EfficientNetV2-B3-Based Stage-Adaptive Attention Encoder–Decoder Network with Multi-Resolution Feature Fusion for Microplastic Segmentation

Karthik Reddy Duddukunta, Pankaj Somkuwar, Kripa Shrestha, Pankaj Lal

Abstract

Microplastic segmentation is a challenging problem due to extreme foreground–background imbalance, heterogeneous particle morphology, translucent materials, and complex background interference. Existing deep learning–based approaches often rely on conventional encoder–decoder designs with limited capacity to simultaneously address scale variability and ambiguous particle boundaries, resulting in moderate segmentation accuracy under realistic microscopy conditions. In this work, we propose a stage-adaptive attention encoder–decoder network based on an EfficientNetV2-B3 backbone for accurate microplastic segmentation. The proposed architecture integrates a Multi-Resolution Fusion Module (MRFM) with resolution-aware attention mechanisms, enabling adaptive feature refinement across different spatial scales and semantic depths. Specifically, edge-aware attention enhances boundary sensitivity at high-resolution stages, coordinate attention captures elongated and spatially distributed structures at intermediate resolutions, and combined channel-spatial attention improves semantic discrimination at deeper layers. These components are embedded within an end-to-end trainable encoder-decoder framework to produce full-resolution pixel-wise segmentation maps. We evaluate the proposed method on a challenging microplastic dataset characterized by strong background clutter and scale variability. Experimental results demonstrate that the proposed model consistently outperforms recent microplastic segmentation approaches, achieving a mean Intersection over Union (mIoU) of 0.8248 and a Dice coefficient of 0.7932. Ablation studies further confirm that multi-resolution feature fusion and stage-adaptive attention provide complementary benefits, leading to improved boundary delineation and enhanced sensitivity to microplastic regions. Overall, the proposed method advances microplastic segmentation by providing a principled and effective solution to long-standing challenges related to scale diversity and boundary ambiguity, thereby enabling more accurate and reliable environmental microplastic analysis.

Keywords: Microplastic segmentation, EfficientNetV2-B3, Multi-resolution feature fusion, Stage-adaptive attention, semantic image segmentation, Optical microscopy

1. Introduction

Microplastic contamination is now recognized as a widespread and persistent environmental challenge affecting aquatic, terrestrial, and freshwater systems. Early investigations revealed extensive accumulation of plastic debris in marine environments and demonstrated that these materials persist for long periods, progressively fragmenting into microscopic particles through physical degradation processes [1]. Subsequent studies showed that the degradation of plastic debris leads to the formation of micro and nanoplastics with enhanced environmental mobility and bioavailability, intensifying concerns regarding their interactions with organisms, pollutant transport, and broader ecological processes [2]. Previous studies have shown that microplastics are not confined to aquatic systems but are also found in terrestrial environments, wastewater streams, and urban settings, where they can modify soil structure, influence hydrological processes, and alter biogeochemical cycling [3,4]. Collectively, these findings underscore the urgent need for accurate, scalable, and reproducible methods for microplastic detection and quantification across diverse environmental matrices.

Despite this growing urgency, microplastic analysis remains largely dependent on labor-intensive laboratory workflows, including manual microscopy, staining, and spectroscopic validation. Although these approaches provide high analytical specificity, they are time-consuming, susceptible to observer bias, and poorly suited for large-scale analysis of microscopic imagery [5,6]. To address these limitations, recent work has explored computer vision–based pipelines for automated microplastic identification, demonstrating that image-based methods can substantially reduce manual effort and improve analytical consistency [7,8,9]. However, achieving reliable pixel-level delineation remains challenging, particularly for thin fibers, translucent fragments, and overlapping microplastics commonly observed in microscopic images.

Deep learning has substantially advanced pixel-wise image analysis, with fully convolutional networks establishing semantic segmentation as an end-to-end dense prediction task [10]. Encoder-decoder architectures, most notably U-Net, further improved localization performance in biomedical and microscopic imagery by combining contextual and spatial information through skip connections [11]. Building on this foundation, U-Net and related architectures have

been applied to microplastic imagery, demonstrating the feasibility of semantic segmentation under challenging background conditions. However, reported segmentation performance remains moderate, and analyses indicate that skip-connection-focused architectural modifications yield only incremental improvements. These findings motivate the integration of attention-based feature refinement to enhance robustness under realistic microscopy conditions [12,13].

Beyond skip-connection design, multi-scale context modeling has become a central component of modern semantic segmentation methods. Atrous convolution and spatial pyramid pooling have demonstrated the importance of capturing contextual information across multiple receptive fields for dense prediction tasks [14]. In parallel, explicit multi-resolution feature fusion has been shown to improve robustness to object size variation and complex morphological structures [15]. Attention mechanisms further enhance feature representations by selectively emphasizing informative spatial and semantic cues. Channel-wise attention models inter-channel dependencies [16], In addition, boundary-aware segmentation studies highlight the importance of explicitly modeling edges and shape information to accurately delineate thin and elongated structures [17,18], which is particularly relevant for microplastic imagery characterized by weak contrast and ambiguous boundaries. Coordinate attention encodes positional information within channel attention, enabling more effective feature representation for dense prediction tasks [28]. Building on this idea, sequential channel–spatial attention further refines both feature selection and spatial localization [29].

In summary, accurate semantic segmentation of microplastics in microscopic imagery remains challenging due to extreme class imbalance, heterogeneous particle morphology, translucent materials, and complex backgrounds. Existing approaches either rely on U-Net-style architectures with limited representational capacity or prioritize computational efficiency over precise boundary delineation. To address these limitations, we propose an attention-driven, multi-resolution semantic segmentation framework built upon an EfficientNetV2-B3 backbone. By integrating explicit multi-resolution feature fusion with complementary attention mechanisms, including channel, spatial, coordinate, and edge-aware attention, the proposed approach enhances contextual discrimination and boundary precision in challenging microscopic microplastic images. Experimental results demonstrate consistent improvements in segmentation performance, as measured by mIoU and Dice score, while maintaining end-to-end trainability and architectural generality.

