Enhancing Sea Turtle Segmentation: A Comparative Study of U-Net, ResNet, and Mask R-CNN with Attention Mechanisms

Chenhao Su
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5503282@ad.unsw.edu.au

Jo Jin
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5510401@ad.unsw.edu.au

Ivan Luk
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5463348@ad.unsw.edu.au

Xuedong Zhang
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z5582401@ad.unsw.edu.au

Yi Han
Computer Science and Engineering
University of New South Wales
Sydney, Australia
z53652201@ad.unsw.edu.au

***Abstract*—This study explores the segmentation of sea turtles in underwater images using U-Net and Mask R-CNN architectures, with modifications to enhance segmentation accuracy. The U-Net model showed strong performance in segmenting fine details, especially for small and irregular structures like flippers and heads, by incorporating a combined loss function and attention mechanisms. Mask R-CNN, with a ResNet-50 backbone, excelled in multi-scale segmentation, particularly in distinguishing individual instances in complex scenes. However, adding attention mechanisms led to mixed results, highlighting sensitivity to noise and background complexity in underwater conditions. The findings indicate the strengths and limitations of each model, suggesting further refinements and potential hybrid approaches for improved performance in real-world applications.**

***Index Terms*—Sea turtle segmentation, U-Net, Mask R-CNN, attention mechanisms, underwater image analysis.**

I. Introduction

Image segmentation, a fundamental task in computer vision, involves dividing images into meaningful regions or objects. This field has evolved significantly, with deep learning approaches largely replacing traditional methods due to their superior accuracy and efficiency, which is especially beneficial in applications like medical imaging, autonomous driving, and environmental monitoring [1]. In the context of wildlife monitoring, segmentation automates the otherwise labour-intensive and error-prone task of manually analysing images, facilitating more accurate and scalable tracking.

This project focuses on developing automated segmentation methods to identify key body parts of sea turtles—such as the head, flippers, and carapace—from photographs. Accurate segmentation of these features is essential for applications in marine biodiversity conservation, individual turtle tracking, and behavioural studies. Given the high demands of manual segmentation in these applications, an effective automated approach is crucial for improving efficiency and reducing human error.

Our research is based on the SeaTurtleID2022 dataset [2], which comprises 8,729 images of 438 unique sea turtles collected over 13 years. This dataset's rich annotations support deep learning model training and robust evaluation. Moreover, its unique open-set splitting method provides a realistic test of model generalisation, simulating real-world conditions by assessing performance beyond closed or random set divisions. This structure is essential for long-term individual recognition and enables us to rigorously evaluate the scalability of segmentation models in authentic scenarios.

The primary aim of this project is to compare the effectiveness of several deep learning-based segmentation models, including U-Net, ResNet, and Mask R-CNN, in accurately identifying these specific regions in sea turtles. Additionally, we explore the role of channel and attention mechanisms in enhancing segmentation precision. Through a comparative analysis of model performances, this study aims to highlight effective techniques and provide insights for improving wildlife image segmentation. This project ultimately contributes to the development of scalable and precise tools in biological image analysis, supporting conservation and monitoring efforts in marine environments.

II. Literature Review

The segmentation of animal images has become a fundamental tool in wildlife conservation, population monitoring, and individual identification. Traditionally, segmentation has relied on classical image processing techniques, such as thresholding, region growing, and edge detection [3] [4]. However, these methods fall short when applied to complex

backgrounds, overlapping objects, or varying poses, which are common in wildlife imagery. To address these challenges, machine learning techniques, including Gaussian Mixture Models (GMM) and Pearson Type VI Distribution (PTVID)-based methods, have been introduced. Studies have shown that PTVID-K and PTVID-H outperform GMM in segmentation accuracy and boundary consistency across various animal images [5], proving effective in enhancing segmentation quality in complex environments. However, machine learning techniques like GMM and PTVID often face limitations in generalisability and scalability, particularly when applied to large and diverse datasets. These methods rely on handcrafted features and statistical assumptions that may not fully capture complex patterns in real-world images, making them sensitive to variations in lighting, texture, and noise, which limits their robustness in natural settings [6].

A. Deep Learning-Driven Segmentation Models

The evolution of deep learning models has significantly enhanced image segmentation capabilities, particularly through convolutional networks and, more recently, Transformer-based architectures. Early breakthroughs in fully convolutional networks (FCNs) paved the way for dense pixel-wise predictions by adapting classification networks such as AlexNet, VGG, and GoogLeNet [7]. FCNs utilised both deep semantic information and shallow appearance features to improve segmentation accuracy, achieving 62.2% mean Intersection over Union (mIoU) on PASCAL VOC.

Building on FCNs, a DeepLab model introduced atrous convolution to control the resolution of feature responses without increasing computation [8], along with atrous spatial pyramid pooling (ASPP) for capturing multi-scale context. Additionally, it incorporated Conditional Random Fields (CRFs) to improve boundary localisation. DeepLab achieved an impressive 79.7% mIoU on the PASCAL VOC 2012 dataset, setting a new standard in semantic segmentation at the time .

U-Net was designed specifically for biomedical image segmentation. It introduced a symmetric encoder-decoder architecture with skip connections that allowed for detailed segmentations even with limited data. On the ISBI cell tracking challenge, U-Net achieved impressive IoU scores of 0.9203 and 0.7756 on the PhC-U373 and DIC-HeLa datasets, respectively, setting a benchmark in medical image segmentation [9]. The model's success inspired further innovations, such as UNet++ [10], which introduced nested dense skip pathways to address the semantic gap between encoder and decoder features, yielding better performance in complex tasks, including liver and lung nodule segmentation, with IoUs of 82.9% and 77.21%.

Attention mechanisms also enhanced segmentation by allowing models to focus selectively on relevant regions [11]. The Attention U-Net incorporated attention gates, which suppressed irrelevant areas and enhanced feature extraction for target structures. On the TCIA Pancreas-CT dataset, this model achieved a Dice Score (DSC) of 0.831, outperforming the standard U-Net's DSC of 0.804, demonstrating its utility in precise medical applications.

