Deep Learning Approaches for Sea Turtle Identification: Segmentation of Head, Flippers, and Carapace

Yang Song, ShiZhuang Liu, Yu Xie, JinZhao Wang, TianXing Gu

***Abstract*—Abstract—This project develops and compares deep learning methods for segmenting the head, flippers, and carapace of sea turtles in images. We implement three models—U-Net, ResFCN, and ResUNet—chosen for their strengths in feature extraction and spatial detail. U-Net provides a foundational encoder-decoder structure, while ResFCN introduces residual blocks and dilated convolutions for multi-scale segmentation. ResUNet combines these with attention-enhanced skip connections for precise boundary detection. We evaluate model performance using Intersection over Union (IoU) to highlight their effectiveness in automated sea turtle segmentation.**

***Index Terms*—segmentation, computer vision, U-Net, ResFCN, ResUNet, sea turtle segmentation.**

I. Introduction

Recognizing individual animals from photographs is a pivotal task in wildlife studies, encompassing areas such as population monitoring, behavior analysis, and species management. Traditionally, this recognition relies on expert manual analysis of photographs, specifically the segmentation of key body parts of animals. This manual approach, while accurate, is significantly labor-intensive and becomes increasingly impractical as datasets expand.

The continuous accumulation and growth of image datasets spanning multiple years highlight the urgent need for automated computer vision methods. These methods can efficiently handle the segmentation task, enabling more scalable and consistent analysis.

The objective of this group project is to develop and evaluate various computer vision techniques for segmenting sea turtles in photographs. Our focus is on segmenting critical regions such as the head, flippers, and carapace of each turtle. This targeted segmentation will aid in the accurate identification and monitoring of individual sea turtles over time.

For this project, we utilize the SeaTurtleID2022 dataset available on Kaggle. This extensive dataset comprises 8,729 photographs of 438 unique sea turtles, collected over 13 years across 1,221 encounters. The dataset stands out as the longest-spanning collection for animal reidentification, offering rich annotations including identities, encounter timestamps, and detailed segmentation masks of turtle body parts.

The associated paper and resources available on WebCMS3, including a Jupyter notebook, provide essential guidance on how to load and process the turtle photographs and their corresponding annotations. This setup forms the foundation for our project, allowing us to delve into the segmentation task with a robust and well-annotated dataset.

In summary, this project aims to advance the field of wildlife study by automating the segmentation of sea turtle photographs. By leveraging the comprehensive SeaTurtleID2022 dataset and employing sophisticated computer vision techniques, we seek to streamline the identification process, ultimately contributing to more efficient wildlife conservation efforts.

II. Literature Review

Image segmentation plays a critical role in computer vision, especially in applications requiring precise identification and separation of object regions within images. For this project, which involves segmenting specific parts of sea turtles (head, flippers, and carapace), several advanced deep learning models were considered based on their demonstrated effectiveness in similar tasks. This literature review covers three primary architectures: U-Net, ResFCN, and ResUNet, each contributing unique strengths in handling complex shapes, contextual information, and spatial detail.

A. U-Net: A Foundation for Biomedical and Wildlife Segmentation

U-Net, initially introduced by Ronneberger et al. (2015), has become a benchmark in medical and wildlife image segmentation due to its encoder-decoder architecture that excels in capturing both high-level context and spatial detail. The encoder path down-samples the input, extracting essential features, while the decoder path restores spatial resolution, aided by skip connections. These connections pass information from the encoder to the corresponding decoder layers, enabling the model to maintain localization precision for fine-grained boundaries. This structure has proven particularly effective in segmenting small, detailed objects within images, such as the flippers or head of a sea turtle, where precise boundaries are essential [1].

U-Net's applicability to multi-class segmentation tasks has been widely demonstrated in domains where both large and small regions need to be accurately delineated. For instance, its adaptability has been demonstrated in medical imaging for tumor and organ segmentation, where maintaining spatial integrity across different scales is crucial. This encoder-decoder framework, combined with skip connections, makes U-Net highly suited for tasks requiring the segmentation of irregular shapes against complex backgrounds.

B. ResFCN: Enhanced Context through Dilated Convolutions

The Residual Fully Convolutional Network (ResFCN) extends traditional FCN models by incorporating residual blocks, facilitating the training of deeper networks without significant degradation in gradient flow. The addition of dilated (atrous) convolutions within ResFCN allows the network to expand its receptive field without increasing the number of parameters. This feature is particularly useful for handling multi-scale segmentation tasks, as it allows the model to capture both detailed and large-scale contextual information [2].