2. Research background

Early automated approaches for microplastic analysis explored image-driven detection and categorization using conventional feature extraction and classification techniques. These studies demonstrated the feasibility of visual analysis while highlighting persistent challenges related to background clutter, particle overlap, and substantial morphological variability in microscopic imagery [5–8]. Although such approaches reduced manual effort, their reliance on object-level representations and heuristic post-processing limited accurate delineation of fine-grained particle boundaries.

The emergence of deep learning significantly advanced pixel-wise image analysis through semantic segmentation. Fully convolutional networks reformulated segmentation as a dense prediction task without handcrafted features [10], while encoder–decoder architectures further improved accuracy by integrating high-level semantic context with low-level spatial detail through skip connections [11]. These architectures proved effective in biomedical and microscopic imaging and were subsequently adopted for microplastic segmentation. However, empirical results indicate that architectural refinements focused primarily on skip-connection redesign yield only incremental performance improvements.

To enhance robustness to scale variation and complex morphology, multi-scale context modeling has been widely explored. Atrous convolution and spatial pyramid pooling enable the capture of contextual information across multiple receptive fields without reducing spatial resolution [14], while explicit multi-resolution feature fusion improves robustness to object size variation and structural diversity [15]. Although particularly relevant for microplastic imagery, where particle size and orientation vary substantially, multi-scale fusion alone often remains insufficient for suppressing background interference and resolving ambiguous boundaries.

Attention mechanisms have further advanced semantic segmentation by enabling adaptive feature recalibration. Channel-wise attention improves representational capacity by modeling inter-channel dependencies [16], while boundary-aware segmentation approaches emphasize explicit modeling of edges and shape information to better delineate thin and elongated structures commonly observed in microscopic microplastic imagery [17,18]. Recent developments in convolutional backbone design focus on improving the trade-off between accuracy and efficiency. EfficientNetV2 achieves this through training-aware scaling strategies and fused convolutional blocks, enabling robust hierarchical feature extraction with reduced computational overhead and faster training convergence [19]. While such backbones offer a solid foundation for dense prediction tasks, studies applying lightweight encoders and feature pyramid–based fusion strategies to microplastic segmentation report only moderate performance gains, indicating that backbone efficiency alone is insufficient without complementary attention or boundary-aware modeling [20].

More recently, dataset-driven efforts have enabled systematic benchmarking of microplastic segmentation methods. A large-scale microscopic microplastic dataset with segmentation annotations was introduced, and baseline models incorporating EfficientNet-based architectures with multi-resolution fusion were systematically evaluated [21]. In parallel, attention-based and boundary-aware frameworks that integrate multi-kernel feature representations have demonstrated improved segmentation performance near object boundaries, highlighting the benefits of jointly modeling contextual diversity and boundary information [25]. These results highlight both the difficulty of the task and the limitations of fusion-only designs. Edge-attention guidance has demonstrated improved boundary awareness by explicitly modeling edge information in early encoding stages and propagating it to deeper layers [27]. Coordinate attention strengthens spatial feature modeling by incorporating positional information into channel attention [28], and combined channel–spatial attention further enhances localization by selectively amplifying salient regions and suppressing irrelevant background responses [29].

Overall, despite notable progress, several gaps remain in existing microplastic segmentation research. Multi-scale feature fusion is frequently applied without adaptive attention mechanisms, and attention- and boundary-aware modeling has not yet been systematically integrated into microplastic-specific segmentation frameworks. These limitations motivate the development of unified architectures that combine modern high-capacity backbones with explicit multi-resolution feature fusion and complementary attention mechanisms to jointly address scale variability, boundary ambiguity, and contextual complexity in microscopic microplastic images.

3. Proposed Model

3.1 Overall Architecture

The input to the network is a 512 × 512 × 3 microplastic images (Figure 1). The EfficientNetV2-B3 encoder extracts hierarchical feature representations at five spatial resolutions, which are subsequently refined using MRFM and stage-adaptive attention mechanisms. Edge attention is applied at high-resolution stages, coordinate attention at the intermediate stage, and channel-spatial attention at deeper stages. The decoder progressively restores spatial resolution through transposed convolutions and attention-refined skip connections, producing a final 512 × 512 × 1 segmentation map representing pixel-wise microplastic probabilities.

The proposed architecture is composed of four principal components: an EfficientNetV2-B3 encoder, MRFM, a stage-adaptive attention mechanism, and a U-Net-style decoder with attention-refined skip connections. Together, these components form an end-to-end trainable segmentation framework capable of producing full-resolution pixel-wise prediction maps at 512 × 512 resolution.