Mask R-CNN became widely adopted for instance segmentation in complex environments due to its robust two-stage architecture. In agriculture, Mask R-CNN has proven to be reliable, delivering high precision and recall even in challenging conditions. Take orchard segmentation, for example: Mask R-CNN hit a precision score of 0.85 and a recall of 0.88 when identifying individual apples [17]. While newer models like YOLOv8 offer a slight bump in accuracy for this specific task, Mask R-CNN's sturdy design makes it a top choice when precision, especially for clearly outlining object boundaries is critical.

In wildlife research, the SeaTurtleID2022 dataset has paved the way for using deep learning to identify and segment different species [2]. Researchers tested models like Mask R-CNN and HTC with backbones such as ResNet-50 and Swin-B. However, the standout was Mask2Former with a ResNet-50 backbone, achieving an impressive mean Average Precision (mAP) of 0.892 and an IoU of 0.977 for segmenting turtles. This result shows just how effective specialised architectures can be in handling complex ecological data with high precision.

The Swin Transformer brings a fresh approach to vision tasks, introducing a Transformer-based architecture with a unique hierarchical structure and a shifted window method. This design lowers the computational demands of self-attention, making Swin Transformer an excellent choice for high-resolution images [12]. It achieved a top mIoU of 53.5 on the ADE20K dataset for semantic segmentation and has been effective across a range of tasks, including object detection and segmentation on the COCO dataset.

These examples highlight the shift in segmentation models from traditional convolutional methods to more advanced Transformer-based architectures. Each of these approaches tackles specific challenges in image segmentation and continues to push performance across various applications.

B. Advanced Vision Foundation Models for Segmentation

Vision foundation models have taken segmentation a step further, using extensive pre-training and in-context learning to improve generalization across domains. The Segment Anything Model (SAM) [13] is a flexible, prompt-based framework that sets a new benchmark for broad segmentation tasks, while still allowing fine-tuning for specific needs. For instance, PerSAM [14] tailors SAM for one-shot segmentation using target-guided attention and semantic prompts. In fact, its fine-tuned version, PerSAM-F, achieved a remarkable mIoU of 95.3 by adjusting just two parameters—showing just how adaptable SAM is for domain-specific segmentation with minimal retraining.

Building on SAM's versatility, models like SegGPT and Matcher are exploring the limits of general-purpose segmentation. SegGPT [15] excels at varied segmentation tasks by learning contextual relationships across samples, allowing it to handle both in-domain and out-of-domain tasks without requiring specialized training. In tests on datasets like YouTube-VOS 2018, SegGPT performed exceptionally in video segmentation

without video-specific data. Similarly, Matcher [16] uses in-context examples for one-shot and few-shot segmentation, achieving a 52.7% mIoU on COCO-20i with a single example, surpassing previous models by 1.6%. These foundation models highlight the growing power of adaptable segmentation solutions for a wide range of challenges.

III. Methods

In our research, we tested various deep learning architectures and configurations to improve segmentation performance on our dataset. Each model was chosen for its unique strengths in segmentation, and we made targeted adjustments to optimise results for our specific application. This section details the motivation and theoretical basis for each chosen method, including U-Net and Mask R-CNN, as well as the implemented modifications.

A. U-Net and U-Net Variants

U-Net, originally designed for biomedical image segmentation [9], has a symmetric encoder-decoder structure with skip connections that allow high-resolution information from the encoder to flow directly to the decoder, enhancing detail preservation during upsampling. This architecture has proven effective for tasks requiring high precision and fine object boundaries. We selected U-Net as our baseline model due to its robust performance in segmentation, particularly with limited labeled data, making it suitable for our application.

Optimisations to U-Net included multiple modifications: baseline adjustments, optimiser changes, backbone integration, and attention mechanisms:

- **Baseline U-Net**: The standard U-Net model served as our baseline, providing a reference for evaluating the impact of subsequent modifications.
- **Loss Function Optimisation**: We incorporated a combined loss function using Cross-Entropy and Focal Tversky Loss. Cross-Entropy loss is commonly used in segmentation tasks to handle pixel-wise classification, while Focal Tversky Loss addresses class imbalance by focusing on challenging pixels, enhancing the performance on hard-to-segment regions. Different body parts (e.g., head, flippers, shell) may vary significantly in pixel count, which could lead the model to focus on larger regions while neglecting smaller ones (such as flippers). Tversky Loss is particularly good at addressing this imbalance by guiding the model's attention to challenging areas, such as partially occluded body parts or edges near the underwater background. This optimisation can improve the model's performance in complex areas, helping it to accurately segment partially obscured turtle parts.
- **Adam with Weight Decay and Learning Rate Scheduler**: To improve optimisation, we employed the Adam optimiser with weight decay, which prevents overfitting by penalising large weights. This is particularly crucial in underwater segmentation tasks due to the variations in lighting and noise. Additionally, we introduced a learning rate scheduler to adapt the learning rate during training, helping the model converge faster and more effectively by gradually reducing the learning rate when a performance plateau is detected. This fine-tuning can enhance edge detection accuracy, helping the model to better recognise the edges of turtles under varying lighting conditions and environments.
- **ResNet-34 as U-Net Backbone**: Instead of the traditional U-Net encoder, we replaced it with ResNet-34, a residual network that provides deeper feature extraction with fewer parameters than ResNet-50, which is especially useful for processing unclear or dark images. The residual connections in ResNet-34 improve gradient flow during training, allowing the network to capture more intricate features and enhancing segmentation accuracy [17]. In images where the turtle is blurry or has a complex background, ResNet-34 can help the model distinguish the turtle from the background. This optimisation likely improves the model's performance on dark or less distinct turtle images, making it more accurate at segmenting the main contours of the turtle.
- **Channel and Spatial Attention Modules**: To further refine feature selection, we added both channel and spatial attention modules within the U-Net architecture. Channel attention emphasises important feature channels, while spatial attention focuses on key spatial regions within the feature maps. Attention mechanisms have been shown to improve segmentation accuracy by directing the network's focus to the most relevant parts of the image [18]. This is beneficial for images where the background is complex, or the turtle's contours are easily confused with the environment. For instance, underwater images might contain various plants or sediment that interfere with segmentation; spatial attention helps the model ignore these background noises, allowing it to capture the edges and local details of the turtle more accurately. It also improves segmentation of clear boundaries, like the flippers and head, leading to smoother and more precise edges.