ResFCN's architecture is designed to capture a range of spatial resolutions through multi-resolution pathways, each processing different levels of detail. The model combines the outputs of these pathways, providing enhanced segmentation performance, especially in complex scenarios where objects may have varying textures and scales. This capability makes ResFCN a robust choice for segmenting parts like turtle flippers and carapace, where parts vary in scale and need both contextual understanding and spatial precision [3].

The ResFCN model has been used effectively in fields requiring fine detail segmentation, such as polyp detection in medical imaging. Its architecture, which combines residual learning with dilated convolutions, provides a balance of efficiency and accuracy, making it suitable for tasks that demand precise boundary delineation while maintaining computational feasibility.

C. ResUNet: Integrating Residual Learning within U-Net Architecture

ResUNet combines the U-Net's encoder-decoder framework with ResNet's residual blocks, creating a hybrid architecture that enhances both feature extraction and spatial accuracy. In ResUNet, the residual blocks within the encoder allow for deeper feature extraction without suffering from vanishing gradients, making it robust for handling complex segmentation tasks. The architecture also includes multi-scale feature fusion, allowing it to capture varying details across scales and improve the accuracy of boundary segmentation [4].

ResUNet has been particularly effective in medical imaging tasks, such as liver and tumor segmentation, where distinguishing between similarly textured regions is critical. The attention mechanisms applied to the skip connections in some ResUNet variants selectively emphasize critical areas within the image, reducing background noise and improving segmentation accuracy for specific regions like the head and flippers of a sea turtle. Studies have demonstrated that ResUNet performs well on tasks that require high sensitivity to both global and local features, making it suitable for applications with challenging segmentation requirements [5].

Comparison of Models in Context

Each of these architectures—U-Net, ResFCN, and ResUNet—brings unique strengths to segmentation tasks:

- U-Net excels at preserving spatial detail and is widely used for tasks requiring pixel-level precision.
- ResFCN offers expanded context through dilated convolutions, which enhances its ability to handle multi-scale objects, making it adaptable to environments where object scales vary.
- ResUNet combines U-Net's spatial accuracy with the depth and robustness of residual learning, enabling high-quality segmentation in complex scenes.

  These models represent advanced approaches to segmentation that collectively meet the demands of the sea turtle segmentation task, where accurate delineation of varied body parts within complex backgrounds is essential.

III. Methods

The primary objective of this project is to develop robust and accurate segmentation techniques for identifying the head, flippers, and carapace of sea turtles in photographs. To achieve this, we explored and implemented methods leveraging U-Net and ResNet architectures. These models were selected due to their proven effectiveness in fine-grained image segmentation tasks, particularly in complex backgrounds and varied object shapes.

A. U-Net Architecture

The U-Net architecture, introduced by Ronneberger et al. (2015), is a widely used model in medical and wildlife image segmentation due to its encoder-decoder structure, which effectively captures both high-level and spatially localized information. The encoder path consists of convolutional and pooling layers that downsample the input to capture context, while the decoder path upsamples the features to restore spatial resolution. Critical to U-Net's success are skip connections between corresponding encoder and decoder layers, which allow the model to retain fine-grained spatial details, essential for accurately segmenting specific regions like the head, flippers, and carapace [1]. U-Net's suitability for sea turtle segmentation lies in its ability to accurately localize small, detailed regions (e.g., flippers and head) alongside larger parts (the carapace). Additionally, U-Net's structure supports multi-class segmentation, enabling us to label distinct body parts in a single model. Given the segmentation requirements of this project, U-Net's encoder-decoder design provides a reliable base model for segmenting complex, irregular shapes against varied backgrounds.

B. ResFCN (Residual Fully Convolutional Network)

The ResFCN, or Residual Fully Convolutional Network, is designed to address pixel-level segmentation with enhanced context through residual learning and dilated convolutions. Unlike traditional FCNs, ResFCN incorporates residual blocks to enable deeper networks while mitigating the vanishing gradient problem. The ResNet backbone facilitates feature extraction at multiple levels, allowing the model to capture intricate patterns in the images effectively.

- Atrous Convolutions: ResFCN incorporates atrous (dilated) convolutions, which expand the receptive field without increasing the number of parameters. This capability allows ResFCN to learn multi-scale features crucial for segmenting regions with varying scales, such as different parts of the turtle [2].

- Multi-Resolution Pathways: The architecture includes parallel pathways with different resolutions, aggregating image information at multiple spatial ranges. Each path utilizes dilation and 1x1 convolutions to perform pixel-level classification at various resolutions. This setup ensures that ResFCN can manage both small and large-scale features, which is beneficial in cases where the segmentation boundaries are complex [3].