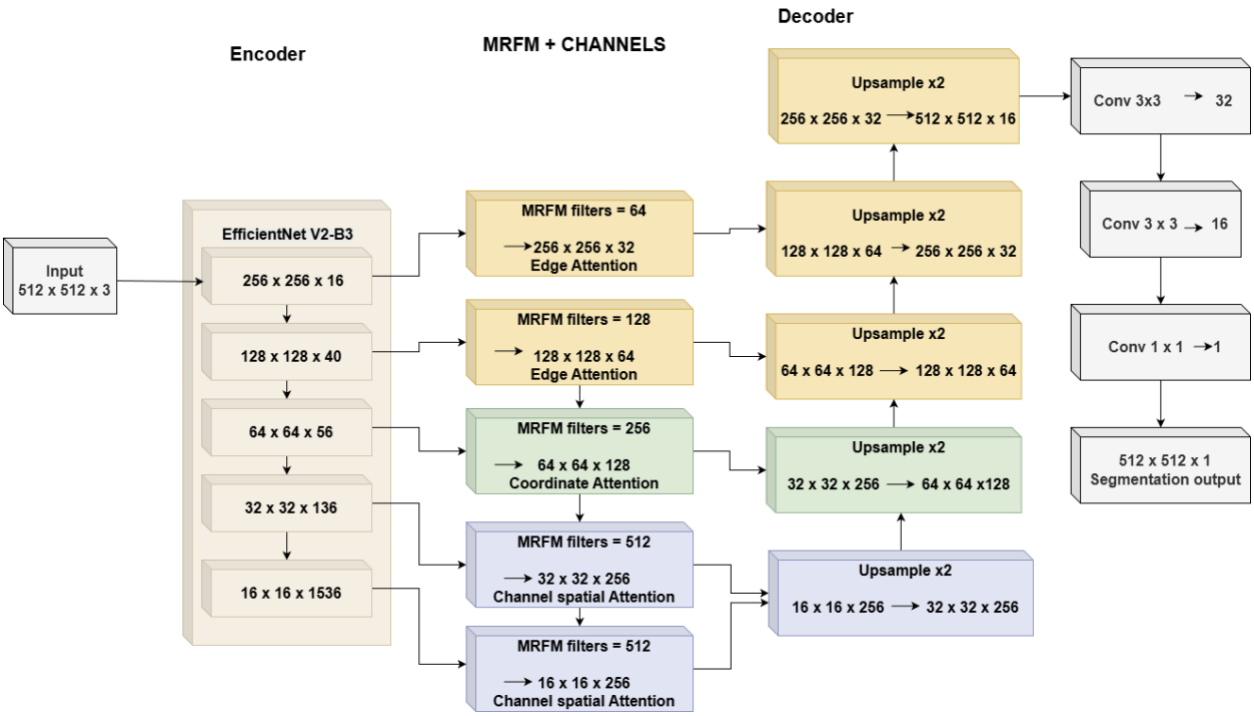

Figure 1. Architecture of the proposed stage-adaptive attention U-Net for microplastic segmentation. The model integrates an EfficientNetV2-B3 encoder, MRFM, and resolution-aware attention mechanisms within an encoder–decoder framework to produce full-resolution segmentation maps.

3.2 Encoder Backbone

We adopt EfficientNetV2-B3 as the encoder backbone due to its strong balance between representational capacity, computational efficiency, and training stability [23]. Recent work on ConvNet modernization demonstrates that well-designed convolutional architectures remain highly effective for dense prediction tasks when feature hierarchy and efficiency are carefully balanced [24]. EfficientNetV2-B3 incorporates training-aware architectural scaling and fused convolutional blocks, enabling efficient extraction of hierarchical features across multiple spatial resolutions. The backbone design further benefits from principles of deep residual learning, which facilitate stable gradient propagation and effective optimization of deep convolutional networks [26]. These findings collectively motivate our choice of EfficientNetV2-B3 as a robust encoder for microplastic segmentation.

We initialize the encoder with pretrained weights and keep it fully trainable, allowing both low-level edge features and high-level semantic representations to adapt to the microplastic domain. Feature maps are extracted at five hierarchical levels with progressively decreasing spatial resolutions:

$$\{E_1, E_2, E_3, E_4, E_5\}, \qquad (1)$$

Equation (1) represents the set of feature maps extracted from successive stages of the encoder backbone. Each feature map E_i corresponds to a different spatial resolution and semantic depth. Earlier feature maps (E_1, E_2) retain high spatial resolution and encode low-level information such as edges and textures, (E_3) encodes mid-level, part-aware, and structurally meaningful features that bridge local texture information, while deeper feature maps (E_4, E_5) encode higher-level semantic concepts with reduced resolution.

Corresponding to spatial resolutions of $256 \times 256, 128 \times 128, 64 \times 64, 32 \times 32, and\ 16 \times 16$, respectively. This hierarchical representation enables the simultaneous capture of fine boundary cues and high-level semantic context, which is essential for accurate segmentation.

3.3 Multi-Resolution Fusion Module (MRFM)

To explicitly address the pronounced morphological diversity of microplastics, we integrate a MRFM at each encoder and decoder stage. Microplastic particles exhibit substantial variability in size, shape, and texture, ranging from thin fibers to irregular fragments, which necessitates feature representations that capture information across multiple receptive field scales.

Given an input feature map X, MRFM applies three parallel convolutional branches:

- A 3×3 convolution for capturing local spatial details,
- A 5×5 convolution for modeling broader contextual structures, and
- A dilated 3×3 convolution ($dilation\ rate = 2$) to enlarge the receptive field without reducing spatial resolution.

This design is inspired by prior studies demonstrating that parallel multi-scale convolutional operations improve robustness to object size variation and complex morphology in semantic segmentation, including atrous convolution and spatial pyramid–based context aggregation strategies [14, 15].

The outputs of these branches are concatenated and compressed using a 1×1 convolution followed by batch normalization and ReLU activation:

$$F_{MRFM} = \emptyset\left(Conv_{1\times1}([B_3 \parallel B_5 \parallel B_d])\right) \qquad equation(2)$$

Equation (2) defines the output of the MRFM. The input feature map is processed in parallel by three convolutional branches:

The symbol $\parallel$ denotes channel-wise concatenation of these branch outputs, A 1×1 convolution is then applied to reduce channel dimensionality and fuse information across scales. The operator $\emptyset(\cdot)$ denotes batch, normalization followed by a ReLU nonlinearity.

To stabilize training and preserve original feature information, we introduce a residual projection:

$$\hat{F} = F_{MRFM} + X_{proj}, \qquad equation(3)$$

Equation (3) introduces a residual connection between the fused multi-resolution features and the original input feature map X. When necessary, X is projected using a 1×1 convolution (X_{proj}) to ensure channel compatibility. Residual learning has been shown to improve optimization stability and gradient propagation in deep convolutional architectures, particularly when integrating complex feature transformations [26].