These modifications were intended to enhance U-Net's ability to handle challenging segmentation tasks by improving its focus on relevant features, optimising training, and balancing class representations within the loss function.

B. Mask R-CNN and Variants

Mask R-CNN is a well-established model for instance segmentation, where each object in an image is independently identified and segmented [19]. This two-stage model includes a region proposal network (RPN) for detecting regions of interest (RoIs) and a mask prediction head for generating pixel-wise segmentation masks for each detected object. We chose Mask R-CNN for tasks that require precise outlining of objects, especially when dealing with overlapping structures or complex boundaries [20].

Here's a breakdown of the Mask R-CNN variants we implemented and tested:

- **Mask R-CNN with ResNet-50 and FPN**: In this setup, Mask R-CNN uses ResNet-50 as its backbone network, combined with a Feature Pyramid Network (FPN). The FPN enhances Mask R-CNN by building a feature pyramid, which captures high-level semantic details across multiple scales. This setup is particularly strong for detecting objects of different sizes, as the FPN helps Mask R-CNN pick up details at both large and small scales. ResNet-50 handles feature extraction effectively, while the FPN preserves important spatial details—key for accurate segmentation. This combo performs well in situations where objects are complex and vary in size, like segmenting turtles in underwater scenes, where they show up in a range of sizes and poses.
- **SEBlock Attention Mechanism**: To push segmentation quality even further, we added SEBlock attention modules to the Mask R-CNN framework. Attention modules, similar to their use in U-Net, direct the model's focus to relevant regions within the RoIs, enhancing precision in challenging areas. The integration of attention modules helps Mask R-CNN prioritise important regions, leading to more accurate and efficient segmentation in capturing clear edges and small structures in complex underwater scenes.

These variants were explored to leverage Mask R-CNN's instance segmentation capabilities while improving its focus on target structures, particularly in cluttered or overlapping regions, through the integration of attention and multi-scale feature extraction.

IV. Experiment and Results

A. Data Exploration and Preprocessing

The dataset we used primarily consists of underwater images of sea turtles with natural blue backgrounds. Sea turtles in these images typically exhibit consistent colour patterns (Figure 1), but significant variability in lighting, scale, orientation, and pose makes segmentation challenging. The turtles have irregular boundaries, especially around the flippers and shells, necessitating precise edge detection. Moreover, underwater conditions often introduce blurriness or noise due to water particles or camera movement. The dataset includes turtles in different orientations, poses, and angles, while turtles may appear at various distances from the camera, causing variations in scale. These factors require the model to be robust against variations in rotation, scale, and partial visibility. Given these challenges, our augmentation strategy involved subtle transformations, aiming to maintain natural appearance while enhancing model robustness.

For U-Net, the dataset was split into 5303 images for training, 1118 for validation, and 1000 for testing. Basic preprocessing steps included resizing each image to a standard input size (128x128). For Mask R-CNN, we used a dataset split of 2400 images for training, 1200 for validation, and 1000 for testing, as Mask R-CNN has a higher learning capacity compared to U-Net, it can achieve effective fitting with a smaller dataset, reducing the risk of overfitting that might occur with larger amounts of data. Training data augmentations included random horizontal flips to increase data diversity, whereas validation and test data underwent basic preprocessing to maintain consistency during evaluation.

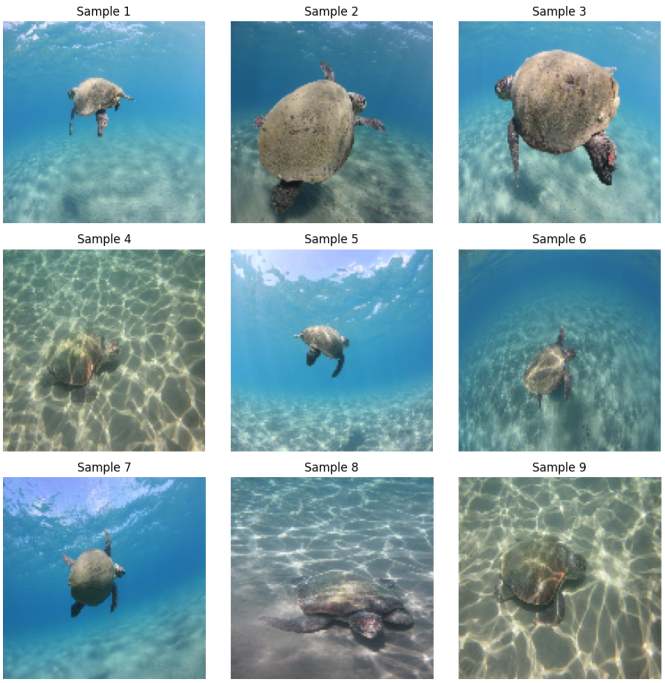

Fig. 1. Dataset Observation

B. Experiment Setup

In our experiments, first, we adjusted the loss function and optimiser for the baseline U-Net, then pursued two enhancement directions: (a) replacing the backbone with ResNet-34, and (b) adding an attention mechanism. For Mask R-CNN, we combined it with a ResNet-50 backbone and subsequently introduced an attention mechanism. Figure 2 outlines the overall experimental process.