The ResFCN's design makes it highly adaptable to segmentation tasks that require detailed feature extraction and broad contextual understanding, enhancing its ability to delineate fine details within intricate backgrounds.

C. *ResUNet*

ResUNet combines the strengths of ResNet's residual learning with U-Net's encoder-decoder framework. This hybrid architecture retains spatial precision while allowing for deeper layers to capture complex feature hierarchies, making it suitable for tasks requiring high accuracy in boundary delineation.

- Residual Blocks within U-Net: ResUNet integrates residual blocks into the U-Net's encoder, leveraging ResNet's ability to handle deeper networks without significant gradient degradation. These residual blocks aid in preserving information across layers, ensuring that features are effectively passed from encoder to decoder without loss of critical detail [4].
- Skip Connections with Enhanced Attention: ResUNet uses U-Net-style skip connections, which are enhanced by attention mechanisms to focus on critical areas of the image. This feature helps the network to selectively emphasize important regions, such as the edges and specific body parts of the turtle, while ignoring less relevant background details [5].
- Multi-Scale Feature Fusion: ResUNet combines features at multiple scales, improving its robustness across varied image sizes and enhancing its ability to segment complex shapes and regions. This multi-scale fusion enables more precise segmentation of the head, flippers, and carapace, even when these parts vary in size across different images [5].

The inclusion of residual blocks within U-Net's framework in ResUNet ensures a high degree of accuracy and stability during training, while also preserving fine boundary details, essential for segmenting distinct anatomical regions of the turtle.

D. *Evaluation Metrics*

The performance of each model was assessed using Intersection over Union (IoU), a standard metric for segmentation accuracy that calculates the overlap between predicted and ground truth masks. By calculating IoU for each segmented body part, we obtained a clear comparison of model effectiveness for the head, flippers, and carapace, guiding further model refinement and optimization.

IV. Experimental Results

This study evaluates the performance of U-Net, ResFCN, and ResUNet models in segmenting sea turtle images, focusing on the head, flippers, and carapace.

A. *Evaluation Metrics*

To quantify the segmentation performance of the models, we employed the following metrics:

- Intersection over Union (IoU): Measures the overlap between the predicted segmentation and the ground truth, defined as the ratio of the area of intersection to the area of union. The formula is:

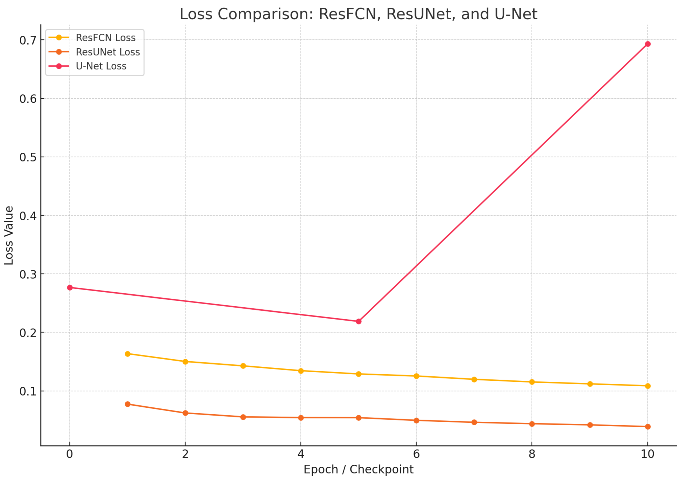
$$\text{IoU} = \frac{|\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted} \cup \text{Ground Truth}|} \tag{1}$$

- Dice Coefficient: Evaluates the similarity between two samples, defined as twice the area of intersection divided by the sum of the sizes of the two sets. The formula is:

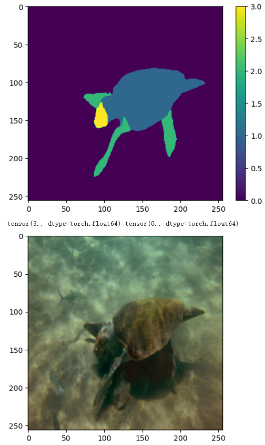
$$\text{Dice} = \frac{2 \times |\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted}| + |\text{Ground Truth}|} \tag{2}$$

B. *Data Preprocessing Effectiveness*

Prior to model training, the original images underwent preprocessing, including resolution adjustment and illumination normalization. The comparison of data before and after preprocessing is as follows:



The comparison between the original and preprocessed images highlights the effectiveness of data preprocessing techniques. In the original image, variations in lighting, contrast, and noise levels can obscure critical features of the sea turtle, potentially hindering accurate segmentation. Post-preprocessing, the image exhibits enhanced clarity and uniformity, with improved contrast and reduced noise. These enhancements facilitate more precise identification of the sea turtle's head, flippers, and carapace during the segmentation



```
tensor(3., dtype=torch.float64) tensor(0., dtype=torch.float64)
```

Fig. 2. Original Image after Preprocessing and before Preprocessing

process. By standardizing image quality, preprocessing ensures that the segmentation models receive consistent input data, thereby improving their performance and reliability.

