Visualization, Roles/Writing - original draft; Writing.
Ravi Kumar Suggala – Supervision, Investigation, Writing - Original Draft, Reviewing and Editing.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Funding Declaration

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Statement

Not Applicable.

Ethical Declarations

This article is a completely original work of its authors; it has not been published before and will not be sent to other publications until the journal's editorial board decides not to accept it for publication.

Acknowledgment

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

Not Applicable.

Abbreviations

The following abbreviations are used in this manuscript:

CNN - Convolutional Neural Network
Dice Coefficient
Faster R-CNN - Faster Region-Based Convolutional Neural Network
HCTSpeckle - Hybrid CNN-Transformer Speckle Denoising Network
RGB - Red Green Blue
U-Net - U-Shaped Convolutional Network
ViT - Vision Transformer
YUV - Luminance and Chrominance Color Space (YUV Color Model)
SE - Squeeze-and-Excitation

References

[1] L. M. Dang, A. S. Sagar, N. D. Bui, L. V. Nguyen, and T. H. Nguyen, "Attention-guided marine debris detection with an enhanced transformer framework using drone imagery," Process Safety and Environmental Protection, vol. 197, Art. no. 107089, 2025.

[2] S. Karthick, and N. Muthukumaran, "Deep RegNet-150 architecture for single image super resolution of real-time unpaired image data," Applied Soft Computing, vol. 162, Art. no. 111837, 2024.

[3] H. Liu et al., "Underwater particle classification detector using Mueller matrix and fluorescence signal," IEEE Journal of Oceanic Engineering, 2025.

[4] N. Deluxni, P. Sudhakaran, M. Alsafyani, and A. Yousef, "Underwater debris segmentation using improved YOLOv12s with recursive efficient layer aggregation and FlashAttention for autonomous underwater vehicle," IEEE Access, vol. 13, pp. 200239–200252, 2025.

[5] P. Russo, F. Di Ciaccio, P. Santaniello, T. Cacace, P. Carcagnì, M. Del Coco, and M. Paturzo, "Encoding holographic data into synthetic video streams for enhanced microplastic detection," IEEE Access, 2025.

[6] S. Galata et al., "Rapid detection of microfibres in environmental samples using open-source visual recognition models," Journal of Hazardous Materials, vol. 480, Art. no. 135956, 2024.

[7] F. Han, J. Yao, H. Zhu, and C. Wang, "Underwater image processing and object detection based on deep CNN method," Journal of Sensors, vol. 2020, Art. no. 6707328, 2020.

[8] Z. He et al., "UISS-Net: Underwater image semantic segmentation network for improving boundary segmentation accuracy of underwater images," Aquaculture International, vol. 32, no. 5, pp. 5625–5638, 2024.

[9] A. A. Alsuwaylimi, "Enhanced YOLOv8-Seg instance segmentation for real-time submerged debris detection," IEEE Access, 2024.

[10] P. Govindarajan et al., "Tracking microplastic pathways: Real-time IoT monitoring for water quality and public health," MethodsX, Art. no. 103674, 2025.

[11] S. Pavithra, "An efficient approach to detect and segment underwater images using Swin Transformer," Results in Engineering, vol. 23, Art. no. 102460, 2024.

[12] K. P. Pushkala and P. Subbulakshmi, "Synergistic integration of vision transformers and advanced segmentation algorithms for panoptic mapping of marine litter," Frontiers in Marine Science, vol. 12, Art. no. 1726472, 2025.

[13] A. Sivaanpu et al., "Speckle noise reduction for medical ultrasound images using hybrid CNN-transformer network," IEEE Access, 2024.

[14] A. Brateanu et al., "Lyt-Net: Lightweight YUV transformer-based network for low-light image enhancement," IEEE Signal Processing Letters, vol. 32, pp. 2065–2069, 2025.

[15] Y. Lee and P. Kang, "Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder," IEEE Access, vol. 10, pp. 46717–46724, 2022.

[16] H. Ni et al., "SDNet: Sandwich decoder network for waterbody segmentation in remote sensing imagery," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 13895–13911, 2025.