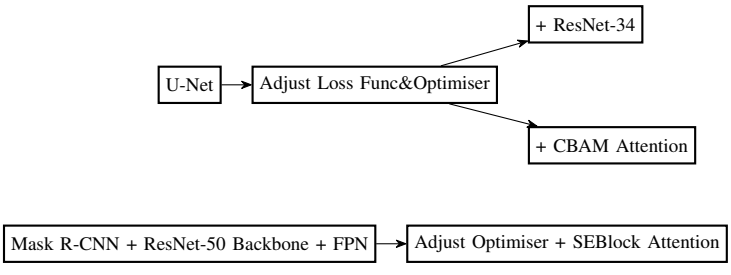

Fig. 2. Workflow of Model Configurations

The training and evaluation of all U-Net models, including the baseline, loss function and optimiser adjustments, ResNet-34 backbone, and CBAM attention, were performed on an NVIDIA GeForce RTX 4070 using PyTorch. Training time was approximately 15 mins for each configuration, with testing

times of 20–40 seconds per model. The training and evaluation of all Mask R-CNN configurations, including the ResNet-50 backbone and SEBlock attention variant, were performed on an NVIDIA GeForce RTX 4090 using PyTorch. Training took approximately 1.5 hours per configuration, with testing times of around 30 minutes per model.

The experiment setup for each to achieve superior segmentation performance is following:

- **Baseline U-Net**: We initially trained a standard U-Net with 23 convolutional layers, including 4 up-convolutional layers in the decoder. The model was configured to use Cross-Entropy Loss and the Adam optimiser with a learning rate of 0.001 for baseline segmentation. The model was trained for 40 epochs, and early stopping was applied to monitor validation loss to ensure convergence.
- **U-Net + Loss Function and Optimiser Adjustments**: To handle class imbalance, we integrated a combined loss function using Cross-Entropy and Focal Tversky Loss (with a Tversky index parameter $\alpha = 0.5$, $\beta = 0.5$, and $\gamma = 0.75$). This combined approach aimed to improve segmentation on complex or imbalanced regions by prioritising harder-to-classify areas. Additionally, the Adam optimiser was enhanced with weight decay of 1e-4, and a ReduceLROnPlateau scheduler with a patience of 3 epochs and a decay factor of 0.5 adaptively adjusted the learning rate based on validation loss trends, ensuring stable convergence. The initial learning rate was set to 0.001 with a minimum learning rate of 1e-6. ResNet and CBAM Attention adjustments are based on this modification.
- **U-Net + ResNet-34 Backbone**: To improve feature extraction, the U-Net's encoder was replaced with a pre-trained ResNet-34 backbone. This architecture uses deep residual connections, which allow the model to capture complex details without overwhelming the parameter count.
- **U-Net + CBAM Attention Mechanism**: We added Channel and Spatial Attention Modules (CBAM) to the U-Net architecture to improve feature learning in the decoder stages. These CBAM modules help the model focus on the most relevant areas within each feature map, which leads to higher-quality segmentation. This is especially useful in situations with a lot of background noise. CBAM helps the model "tune out" irrelevant details and concentrate on the important parts.
- **Mask R-CNN + ResNet-50 Backbone + FPN**: We also tested a Mask R-CNN model with a ResNet-50 backbone paired with a FPN. The FPN enables multi-scale feature extraction, which is essential for accurately identifying objects of various sizes within an image. This setup performs especially well when object sizes vary widely, making it ideal for applications with diverse scale requirements. The training process used Stochastic Gradient Descent (SGD) as the optimiser, with a learning rate of 0.005, momentum of 0.9, and weight decay of 0.0005. A StepLR scheduler reduced the learning rate by a factor of 0.1 every three epochs. The model was trained for 8 epochs.
- **Mask R-CNN + SEBlock Attention Mechanism**: To refine feature representation in Mask R-CNN, we introduced SEBlock (Squeeze-and-Excitation) modules into specific layers (layers 2 and 3) of the ResNet-50 backbone. SEBlock recalibrates channel-wise feature weights, helping the model focus on the most informative channels, particularly within the Feature Pyramid Network. This modification aimed to improve segmentation accuracy by selectively enhancing critical features. Adam optimiser was applied in this model.

Through these setups, our models were trained and evaluated under optimised configurations tailored to the unique challenges of sea turtle segmentation. These modifications in both U-Net and Mask R-CNN aimed to balance model complexity and performance, addressing specific dataset characteristics such as background noise, class imbalance, and variability in object size and orientation.

C. Results

We evaluated models aiming to maximise mean Intersection over Union (mIoU) across three categories: turtle body, flippers, and head. Table I summarises the mIoU scores for each model and its specific configuration.

TABLE I
MIOU SCORES ACROSS DIFFERENT MODELS AND CONFIGURATIONS FOR TURTLE SEGMENTATION.

Model	Turtle	Flippers	Head	mIoU
Baseline U-Net	0.8658	0.7221	0.7592	0.7824
U-Net + Loss/Optimiser	0.8840	0.7621	0.8055	0.8172
U-Net + ResNet-34	0.8740	0.7354	0.7600	0.7898
U-Net + CBAM Attention	0.8938	0.7830	0.8180	0.8316
Mask R-CNN + ResNet-50	0.9307	0.8941	0.9102	0.9117
Mask R-CNN + Attention	0.7832	0.7340	0.8269	0.7814

Initially, a baseline U-Net yielded mIoU scores of 0.8658 for the turtle body, 0.7221 for the flippers, and 0.7592 for the head, with training loss of 0.0261 and validation loss of 0.0503. This model demonstrated solid performance but left room for improvement, particularly in segmenting the flippers and head.

We modified the loss function by combining cross-entropy with focal Tversky loss, and replaced the Adam optimiser with AdamW and incorporated a learning rate scheduler. This optimisation improved mIoU scores across all classes, with the turtle body reaching 0.8840, flippers 0.7621, and head 0.8055, leading to an overall mean mIoU increase from 0.7824 to 0.8172. The model's final training loss was 0.1105, and the validation loss was 0.2168 at epoch 30/40, indicating a well-generalised model with reduced class imbalance issues, particularly benefiting segmentation accuracy in complex areas like flippers and head. The use of early stopping helped in halting training at an optimal point, preventing overfitting.