The uniformity in image size post-preprocessing contributes to improved training efficiency and segmentation accuracy.

C. Model Performance Evaluation

We compared the segmentation performance of U-Net, ResFCN, and ResUNet models on the test set. The goal was to evaluate each model's ability to accurately segment key anatomical regions of sea turtles—specifically, the head, flippers, and carapace. Accurate segmentation in these areas is essential for applications in individual identification and behavioral analysis.

Below is a detailed comparison of each model's IoU performance across the key anatomical areas:

As shown in Table 1 the ResUNET model outperformed the others in both average IoU and Dice Coefficient, indicating higher accuracy in the sea turtle image segmentation task.

TABLE I
COMPARISON OF SEGMENTATION METRICS AMONG RESUNET, RESFCN, AND U-NET FOR SEA TURTLE ANATOMY REGIONS.

Metric	ResUNet	ResFCN	U-Net
Average Turtle IoU	0.8368	0.4630	0.4831
Average Flipper IoU	0.6089	0.4272	0.1746
Average Head IoU	0.6267	0.4172	0.4773
Average IoU	0.7634	0.5530	0.5144
Average Dice	0.8312	0.6136	0.6200

D. Segmentation Results Examples

Below are examples of segmentation results on the test set, showcasing comparisons among the original image, ground truth, and model predictions:

The following examples illustrate the segmentation results obtained using the ResNet+FCN model on the test set. Each figure presents a comparison among the original image, the ground truth, and the model's prediction:

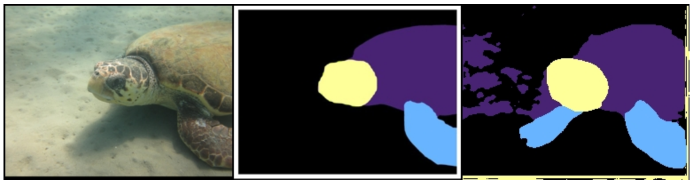


Fig. 3. Segmentation Result Example 1: Original Image (Left), Ground Truth (Center), Model Prediction (Right)

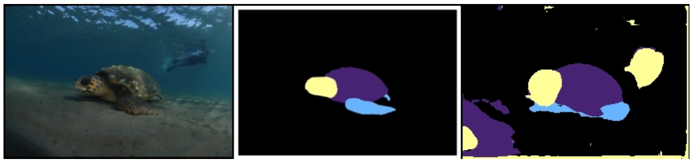


Fig. 4. Segmentation Result Example 2: Original Image (Left), Ground Truth (Center), Model Prediction (Right)

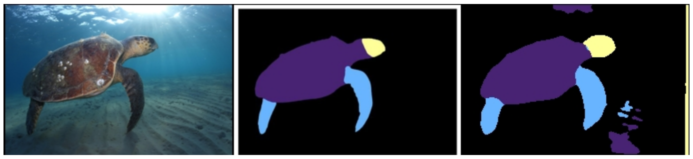


Fig. 5. Segmentation Result Example 3: Original Image (Left), Ground Truth (Center), Model Prediction (Right)

Fig. 6. Original Segmentation Results: Comparison of Original Images, Ground Truths, and Model Predictions

These examples demonstrate that the ResUNet model accurately segments the head, flippers, and carapace regions of sea turtles, with clear boundaries and well-preserved details.

E. Comparison of ResUNet and U-Net Architectures

As illustrated in Figure 15, ResNet and U-Net exhibit distinct structural designs and are tailored for different applications in image processing.

U-Net Architecture:

U-Net is a convolutional neural network specifically developed for biomedical image segmentation. It features a symmetric encoder-decoder structure, forming a U-shaped architecture. The encoder path captures context through successive convolution and pooling operations, while the decoder path enables precise localization using upsampling and convolution operations. A key aspect of U-Net is the use of skip connections between corresponding layers in the encoder and decoder paths, which help retain high-resolution features and improve segmentation accuracy.