[17] "Microplastic_100 dataset," Roboflow, 2023. [Online]. Available: https://universe.roboflow.com/kueranan/microplastic_100

[18] Q. Jiang, Y. Chen, G. Wang, and T. Ji, "A novel deep neural network for noise removal from underwater image," Signal Processing: Image Communication, vol. 87, Art. no. 115921, 2020.

[19] M. Guan et al., "DiffWater: Underwater image enhancement based on conditional denoising diffusion probabilistic model," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 2319–2335, 2023.

[20] I. S. Chandra et al., "CNN based color balancing and denoising technique for underwater images: CNN-CBDT," Measurement: Sensors, vol. 28, Art. no. 100835, 2023.

[21] W. Zhang et al., "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," IEEE Transactions on Image Processing, vol. 31, pp. 3997–4010, 2022.

[22] T. Wang et al., "Underwater image enhancement based on optimal contrast and attenuation difference," IEEE Access, vol. 11, pp. 68538–68549, 2023.

[23] W. Zhang et al., "Underwater image enhancement via weighted wavelet visual perception fusion," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 4, pp. 2469–2483, 2023.

Authors' Profiles

B. Vimala Victoria completed her M.Tech degree from SVECW College, Bhimavaram. She has successfully qualified for the APSET examination, demonstrating her academic excellence and research aptitude. She is currently working as an Assistant Professor at a reputed institute while also Currently, she is a Research Scholar at Woxsen University. Her research focuses on the detection of microplastics in underwater environments using image processing and deep learning techniques. She has actively contributed to the research community by publishing papers in various national and international conferences and journals. Her areas of interest include image processing, deep learning, and environmental monitoring applications.

Kamil Reza Khondakar earned his doctoral degree from the University of Queensland, Australia. His research interests lie at the intersection of Lab-on-Chip, Cancer Diagnostics, Biosensors, Machine Learning, and AI & IoT-enabled Sensors, with a strong emphasis on developing advanced technologies for next-generation healthcare applications. His expertise spans lab-on-chip systems, immunoassays, nanosensors, spectroscopy, and data-driven approaches for medical diagnostics and monitoring. Dr. Khondakar has authored over 60 peer-reviewed publications and edited three academic books, reflecting his significant contributions to the scientific community. He is a recipient of the prestigious Government of India DST-SRG research grant. His research has garnered over 1,583 citations, with an h-index of 25 and an i10-index of 32, demonstrating strong academic impact. In addition to his research and teaching roles, he actively serves as an editor and reviewer for several Scopus-indexed journals, including Frontiers in Nanotechnology and Computational Biomedicine.

Ravi Kumar Suggala is currently working as an Assistant Professor in the Department of Information Technology at Shri Vishnu Engineering College for Women, Bhimavaram, Andhra Pradesh, India. He holds a Ph.D. degree in Computer Science and Engineering from Centurion University of Technology and Management with a focus on machine learning and deep learning applications in health-related fields. With over 15 years of experience in academia, Dr. Suggala has distinguished himself in various administrative roles and is actively involved in research and development initiatives. His expertise spans a range of areas, including machine learning, deep learning, and quantum computing, cyber security and DDoS Attacks in IOT networks with a strong focus on developing innovative solutions and advancing computational techniques in these fields. Dr. Suggala is actively engaged in professional bodies and contributes to the growth of the academic community through various roles. His current research interests include machine learning, deep learning, image processing, quantum computing, and optical character recognition (OCR) for reading and data extraction, with an emphasis on applying these technologies to solve real-world challenges in healthcare, data analytics, and image processing.

How to cite this paper:

Badugu Vimala Victoria, Kamil Reza Khondakar, Ravi Kumar Suggala, "HCTSpeckle and Lightweight YUV Transformer with ViT Encoder-Sandwich Decoder Network for Underwater Microplastic High Resolution Image Segmentation", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.4, pp. 65-81, 2026. DOI:10.5815/ijigsp.2026.04.04