In an effort to leverage richer feature extraction capabilities, we modified the U-Net with a ResNet-34 backbone. However, this change resulted in a slight decrease in overall performance, with a mean mIoU dropping to 0.7898. The model achieved different mIoU scores for each part of the turtle: 0.8740 for the body, 0.7354 for the flippers, and 0.7600 for the head. The final training and validation losses for the model settled at 0.1121 and 0.2573 by epoch 26 out of 40. This indicates that, while the model did a solid job capturing the main features, it struggled a bit with the finer details, especially when it came to segmenting smaller structures. This drop in performance could be due to a mismatch between the ResNet-34 backbone's strengths and the specific needs of our dataset.

Adding the CBAM made a big difference. This adjustment led to an improved mean mIoU of 0.8316 overall. With CBAM, the model's mIoU scores went up to 0.8938 for the turtle body, 0.7830 for the flippers, and 0.8180 for the head. Final training and validation losses were 0.1236 and 0.1943 by epoch 40 out of 40, reflecting a well-balanced model that could effectively focus on important spatial features. The attention layers helped the model home in on relevant areas and ignore background noise, improving accuracy in challenging spots. We didn't need early stopping here because the model kept improving up to the last epoch.

The highest mIoU scores in our study came from the Mask R-CNN with a ResNet-50 backbone and FPN. This setup reached an impressive 0.9307 for the turtle body, 0.8941 for the flippers, and 0.9102 for the head. Training and validation losses were recorded at 0.2384 and 0.3233, respectively, by epoch 8 out of 8. This combination excelled at instance segmentation, handling multi-scale features and complex underwater environments with ease. The model's strong generalisation ability, as shown by the validation loss, highlights its suitability for accurately segmenting distinct regions of the turtle.

When we introduced SEBlock attention to Mask R-CNN, the mIoU scores dropped to 0.7832 for the turtle body, 0.7340 for the flippers, and 0.8269 for the head, with an overall mIoU of 0.7814. The final training and validation losses were 0.4060 and 0.6264, respectively, at epoch 8/8. The higher validation loss suggests that the added attention mechanism may have led the model to focus on irrelevant features, thereby reducing its effectiveness in segmenting the target areas. This increase in validation loss indicates potential overfitting to background noise or less relevant features, which diminished the model's ability to generalise effectively.

V. Discussion

Both U-Net and Mask R-CNN demonstrate strong segmentation performance on high-quality images (Figure 3), where the underwater environment is clear, well-lit, and the turtle's body outline is distinct. In such conditions, both models accurately capture the boundaries of key body parts, including the head, flippers, and carapace, with minimal deviation from the ground truth masks. The segmentation masks reveal high IoU scores for all body parts, indicating that both U-Net and Mask R-CNN effectively handle images with bright lighting and clear contours. This suggests that under optimal conditions, both models are well-suited for precise segmentation tasks, showcasing their strengths in capturing detailed features of the turtle's anatomy.

Fig. 3. Good Segmentation Result

A. U-Net Performance

U-Net's segmentation results showed notable successes in capturing fine details. In successful segmentations, U-Net accurately delineated these regions, achieving clear, well-defined edges that underscore its effectiveness in detailed segmentation tasks. Representative examples illustrate that U-Net maintained the structural integrity of the turtle's body and appendages. This success can be attributed to the skip connections in U-Net's architecture, which allow it to combine high-resolution encoder features with decoder layers, preserving the fine-grained information necessary for accurately segmenting complex shapes.

However, U-Net also exhibited segmentation failures, particularly in cases where it struggled with blurred or incomplete boundaries (Figure 4). These failures were more frequent

in images with challenging underwater lighting or blurriness caused by particles in the water. Such conditions reduce the clarity of pixel-level features, making it difficult for U-Net to precisely distinguish the turtle from the background. Additionally, U-Net's reliance on lower-level contextual information may limit its ability to capture high-level spatial relationships, leading to missed or inaccurately defined boundaries, especially in regions with irregular shapes like flippers or areas partially obscured by shadows or reflections.

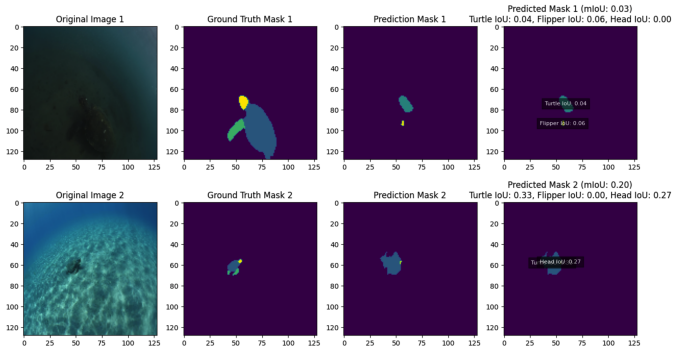

Fig. 4. U-Net Baseline Cannot Fully Recognise Complex Underwater Conditions

To enhance U-Net's performance, we implemented a series of optimisations aimed at improving its segmentation capabilities in challenging areas. Initially, we combined cross-entropy with focal Tversky loss, which resulted in a notable improvement in segmenting difficult regions like fins and heads. This enhancement allowed the model to better handle class imbalance and achieve higher overall accuracy. In the second optimisation attempt, performance took a slight dip. This may be due to a mismatch between ResNet's high-level feature extraction—originally trained on broad datasets like ImageNet—and the specific needs of turtle segmentation. ResNet tends to focus on broader, more general features, which can sometimes miss the fine details needed for accurately segmenting smaller structures. The depth of this model can sometimes increase the risk of overfitting, especially when working with a smaller dataset.

Adding a CBAM attention mechanism, however, led to the best outcomes. CBAM allowed the model to zero in on the most important spatial features, improving segmentation accuracy across all parts of the turtle. This result highlights how attention layers can help direct the model's focus, capturing the detail needed for complex segmentation tasks. With CBAM, the model did a better job distinguishing and capturing fine details, especially around intricate edges, by filtering out background noise and concentrating on essential regions.