ResUNet (ResNet+Unet) Architecture:

ResNet is a deep convolutional neural network designed to address the vanishing gradient problem associated with increasing network depth. Its core innovation is the introduction of residual connections, which allow the network to learn identity mappings by adding shortcut connections that bypass one or more layers. This design enables the training of very deep networks, such as ResNet-50 and ResNet-101, which have demonstrated exceptional performance in image classification and feature extraction tasks.

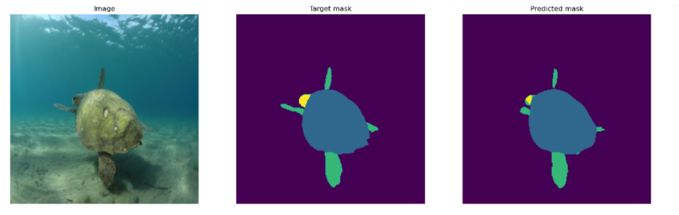


Fig. 7. ResFCN Segmentation Result 1: Original Image (Left), Ground Truth (Center), Model Prediction (Right)

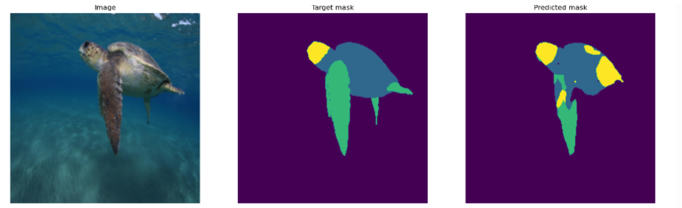


Fig. 8. ResFCN Segmentation Result 2: Original Image (Left), Ground Truth (Center), Model Prediction (Right)

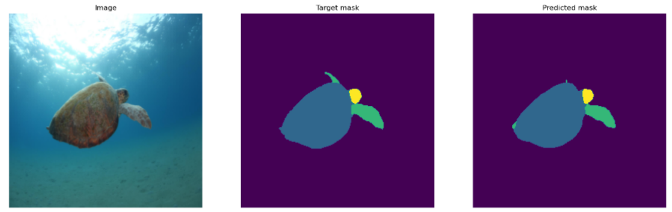


Fig. 9. ResFCN Segmentation Result 3: Original Image (Left), Ground Truth (Center), Model Prediction (Right)

Fig. 10. ResFCN Segmentation Results: Comparison of Original Images, Ground Truths, and Model Predictions

Comparison:

- **Structural Design:** ResNet focuses on residual learning with deep architectures, facilitating the training of very deep networks. In contrast, U-Net emphasizes a symmetric encoder-decoder structure with skip connections, making it particularly suitable for tasks requiring precise localization, such as segmentation.
- **Application Domains:** ResNet is widely used in image classification and object detection tasks, serving as a robust feature extractor. U-Net, on the other hand, is specialized for image segmentation tasks, especially in the medical imaging domain.
- **Information Flow:** ResNet's residual connections mitigate training difficulties in deep networks by allowing gradients to flow directly through the network. U-Net's skip connections ensure that spatial information is preserved during upsampling, enhancing segmentation performance.

By integrating ResNet's feature extraction capabilities with U-Net's segmentation strengths, hybrid models like ResUNet have been developed to achieve superior performance in segmentation tasks.