3.4 Stage-Adaptive Attention Mechanism

A central contribution of this work is the Microplastic Attention block, which we propose as a stage-adaptive attention mechanism that selects attention strategies according to feature resolution and semantic content. Unlike conventional approaches that apply a uniform attention module throughout the network, we tailor the attention mechanism to the characteristics of features at different depths, thereby improving robustness to scale variation and boundary ambiguity in microplastic images.

3.4.1 Edge Attention (Early Stages)

At high-resolution stages ($256 \times 256, 128 \times 128$), feature maps retain rich spatial detail and predominantly capture edges, gradients, and fine-scale textures. Motivated by prior work demonstrating the effectiveness of edge-attention guidance for improving boundary delineation in segmentation networks [27], we introduce an edge-aware attention mechanism to enhance boundary sensitivity. Given an input feature map F, an edge-focused spatial attention map is generated by first applying a 3×3 convolution to capture local gradient and boundary responses, followed by

normalization and non-linear activation. A subsequent 3×3 convolution and sigmoid activation produce a normalized attention map $A_{edge} \epsilon[0,1]$, which selectively highlights edge-relevant spatial locations:

$$A_{edge} = \sigma\left(Conv_{1\times1}\left(\emptyset\left(Conv_{3\times3}(F)\right)\right)\right), \quad equation(4)$$

Equation (4) defines the spatial edge-attention map A_{edge} , where a 3×3 convolution extracts local boundary cues from the input feature map, and a subsequent 1×1 convolution with sigmoid activation produces normalized attention weights.

$$\hat{F} = F \odot A_{edge} \quad equation(5)$$

Equation (5) applies the learned edge-attention map to the input feature map via element-wise multiplication, yielding an edge-enhanced feature representation.

This operation adaptively amplifies edge-dominant regions while suppressing less informative background responses. As shown in prior medical image segmentation studies [27], such edge-attention guidance provides effective fine-grained constraints for boundary modeling. In our framework, this mechanism improves the delineation of thin, fragmented, and low-contrast microplastic boundaries in microplastic imagery.

3.4.2 Coordinate Attention (Middle Stage)

At the intermediate resolution (64×64), features encode object shape and spatial layout. We incorporate Coordinate Attention to capture long-range spatial dependencies while preserving positional information [28]. Directional pooling is defined as:

$$z_h(h) = \frac{1}{W}\sum_{w=1}^{W} F(h,w), \; z_w(w) = \frac{1}{H}\sum_{h=1}^{H} F(h,w) \quad equation(6)$$

Where H and W denote the height and width of the feature map.

Equation (6) defines the directional pooling operations used in the coordinate attention mechanism. Instead of global pooling, features are aggregated separately along the horizontal and vertical directions.

This mechanism is particularly effective for elongated microplastic structures such as fibers.

3.4.3 Channel-Spatial Attention (Deep Stages)

At deeper stages $(32 \times 32 \; and \; 16 \times 16)$, features become highly semantic but spatially coarse. To enhance discriminative capability, we apply a combined channel–spatial attention mechanism, inspired by CBAM [29]. Channel attention is computed as:

$$A_c = \sigma\left(MLP\left(GAP(F)\right) + MLP\left(GMP(F)\right)\right) \quad equation(7)$$

Equation (7) computes channel-wise attention weights by combining global average pooling (GAP) and global max pooling (GMP). These pooled descriptors capture complementary statistical information about each channel and are processed by a shared multilayer perceptron (MLP). A sigmoid activation normalizes the output.

This mechanism emphasizes informative feature channels while suppressing irrelevant ones, improving semantic discrimination in deep network layers.

$$A_S = \sigma(Conv_{7\times7}([Avg(F_c) \parallel Max(F_c)]) \quad equation(8)$$

Equation (8) defines the spatial attention mechanism applied after channel refinement. Channel-refined features F_c are aggregated using spatial average and max pooling, concatenated, and processed by a 7×7 convolution followed by a sigmoid activation.

Spatial attention enables the model to focus on salient spatial regions, improving localization accuracy and suppressing visually similar background structures.

3.5 Decoder and Output layer

We design the decoder following a U-Net-style architecture that progressively restores spatial resolution using transposed convolutions [11]. At each decoding stage, we integrate MRFM and stage-adaptive attention, ensuring that relevant boundary and contextual information is preserved while suppressing irrelevant background features.

Finally, we upscale the decoder output to the original image resolution and apply convolutional layers to generate a single-channel segmentation map. A sigmoid activation produces pixel-wise probabilities for binary microplastic segmentation.

4. Experimental setup

4.1 Dataset Description

The microplastic dataset exhibits substantial variability in both particle appearance and background characteristics. Microplastic particles vary widely in shape, size, orientation, and contrast, while background regions often contain reflective container surfaces, circular boundaries, embossed characters, and residual debris. Such background structures frequently resemble microplastic particles in texture and intensity, increasing the difficulty of accurate foreground-background separation. [5, 6]