A comparative analysis between the U-Net baseline model and the CBAM-enhanced U-Net model further highlights these improvements. As illustrated in Figure 5, the CBAM model demonstrates significantly better segmentation performance in challenging underwater conditions, especially in capturing fine details in environments that are dark, blurry, or have irregular lighting. By incorporating CBAM, the modified model is

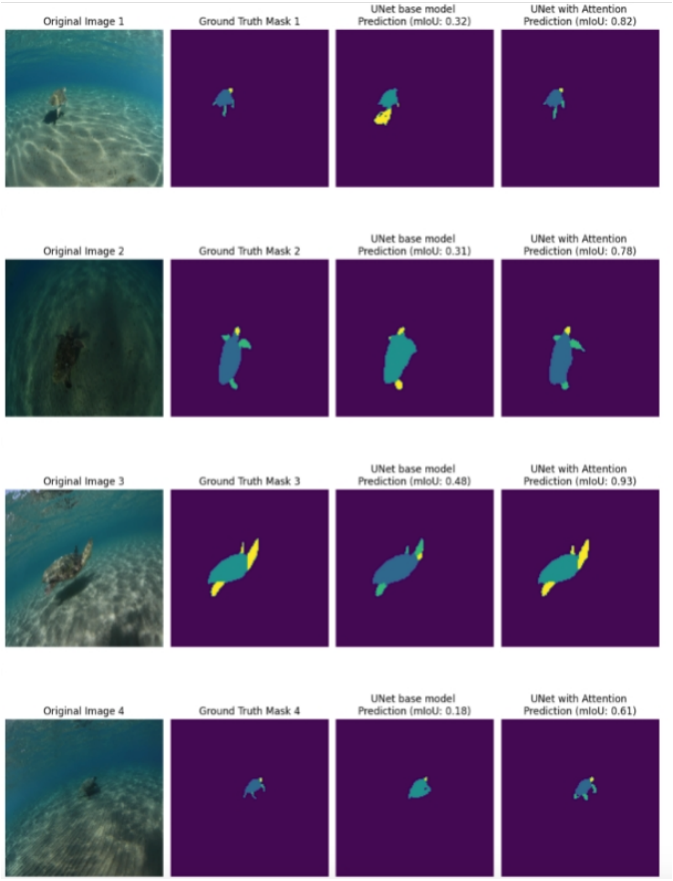

Fig. 5. U-Net Baseline and CBAM-Enhanced Model in Complex Underwater Conditions

able to more effectively isolate the turtle's body from the noisy background, maintaining structural integrity even under adverse visual conditions. This enhancement underscores CBAM's utility in improving segmentation robustness and fine detail accuracy, making the model better suited for the complexities of underwater imagery.

For sea turtle segmentation, the U-Net + CBAM enhanced the ability to capture detailed segmentations, which are beneficial for maintaining edge clarity and handling small datasets effectively. However, U-Net's limitations in generalising to complex scenes or dealing with partial occlusions present challenges. The model may struggle to maintain accuracy in instances where the turtle's body is only partially visible or obscured, affecting its consistency across varied real-world scenarios (Figure 6). This issue has been effectively addressed in the optimised Mask R-CNN.

B. Mask R-CNN Performance

Mask R-CNN demonstrated strong performance in segmenting distinct parts of the sea turtle, particularly excelling in cases that required differentiating individual objects. In examples with clear object boundaries, Mask R-CNN accurately identified areas such as the head, body, and flippers, achieving high mIoU scores. The model's instance segmentation capa-

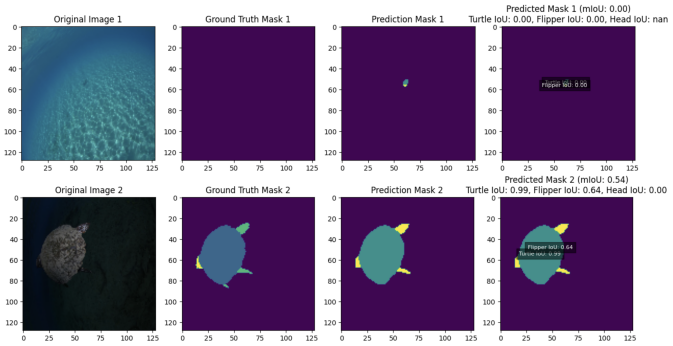

Fig. 6. Cases Opitimised U-Net Cannot Solved

bilities allowed it to capture complex features in structured and well-defined images. Also, the optimised Mask R-CNN displayed remarkable resilience in highly challenging underwater environments (e.g., dark and blurry settings), performing segmentation at a level even surpassing some ground truth annotations (Figure 7). In certain cases, the model segmented the turtle more accurately than ground truth masks, showing its capacity to capture subtle features missed in some labeling.

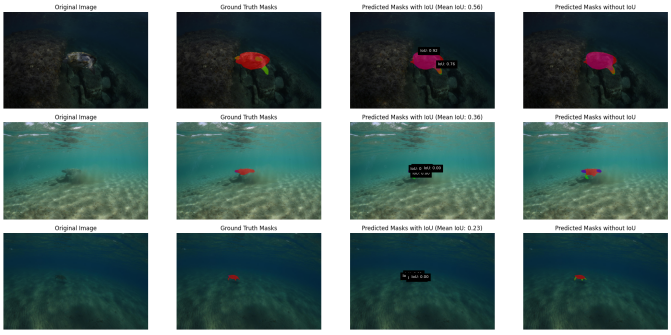

Fig. 7. Optimised Mask R-CNN Excelled in Dark and Blurry Environment

However the variability in turtle poses, lighting conditions, and underwater blurriness impacted the model's segmentation accuracy. When turtles appeared in unusual orientations or were partially occluded, Mask R-CNN occasionally failed to capture the entire object, particularly smaller parts like flippers or the head, such as those taken from unusual angles with only the head and flippers shown, as seen in 8. Additionally, due to the relatively high model complexity and computational demands, Mask R-CNN exhibited sensitivity to overfitting, which might explain the disparity in performance across different scenes. This sensitivity limits the model's ability to generalise effectively on diverse images and may lead to inconsistent results on underrepresented classes or features within the dataset.