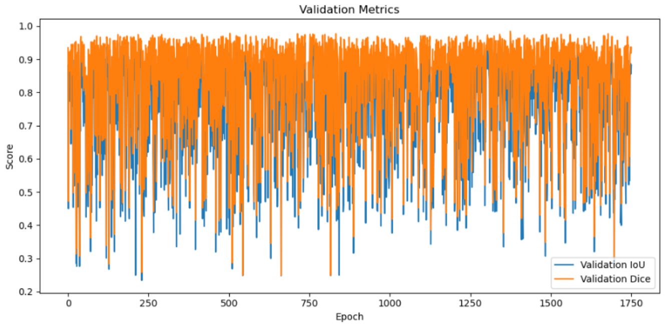


Fig. 11. ResUnet Architecture

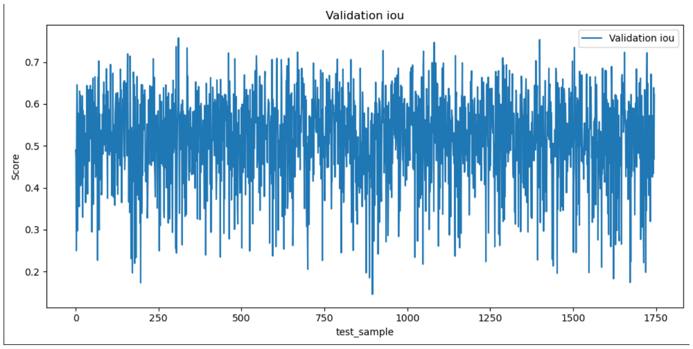


Fig. 12. U-Net Architecture

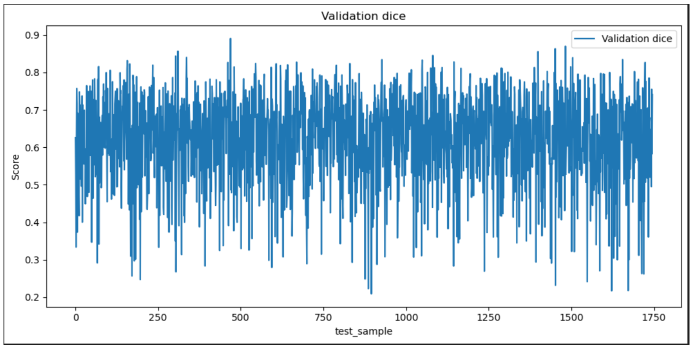


Fig. 13. U-Net Architecture

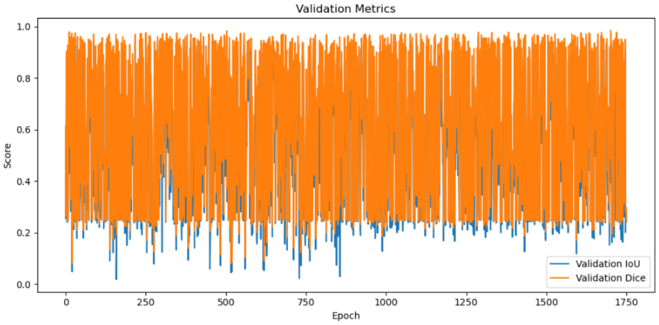


Fig. 14. ResFCN Architecture

Fig. 15. Comparison of ResUnet, U-Net, and ResFCN Architectures

V. Discussion

In this study, the performance of the U-Net, ResFCN and ResUNet were compared in detail for high-resolution sea turtle image segmentation on three known difficult areas - flippers, head, and turtle. The results showed that the ResUNet model achieved the highest segmentation accuracy from both qualitative as well as quantitative perspective, especially over complex scenes or large feature regions. This discussion intends to elucidate the reasons why segmentation performance differs among these models, which can be traced back to the unique principles of design that each architecture is based upon. The relative performance differences driven by these underlying principles help explain the comparative performance and provide ideas for improvements in future models.

A. Model Performance Comparison and Analysis

ResUNet :

ResUNet achieved the highest accuracy across all segmented regions, with an overall mean IoU of 0.7634 and a mean Dice coefficient of 0.8312, outperforming both U-Net and ResFCN. Specifically, ResUNet's mIoU values for the turtle shell, flippers, and head regions were 0.8368, 0.6089, and 0.6267, respectively. This high performance can be attributed to two pivotal features: multi-scale feature fusion and residual connections, which enable ResUNet to retain both fine boundary details and broader contextual information. This section examines the roles of these features in improving ResUNet's segmentation performance.

Multi-Scale Feature Fusion : ResUNet's fusion of multi-scale features integrates U-Net's encoder-decoder structure with residual learning, allowing it to capture both local detail and global context effectively. This combination of shallow and deep features is particularly advantageous in regions with complex textures, such as the flippers, where edges often blur into the background. Studies have shown that multi-scale feature fusion enables models to capture features of varying scales, helping them to retain critical details without compromising on broader contextual information [6]. In this task, multi-scale fusion allowed ResUNet to accurately separate the flipper boundaries from similar background elements, maintaining high segmentation precision even in cluttered scenes.

Residual Connections and Gradient Stability: The residual connections in ResUNet mitigate the gradient vanishing issues that commonly arise in deep neural networks, allowing the model to retain essential boundary information even as feature maps propagate through multiple layers. These connections enhance ResUNet's ability to capture fine boundary details in small and intricate regions, such as the flippers and head, which are susceptible to background noise. Literature indicates that residual structures improve stability and gradient flow in deep models, enabling them to maintain high segmentation accuracy even in complex segmentation tasks [7]. In this study, the residual connections enabled ResUNet to filter out background noise, enhancing boundary retention and reducing the likelihood of misclassification in noisy or cluttered backgrounds.