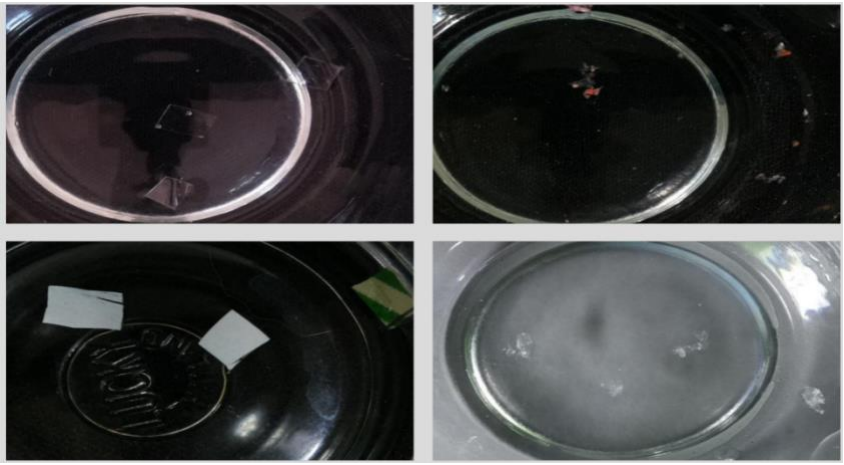

Figure 2: Representative microplastic samples illustrating dataset variability (Source: Microplastic Dataset)

The microplastic particles typically occupy a small proportion of the image area, resulting in pronounced foreground background imbalance (Figure 2). The presence of uneven illumination, partial transparency, and reflective boundaries further complicates particle delineation. Similar challenges have been reported in recent microplastic datasets, where background clutter and visual similarity between particles and surrounding structures were identified as major obstacles for automated segmentation. These characteristics motivate the development of segmentation models that are robust to scale variation, orientation changes, illumination inconsistency, and background interference, which are common in practical microplastic imaging scenarios [5, 6].

4.2 Data Augmentation

To enhance generalization under limited annotated data and diverse microplastic imaging conditions, we employ a two-stage data augmentation strategy consisting of geometric and photometric transformations. All augmentation operations are applied in a label-preserving manner, ensuring strict spatial correspondence between input images and their associated segmentation masks.

Microplastic images are highly sensitive to variations introduced during sample preparation, illumination conditions, and imaging hardware. Prior studies on microscopy-based microplastic segmentation have demonstrated that data augmentation is essential for improving robustness and mitigating overfitting when training deep learning models on limited datasets [5, 12].

4.2.1 Geometric Transformations

Geometric transformations are applied to simulate arbitrary particle orientations and scale variations commonly observed during microplastic imaging and sample handling. These include random rotation, zooming, and flipping, which encourage the network to learn rotation and scale-invariant representations suitable for segmenting microplastic particles with diverse morphologies [9, 12].

4.2.2 Photometric Transformations

Photometric transformations are employed to model appearance variations caused by illumination changes, optical nonlinearity, and sensor noise inherent to microplastic imaging systems. Adjustments to brightness, contrast, gamma, and noise levels improve robustness to intensity fluctuations and background variability, as reported in prior microscopy segmentation studies [5-6]. All geometric transformations are applied identically to both images and segmentation masks, whereas photometric transformations are applied only to images to preserve ground-truth integrity.

Table 1 Summary of the data augmentation strategies used in this study.

Category	Augmentation	Parameter Range
Geometric	Random Rotation	-15° to $+15^{\circ}$
Geometric	Random Zoom	0.9 – 1.1
Geometric	Horizontal Flip	Probability = 0.5
Geometric	Vertical Flip	Probability = 0.5
Photometric	Brightness Shift	±0.05
Photometric	Contrast Adjustment	0.95 – 1.05
Photometric	Gamma Correction	0.95 – 1.05
Photometric	Gaussian Noise	$\sigma = 0.005$

4.3 Evaluation Metrics

Segmentation performance is evaluated using widely adopted pixel-wise binary semantic segmentation metrics, which provide complementary insights into model behavior under imbalanced foreground-background conditions typical of microplastic imagery [12, 24].

Let TP**,** FP**,** TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively, computed at the pixel level.

4.3.1 Dice Coefficient

The Dice coefficient measures the degree of overlap between the predicted segmentation mask and the ground-truth annotation:

$$Dice = \frac{2TP}{2TP + FP + FN} \qquad equation(9)$$

The Dice metric is particularly effective for evaluating segmentation performance on small or thin structures, such as microplastic fibers, where accurate boundary delineation is critical [5, 9].

4.3.2 Mean Intersection over Union (mIoU)

Intersection over Union (IoU), also referred to as the Jaccard Index, is defined as:

$$IoU = \frac{TP}{TP + FP + FN} \qquad equation(10)$$

For binary segmentation, the reported mIoU corresponds to the IoU averaged over foreground and background classes. This metric penalizes both over-segmentation and under-segmentation and is widely used as a benchmark for microplastic segmentation models [12, 24].

4.3.3 Precision

Precision quantifies the proportion of correctly predicted microplastic pixels among all pixels classified as microplastics:

$$Precision = \frac{TP}{TP + FP} \qquad equation(11)$$

High precision indicates a low false-positive rate, which is important for minimizing incorrect labeling of background artifacts as microplastics [6].

4.3.4 Recall

Recall (sensitivity) measures the proportion of actual microplastic pixels that are correctly identified:

$$Recall = \frac{TP}{TP + FN} \quad equation(12)$$

High recall is critical in environmental microplastic analysis, as missed detections may lead to systematic underestimation of microplastic abundance [5, 9].

4.3.5 F1-Score

The F1-score represents the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad equation(13)$$

This metric provides a balanced evaluation when precision and recall exhibit a trade-off, which commonly occurs in segmentation of low-contrast microplastic structures [12].

5. Results

5.1 Qualitative Results and Error Analysis

The qualitative results demonstrate that the proposed model can accurately delineate microplastic particles under challenging imaging conditions, including uneven illumination, partial transparency, and strong background interference caused by reflective container boundaries. Figure 3 presents representative qualitative segmentation results produced by the proposed model. For each example, the original microplastic image, the corresponding ground-truth annotation, the predicted segmentation mask, and the associated error map highlighting true positives, false positives, and false negatives are shown. The Error map analysis indicates that the remaining misclassifications primarily occur near ambiguous particle boundaries or in regions where microplastics exhibit visual similarity to background debris.