In some instances, parts of the turtle were segmented inaccurately due to background interference, leading to false negatives or misclassified segments. One possible reason could be the model's tendency to overfit to background noise, particularly after incorporating SEBlock attention. While the attention mechanism was intended to focus on relevant regions,

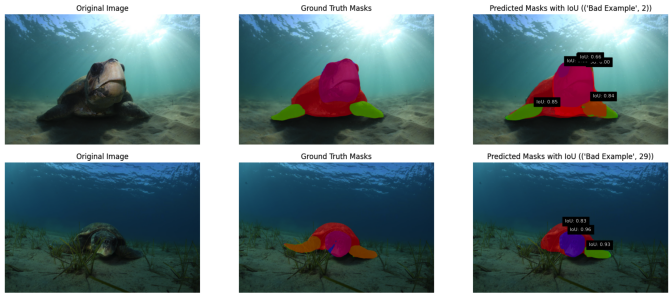

Fig. 8. Mask R-CNN Struggles with Capturing Turtle from Special Angels

it may have inadvertently highlighted irrelevant features, resulting in a decline in generalisation on variable or smaller datasets. The SEBlock attention appears to have amplified features indiscriminately, which could introduce noise from the environment, thereby affecting the model's segmentation performance in challenging scenes.

The optimised Mask R-CNN also struggled with distinguishing real turtles from complex reflections (9), such as those created by mirror-like surfaces underwater. In these instances, the model often misidentified reflections as additional turtles, segmenting both the real turtle and its reflection. This limitation highlights the difficulty of distinguishing between objects and their mirrored counterparts in underwater images.

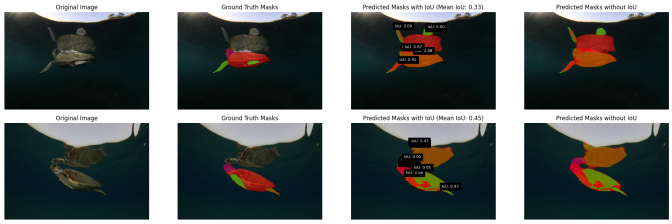

Fig. 9. Difficulty Distinguishing Real Turtle from Reflection in Water

Additionally, when multiple turtles appeared in the frame, Mask R-CNN could not differentiate which turtle should be segmented as the target and which ones could be ignored (10). This limitation emphasises the need for enhanced instance recognition or guidance in multi-object scenes, especially in applications where only specific subjects are of interest.

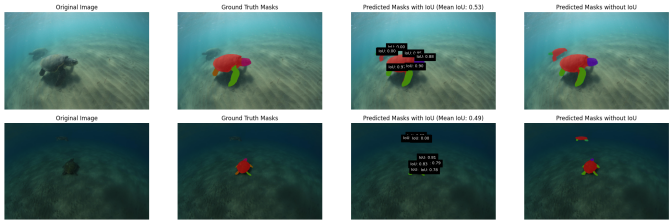

Fig. 10. Inability to Identify Target Turtle Among Multiple Instances

Mask R-CNN's main advantage lies in its strength in instance segmentation, where it effectively handles complex scenes with multiple turtles. Its disadvantages, however, in-

clude a high computational cost and a propensity to overfit on background noise, particularly when augmented with SEBlock attention. Although SEBlock was intended to refine the model's focus on relevant features, it introduced noise that complicated segmentation in more variable underwater scenes.

C. Cross-model Comparative Analysis

Mask R-CNN generally outperformed U-Net across most metrics, as reflected in the higher mIoU, particularly for the turtle's body and flippers. The stronger performance of Mask R-CNN is likely due to its two-stage architecture, especially in scenes with multiple turtles or complex backgrounds. In contrast, U-Net's encoder-decoder structure, while effective in capturing fine-grained details through skip connections, may lack the contextual understanding necessary for segmenting objects that vary in shape and orientation, as commonly seen in underwater scenes.

Within the U-Net variants, each optimisation provided unique advantages and drawbacks. The adjustment of the loss function and optimiser significantly improved U-Net's mIoU scores by focusing on hard-to-classify pixels, helping with challenging regions like flippers and heads. However, the integration of ResNet-34 as a backbone for feature extraction did not yield the expected improvement, likely due to a mismatch between pretrained features and the specific characteristics of the dataset. The CBAM attention mechanism enhanced U-Net's performance by focusing on spatially important features, particularly aiding in distinguishing small details, which improved the segmentation quality of complex turtle structures.

In the Mask R-CNN configurations, the baseline with a ResNet-50 backbone exhibited robust performance, especially in handling multi-scale features. However, adding SEBlock attention, while intended to improve focus on relevant regions, seemed to introduce noise. This outcome suggests that SEBlock might have been too aggressive, potentially causing overfitting to background textures and thus affecting the model's ability to generalise to unseen images with differing environmental conditions.

Comparing the two models, Mask R-CNN's higher computation requirements may partially explain its superior results; its ability to leverage more extensive feature hierarchies allows it to handle complex scenes better than U-Net, especially when individual objects are less prominent. However, U-Net's simplicity and efficiency make it advantageous when computational resources are limited. In underwater images where visibility and lighting vary, Mask R-CNN demonstrated greater resilience, likely due to its instance-segmentation approach. Conversely, U-Net's reliance on pixel-level information without the broader contextual insights of Mask R-CNN limited its adaptability to challenging scenes.

D. Limitations and Improvement

These models demonstrated their strengths in underwater wildlife image segmentation; however, certain limitations reveal areas for improvement. Specifically, enhancing U-Net's ability to handle complex scenes with occlusions and partial visibility, and improving Mask R-CNN's capability to differentiate real turtles from reflections and selectively target specific turtles in multi-object scenes, are crucial steps forward.