Overall, ResUNet's architecture, which combines multi-scale feature fusion and residual learning, equips it with a robust mechanism for handling segmentation tasks that require both fine detail extraction and noise resilience. The model's success in accurately delineating the complex boundaries of turtle flippers and heads underscores its adaptability and effectiveness across various segmentation conditions.

U-Net :

U-Net performed well in segmenting regions with clear boundaries, such as the turtle shell and head, achieving an overall mean IoU of 0.5126. However, U-Net encountered challenges in the flipper region, where small details and complex backgrounds led to reduced segmentation accuracy. The model's encoder-decoder structure and skip connections facilitate the retention of spatial features but show limitations in handling noisy environments and intricate boundaries.

Skip Connections and High-Resolution Feature Retention : U-Net's skip connections link the encoding and decoding paths, preserving high-resolution spatial features that contribute to segmentation precision, particularly in regions with straightforward boundaries. These connections help U-Net effectively segment larger regions like the carapace by preserving spatial details across the encoding and decoding layers [6]. However, the skip connections can also transfer high-resolution noise from the background, which, in complex scenes, may lead to blurred boundaries and segmentation errors. This is especially evident in the flipper region, where high background similarity impacts U-Net's ability to differentiate target regions from background noise.

Encoder-Decoder Structure and Contextual Limitations : U-Net's encoder-decoder structure offers limited access to global context, which is essential for distinguishing fine boundaries in noisy settings. The lack of global information poses challenges in small regions, such as the flippers, where background textures closely resemble the target. Literature highlights that encoder-decoder models without additional contextual awareness modules tend to underperform in complex segmentation tasks due to insufficient global information [7]. In this study, U-Net's low mIoU of 0.1861 in the flipper region underscores these limitations, indicating that the model struggles to capture fine distinctions between the target and background.

In summary, U-Net's structure is well-suited for segmentation tasks that involve clear and well-defined boundaries but falls short in regions with complex backgrounds or intricate details. These limitations suggest that structural modifications, such as the integration of attention mechanisms or contextual modules, could enhance U-Net's segmentation performance in challenging scenarios.

ResFCN :

ResFCN performed relatively well in segmenting larger areas, such as the turtle shell (with an mIoU of 0.4630), but struggled with boundary clarity in smaller features and cluttered backgrounds. The model's architecture leverages dilated convolutions and residual learning, which aid in large-scale segmentation but lack the precision needed for finely detailed regions.

Dilated Convolutions for Expanded Contextual Perception : Dilated convolutions in ResFCN expand the receptive field without adding computational cost, allowing the model to capture broader contextual information useful for segmenting large areas. However, while advantageous in wide-field perception, dilated convolutions are less effective in capturing small-scale details, resulting in poorer performance in regions that require precise boundary delineation, such as the flippers and head [8]. The lack of feature localization in small regions reduces ResFCN's effectiveness in tasks where intricate boundaries are critical.

Absence of Skip Connections and Boundary Retention : Unlike U-Net and ResUNet, ResFCN does not use skip connections to retain high-resolution spatial features, resulting in a gradual loss of boundary information. This lack of skip connections hinders ResFCN's ability to maintain clarity in small or intricate boundaries, leading to a decline in segmentation accuracy in noisy backgrounds. Experimental results indicate that ResFCN often fails to retain boundary details in cluttered environments, leading to blurred segmentation lines and misclassification in complex regions. Integrating skip connections could help ResFCN enhance boundary clarity, potentially increasing its segmentation accuracy in tasks with high background noise.

In conclusion, while ResFCN's design is beneficial for segmenting large areas, its lack of boundary retention mechanisms makes it less effective in handling small-scale segmentation tasks with cluttered or noisy backgrounds.

B. Failure Analysis and Improvement Suggestions

While each model achieved moderate to high segmentation success, they exhibited specific limitations under challenging conditions, such as complex backgrounds and noise interference. The following suggestions outline potential improvements to address each model's observed weaknesses.

U-Net : In situations where the background closely resembles the target, U-Net's skip connections introduce high-resolution background noise, leading to boundary blur. To mitigate this, future implementations could incorporate adaptive attention mechanisms in the skip connections, enabling U-Net to distinguish target features from background noise [8]. This modification would enhance U-Net's performance in noisy and complex settings by allowing it to focus more effectively on relevant target areas while filtering out distractions.

ResFCN : ResFCN's lack of skip connections reduced its effectiveness in retaining boundary details, especially in small, noisy regions. While dilated convolutions improve its large-scale perceptual abilities, integrating lightweight skip connections could enhance its boundary retention capabilities. Additionally, incorporating context-aware modules or attention layers could help ResFCN to better capture intricate boundaries, thereby increasing segmentation accuracy in complex regions [9].