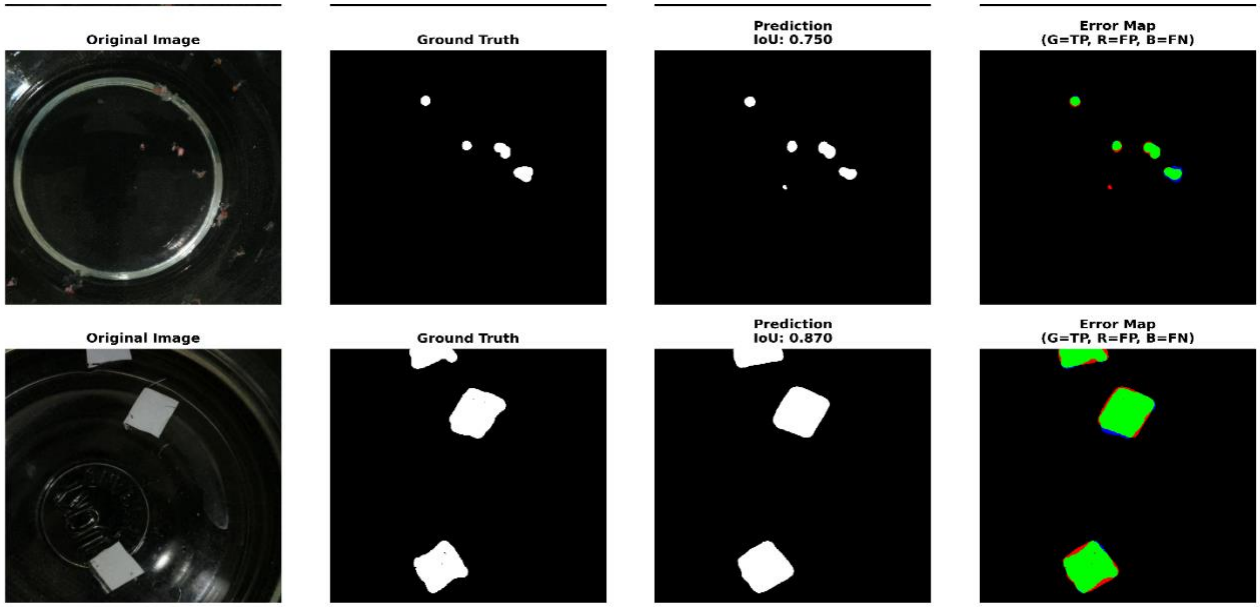

Figure 3 Representative qualitative segmentation results and corresponding error maps produced by the proposed model.

5.2 Ablation Study

To examine the contribution of individual architectural components, an ablation study is conducted by progressively introducing attention mechanisms and the proposed MRFM into a baseline EfficientNetV2-B3 encoder–decoder architecture. All ablation models are trained, validated, and tested under identical experimental settings to ensure a fair comparison.

5.2.1 Experimental Configurations

Three model configurations are evaluated:

1. EfficientNetV2-B3 (Baseline): A U-Net–like segmentation network using EfficientNetV2-B3 as the encoder, without MRFM or attention mechanisms.
2. EfficientNetV2-B3 + Attention: The baseline model augmented with attention modules only.
3. EfficientNetV2-B3 + MRFM + Attention (Proposed): The complete architecture integrating MRFM and stage-adaptive attention mechanisms.

5.2.2 Training performance Analysis

During training, the baseline model achieves a Dice score of 0.7983 (Table 2). Introducing attention mechanisms leads to modest performance improvements, indicating improved feature focusing during optimization. The proposed model achieves the highest training performance, with a Dice score of 0.8158 and an mIoU of 0.8362, suggesting that the joint integration of multi-resolution feature fusion and attention mechanisms enhances feature representation and learning stability. These observations are consistent with prior findings in multi-scale and attention-based segmentation frameworks [17,18].

Table 2 Reports the training performance of the evaluated models.

Model	Dice	mIoU	Precision	Recall	F1
EfficientNetV2-B3	0.7983	0.8232	0.8090	0.7879	0.7983
EfficientNetV2-B3 + Attention	0.8030	0.8265	0.8078	0.7982	0.8030
EfficientNetV2-B3 + MRFM + Attention (Proposed)	0.8158	0.8362	0.8210	0.8107	0.8158

5.2.3 Testing Performance Analysis

As shown Table 3, the proposed model achieves the highest Dice and mIoU scores, along with the highest recall, indicating improved sensitivity to microplastic regions under unseen microplastic imaging conditions. While the attention-only configuration yields recall values comparable to the baseline, it does not improve Dice or mIoU, suggesting that attention mechanisms alone provide limited benefit without explicit multi-scale feature modeling. In contrast, the integration of MRFM leads to consistent gains in recall and overlap-based metrics, demonstrating its effectiveness in capturing scale-variant microplastic characteristics such as thin fibers and irregular fragments.

Table 3 reports the testing performance of the evaluated models.