To tackle U-Net's weaknesses, especially its struggles with busy backgrounds and occluded areas, we could try targeted data augmentation for underrepresented classes and add specialised attention mechanisms. These changes would help the model focus on finer details without risking overfitting. Another approach could be experimenting with deeper backbone networks, carefully balancing the added complexity to improve the model's ability to handle complex edges and retain clarity in challenging underwater conditions.

For Mask R-CNN, improving segmentation in scenes with reflections or multiple turtles might come down to boosting its ability to recognize individual instances. Techniques like using focal loss or fine-tuning specifically on isolated turtle instances could help the model focus on critical features, like flippers and heads. Additionally, rethinking the placement of SEBlock—perhaps targeting higher feature extraction layers—could help reduce noise and improve clarity. These tweaks would allow the model to handle ambiguous underwater scenes more accurately, boosting its performance in complex environments.

VI. Conclusion

In this study, we tested and optimized U-Net and Mask R-CNN for segmenting sea turtles, focusing on the unique challenges of underwater imagery.

With focal Tversky loss and CBAM attention, U-Net demonstrated a strong ability to capture fine details and smaller structures, achieving good accuracy in segmenting features like flippers and heads. However, it struggled when parts of the body were occluded, suggesting that U-Net's encoder-decoder structure might benefit from additional contextual layers to better capture high-level features in diverse underwater scenes. Despite these limitations, U-Net's efficiency and focus on detail make it a solid choice for simpler segmentation tasks, particularly in resource-limited settings.

Mask R-CNN, with a ResNet-50 backbone, achieved the highest mIoU across all segments and excelled at instance segmentation in scenes with multiple turtles or objects of varying scales. Its two-stage architecture and Feature Pyramid Network (FPN) enabled it to handle complex underwater environments effectively. However, adding SEBlock attention didn't produce the expected improvement; instead, it introduced more background noise, likely by emphasizing irrelevant features. This shows the importance of carefully positioning and tuning attention mechanisms for complex segmentation tasks.

Overall, Mask R-CNN's strong instance segmentation abilities make it ideal for applications that demand high accuracy in diverse, occluded environments, such as large-scale turtle tracking and behavioral studies. Meanwhile, U-Net's straightforward structure and precision with fine details make it better suited for tasks where efficiency is a priority over extensive contextual analysis. This comparison highlights the strengths and limitations of each model, providing insights that can

guide the development of more scalable and accurate tools for marine conservation image analysis.

Building on these findings, a promising direction for future work could be developing a hybrid model that combines U-Net's strength in capturing fine details with Mask R-CNN's powerful instance segmentation capabilities. This could lead to better segmentation performance in crowded underwater environments. According to our experiments on attention mechanisms, we found there are still rooms for attention mechanism explorations. Advanced attention mechanisms and adaptive placement strategies could be explored to refine the model's focus, ensuring that attention layers contribute effectively without introducing background noise.

Addressing data imbalance through targeted data augmentation and synthetic data generation could improve model sensitivity to underrepresented features, such as flippers or partially occluded parts. Techniques like focal Tversky loss and selective fine-tuning on specific parts could further enhance segmentation accuracy for small or challenging structures.

Finally, improving multi-scale feature fusion in both models could aid in detecting small and complex features, particularly in scenarios with varied turtle sizes and orientations. These enhancements would contribute to more robust and generalisable models for underwater segmentation tasks, supporting large-scale biodiversity monitoring and conservation efforts.

REFERENCES

[1] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848.

[2] L. Adam, V. Čermák, K. Papafitsoros, L. Picek. SeaTurtleID2022: A long-span dataset for reliable sea turtle re-identification. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7146-7156.

[3] W. Zhou, X. Du and S. Wang, Techniques for Image Segmentation Based on Edge Detection, 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 2021, pp. 400-403, doi: 10.1109/CEI52496.2021.9574569.

[4] Erwin, Erwin & Masyarif, Saparudin & Nevriyanto, Adam & Purnamasari, Diah. (2018). Performance Analysis of Comparison between Region Growing, Adaptive Threshold and Watershed Methods for Image Segmentation.

[5] Rao, K. S., Sekhar, P. C., & Rao, P. S. (2014). Image Segmentation for Animal Images using Finite Mixture of Pearson Type VI Distribution. Global Journals.

[6] Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. Morgan & Claypool Publishers.

[7] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3431-3440.

[8] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848.

[9] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234-241.

[10] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. ArXiv Preprint ArXiv:1807.10165.

[11] Oktay, Ozan & Schlemper, Jo & Folgoc, Loic & Lee, Matthew & Heinrich, Mattias & Misawa, Kazunari & Mori, Kensaku & McDonagh, Steven & Hammerla, Nils & Kainz, Bernhard & Glocker, Ben & Rueckert, Daniel. (2018). Attention U-Net: Learning Where to Look for the Pancreas. 10.48550/arXiv.1804.03999.

[12] Liu, Ze & Lin, Yutong & Cao, Yue & Hu, Han & Wei, Yixuan & Zhang, Zheng & Lin, Stephen & Guo, Baining. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 10.48550/arXiv.2103.14030.

[13] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., Dollár, P., & Girshick, R. (2023). Segment Anything. arXiv. https://arxiv.org/abs/2304.02643

[14] Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Qiao, Y., Gao, P., & Li, H. (2023). Personalize Segment Anything Model with One Shot. ArXiv Preprint ArXiv:2305.03048.

[15] Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., & Huang, T. (2023). SegGPT: Segmenting Everything in Context. ArXiv Preprint ArXiv:2303.03284.

[16] Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., & Shen, C. (2023). Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching. ArXiv Preprint ArXiv:2305.13310.

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. arXiv. https://arxiv.org/abs/1706.03762

[19] He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask r-CNN. In Proceedings of the IEEE international conference on computer vision (pp. 2961–2969).

[20] Sapkota, R., Ahmed, D., & Karkee, M. (2024). Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments. Artificial Intelligence in Agriculture, 13, 84–99. https://doi.org/10.1016/j.aiia.2024.07.001