Model	Dice	mIoU	Precision	Recall	F1
EfficientNetV2-B3	0.7855	0.8194	0.7760	0.7953	0.8221
EfficientNetV2-B3 + Attention	0.7829	0.8176	0.7700	0.7963	0.8208
EfficientNetV2-B3 + MRFM + Attention (proposed)	0.7932	0.8248	0.7648	0.8238	0.8224

Overall, our experimental results demonstrate that multi-resolution feature fusion and stage-adaptive attention play complementary roles in microplastic segmentation. While attention mechanisms enhance feature focusing and sensitivity, the MRFM provides essential scale-aware representations that enable robust segmentation of microplastics with diverse morphologies. By aligning feature fusion and attention strategies with resolution-specific semantic characteristics, the proposed architecture achieves more reliable segmentation performance under realistic microplastic conditions, including cluttered backgrounds, translucent materials, and severe class imbalance

6. Discussion

6.1 Comparison with Existing Methods

Beyond quantitative improvements, these findings have important implications for automated microplastic analysis in environmental monitoring workflows, as improved segmentation accuracy enables more reliable microplastic quantification, size distribution analysis, and morphological characterization while reducing missed detections and analyst-dependent variability. This enhanced capability supports high-throughput and reproducible microplastic assessment, facilitating large-scale environmental surveys and contributing to more accurate contamination monitoring and evidence-based environmental decision-making. To rigorously assess segmentation performance, the proposed model is evaluated under the experimental setting described in this work using the mIoU and Dice coefficient metrics well suited for highly imbalanced foreground–background segmentation tasks and its performance is systematically compared against representative optical microscopy-based microplastic segmentation approaches reported in the literature.

Table 4 Comparison of segmentation performance (mIoU) across different microplastic segmentation models

Model	**mIoU**
MP-Net (U-Net)	0.617
EfficientNetv2-B3 + MRFMx2	0.631
MNv4-Conv-M-FPN	0.690
EfficientNetV2B3-MRFM-Stage Adaptive Attention	**0.8248**

The MP-Net (U-Net) model proposed by Park et al. [13] demonstrates the applicability of fully convolutional architectures to microplastic segmentation using fluorescence microscopy. However, its reliance on a single-scale U-Net backbone limits its ability to effectively handle pronounced scale variability and boundary ambiguity commonly observed in microplastic images, resulting in a reported mIoU of 0.617, as summarized in Table 4. More recently, Yao et al. [20] proposed MNv4-Conv-M-FPN, a lightweight multi-scale architecture incorporating feature pyramid networks, reporting an mIoU of 0.690. While effective in capturing hierarchical contextual information, this approach applies uniform feature refinement across network stages and does not adapt attention strategies to resolution-specific semantic content. Lee et al. [21] extend EfficientNet-based encoders by integrating a MRFMx2, achieving an mIoU of 0.631 (Table 4). These results highlight the benefit of multi-scale feature aggregation for microplastic segmentation. Nevertheless, the absence of attention-based feature refinement may reduce the model's ability to suppress background interference and selectively emphasize microplastic regions, particularly in visually complex microplastic environments. In contrast, the proposed model achieves a higher mIoU of 0.8248, as reported in Table 4, suggesting that combining multi-resolution feature fusion with stage-adaptive attention mechanisms provides more effective feature discrimination across spatial scales and semantic depths. It should be noted that the performance values for existing methods are taken from their respective publications and are included here to provide contextual comparison rather than direct experimental re-evaluation.

6.2 Effect of Multi-Resolution Fusion and Stage-Adaptive Attention

The improved segmentation performance of the proposed model can be attributed to the complementary roles of multi-resolution feature fusion and stage-adaptive attention. MRFM enables effective aggregation of contextual information across multiple spatial resolutions, which is essential for capturing microplastic particles with diverse sizes, shapes, and orientations. However, multi-scale fusion alone is insufficient to fully address background interference and ambiguous particle boundaries. Stage-adaptive attention further enhances segmentation robustness by tailoring feature refinement strategies to resolution-specific semantic characteristics. Attention mechanisms applied at different stages improve boundary sensitivity at high-resolution layers and strengthen semantic discrimination at deeper layers. This adaptive feature modulation allows the network to selectively emphasize microplastic regions while suppressing visually similar background artifacts, leading to more accurate pixel-wise segmentation under challenging microplastic conditions.

6.3 Environmental Implications for Microplastic Monitoring and Assessment

From an environmental perspective, the methodological advances presented in this study have direct implications for the quantification, comparability, and interpretation of microplastic contamination across environmental systems. Accurate micro identification and delineation of microplastics remain a critical bottleneck in environmental assessments, as conventional workflows rely heavily on manual microscopy and expert-driven interpretation, which are time-consuming, prone to observer bias, and difficult to scale [30]. These limitations had always constrained sample throughput and introduced uncertainty into reported microplastic abundances, particularly when dealing with heterogeneous particle morphologies and visually complex backgrounds commonly encountered in environmental samples. Variability in image interpretation, annotation practices, and analytical protocols has been widely recognized as a major obstacle to cross-study comparability and large-scale synthesis of microplastic data [31]. By substantially improving pixel-level segmentation accuracy under realistic microscopy conditions, the proposed framework enables more reliable and reproducible extraction of microplastic occurrence, abundance, and morphology, thereby addressing a key methodological challenge in environmental microplastic monitoring. Likewise, by improving boundary fidelity under challenging imaging conditions, the proposed model enhances the reliability of downstream morphological analyses and supports more accurate environmental characterization of microplastic pollution. Therefore, by enabling accurate, scalable, and reproducible microplastic delineation under realistic microscopy conditions, the proposed

framework supports more reliable contamination assessments and provides a technological foundation for evidence-based environmental management, regulatory assessment, and policy development in the context of plastic pollution.

7. Conclusion

In this work, we presented a stage-adaptive attention U-Net framework for microplastic semantic segmentation, designed to address key challenges such as extreme class imbalance, heterogeneous particle morphology, weak contrast, and complex background interference. The proposed architecture integrates an EfficientNetV2-B3 encoder with a MRFM and resolution-aware attention mechanisms within a unified encoder–decoder framework. By explicitly modeling multi-scale contextual information and aligning attention strategies with resolution-specific semantic characteristics, the proposed method enhances both boundary delineation and semantic discrimination. Experimental results on a challenging microplastic dataset demonstrate that the proposed approach consistently outperforms representative existing methods, achieving notable improvements in mIoU and Dice coefficient. Ablation studies further indicate that multi-resolution feature fusion and stage-adaptive attention play complementary roles in improving segmentation accuracy and robustness.

While the proposed framework demonstrates strong segmentation performance, several aspects warrant further investigation. The design emphasizes robustness through the integration of multi-resolution fusion and stage-adaptive attention mechanisms, and future work may explore efficiency-oriented or streamlined variants to facilitate broader deployment. In addition, the encoder relies on ImageNet pretraining, which may not fully capture domain-specific characteristics of microplastic imagery, suggesting potential benefits from microscopy-specific or domain-adaptive pretraining strategies. The current stage-adaptive attention configuration is guided by spatial resolution and empirical observations rather than being learned dynamically, which may limit flexibility when transferring the framework to datasets with substantially different imaging conditions. Finally, performance comparisons with prior methods are based on reported results from the literature and are intended to provide contextual reference rather than direct experimental re-evaluation. Future research will also investigate multi-class microplastic segmentation, cross-domain generalization, and integration with downstream quantification and environmental assessment pipelines.

7. References

[1] Thompson, R. C., Olsen, Y., Mitchell, R. P., Davis, A., Rowland, S. J., John, A. W., ... & Russell, A. E. (2004). Lost at sea: where is all the plastic? *Science*, *304*(5672), 838-838.

[2] Andrady, A. L. (2011). Microplastics in the marine environment. *Marine pollution bulletin*, *62*(8), 1596-1605.

[3] Dümichen, E., Eisentraut, P., Bannick, C. G., Barthel, A. K., Senz, R., & Braun, U. (2017). Fast identification of microplastics in complex environmental samples by a thermal degradation method. *Chemosphere*, *174*, 572-584.

[4] Rillig, M. C., & Lehmann, A. (2020). Microplastic in terrestrial ecosystems. *Science*, *368*(6498), 1430-1431.

[5] Cabaneros, S. M., Chapman, E., Hansen, M., Williams, B., & Rotchell, J. (2025). Automatic pre-screening of outdoor airborne microplastics in micrographs using deep learning. *Environmental Pollution*, *372*, 125993.

[6] Huang, H., Cai, H., Qureshi, J. U., Mehdi, S. R., Song, H., Liu, C., & Sun, Z. (2023). Proceeding the categorization of microplastics through deep learning-based image segmentation. *Science of the Total Environment*, *896*, 165308.

[7] Lorenzo-Navarro, J., Castrillon-Santana, M., Sanchez-Nielsen, E., Zarco, B., Herrera, A., Martinez, I., & Gomez, M. (2021). Deep learning approach for automatic microplastics counting and classification. *Science of the Total Environment*, *765*, 142728.

[8] Shi, B., Patel, M., Yu, D., Yan, J., Li, Z., Petriw, D., & Howe, J. Y. (2022). Automatic quantification and classification of microplastics in scanning electron micrographs via deep learning. *Science of The Total Environment*, *825*, 153903.

[9] Royer, S. J., Wolter, H., Delorme, A. E., Lebreton, L., & Poirion, O. B. (2024). Computer vision segmentation model-deep learning for categorizing microplastic debris. *Frontiers in Environmental Science*, *12*, 1386292.

[10] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

[11] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.

[12] Lee, G., & Jhang, K. (2021). Neural network analysis for microplastic segmentation. *Sensors*, *21*(21), 7030.

[13] Park, H. M., Park, S., de Guzman, M. K., Baek, J. Y., Cirkovic Velickovic, T., Van Messem, A., & De Neve, W. (2022). MP-Net: Deep learning-based segmentation for fluorescence microscopy images of microplastics isolated from clams. *Plos one*, *17*(6), e0269449.

[14] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

[15] Ibtehaz, N., & Rahman, M. S. (2020). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural networks*, *121*, 74-87.

[16] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

[17] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

[18] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018, September). Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis* (pp. 3-11). Cham: Springer International Publishing.

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

[20] Yao, Y., Xu, W., & Fan, H. (2025). A Deep Learning Approach for Microplastic Segmentation in Microscopic Images. *Toxics*, *13*(12), 1018.

[21] Lee, G., Jung, J., Moon, S., Jung, J., & Jhang, K. (2024). Microscopic Image Dataset with Segmentation and Detection Labels for Microplastic Analysis in Sewage: Enhancing Research and Environmental Monitoring. *Microplastics*, *3*(2), 264-275.

[22] Lee, K. S., Chen, H. L., Ng, Y. S., Maul, T., Gibbins, C., Ting, K. N., ... & Camara, M. (2022). U-Net skip-connection architectures for the automated counting of microplastics. *Neural Computing and Applications*, *34*(9), 7283-7297.

[23] Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning* (pp. 10096-10106). PMLR.

[24] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976-11986).

[25] Zhou, X., Wu, G., Sun, X., Hu, P., & Liu, Y. (2024). Attention-based multi-kernelized and boundary-aware network for image semantic segmentation. *Neurocomputing*, *597*, 127988.

[26] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[27] Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., & Shao, L. (2019, October). Et-net: A generic edge-attention guidance network for medical image segmentation. In International conference on medical image computing and computer-assisted intervention (pp. 442-450). Cham: Springer International Publishing.

[28] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13713-13722).

[29] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

[30] Gqomfa, B., Dlangamandla, C., Ntwampe, S. K., Chidi, B. S., & Maphanga, T. (2025). Current techniques for identifying, quantifying, and characterizing micro and nanoplastics with emphasis on strengths, limitations, and challenges. Discover Environment, 3(1), 282.

[31] Semensatto, D., Passos, C. C., Bicalho, C. S., Mendes-Silva, L. P., & Labuto, G. (2025). Methodological similarities and discrepancies among studies on microplastics in South American continental aquatic environments. Anais da Academia Brasileira de Ciências, 97(Suppl. 3), e20241459.