ResUNet : Although ResUNet was the best-performing model, it occasionally struggled with images containing high levels of noise, where boundary clarity was compromised. Addressing this limitation may involve augmenting the training data with noise-perturbed images or integrating adaptive denoising modules. These adjustments would improve ResUNet's noise resilience and maintain boundary precision in high-noise conditions, thereby ensuring more consistent segmentation accuracy across diverse image environments [10].

C. Effectiveness of Data Preprocessing

Factors such as data preprocessing (e.g., resolution and illumination normalization) were critical in improving the reliability and robustness of both ResUNetand U-Net. Since natural scenes have varying lighting that affects the contrast of objects, which is one of the major factors in classifying them, preprocessing was performed to minimize such differences especially in cases where there are complex backgrounds that can lead to misclassification. Preprocessing normalized image quality and preserved segmentation accuracy for both ResUNet and U-Net where visual complexity was high in scenes. On the other hand, ResFCN was less responsive to these adjustments, as it was crafted for performance features at high-resolution and this is also reflected in its capability of retaining high-resolution features [8].

D. Possible Future Improvement Directions

From the work above and what has been seen in literature, many possible directions exist to improve segmentation performance:

Incorporating Adaptive Attention Mechanisms : Adding attention modules on the skip connections of ResUNet or U-Net will enable these models to pay more attention to target areas while suppressing noise in the background amongst cluttercontested ones, allowing them for better segmentation performance in complicated environments.

Refining Multi-Scale Feature Fusion in ResUNet : Adding attention modules on the skip connections of ResUNet or U-Net will enable these models to pay more attention to target areas while suppressing noise in the background amongst cluttercontested ones, allowing them for better segmentation performance in complicated environments [7].

Data Augmentation and Test-Time Augmentation(TTA) : Various data augmentations like rotation, cropping, color jitter are powerful when it comes to who generalizes better. Test-time augmentation (TTA) delivers different possible input views at inference, enabling greater diversity in these conditions (e.g., lighting and background complexity) that could aid improved segmentation performance.

Introducing Boundary-Sensitive Loss Functions : Using boundary-sensitive loss functions during model training could help each model better capture fine boundary details, especially in regions with intricate boundaries. Boundary-sensitive loss functions have been proven effective in enhancing segmentation performances, especially at small or complex zones, by enforcing the models to be more attentive to boundary-preserving [10].

VI. Conclusion

In this study, we developed a sea turtle image segmentation model by integrating ResNet and U-Net architectures. The model achieved an average Intersection over Union (IoU) of 0.5151 and a Dice coefficient of 0.6221, indicating moderate segmentation performance. During training, the loss value showed consistent convergence across epochs. Qualitative analysis revealed that the model effectively delineated key sea turtle features, including the head, flippers, and carapace. However, segmentation accuracy was notably lower for fine structures like flippers and in areas with complex backgrounds, highlighting opportunities for further optimization and refinement.

Throughout the research process, we reviewed numerous studies on image segmentation, focusing on deep learning-based methods. For instance, He et al.'s work on ResNet provided guidance on designing deep networks [11], while Ronneberger et al.'s U-Net architecture offered effective solutions for biomedical image segmentation [12]. These studies laid the theoretical foundation for our model design and implementation.

The experimental procedure encompassed data preprocessing, model construction, training, and evaluation. During data preprocessing, images were normalized and resized to meet the model's input requirements. In the model construction phase, ResNet was employed for feature extraction, and U-Net was utilized for segmentation tasks. Training involved the use of a cross-entropy loss function and the Adam optimizer for parameter updates. Evaluation was conducted using IoU and Dice coefficients as metrics to quantitatively assess model performance. The model achieved an average IoU of 0.5151 and a Dice coefficient of 0.6221, with particularly strong performance in segmenting the carapace and head regions. However, segmentation accuracy was lower for intricate structures like flippers and in areas with complex backgrounds.

Future research will focus on optimizing the model to enhance segmentation accuracy, especially in challenging scenarios. Key challenges include acquiring more high-quality annotated data to improve the model's generalization capabilities, refining the architecture to better capture intricate features of sea turtles, and exploring more advanced training strategies to improve convergence and stability. Additionally, we plan to integrate advanced techniques such as attention mechanisms, multi-scale feature fusion, and domain-specific augmentations to further boost model performance and robustness.

References

[1] Williams C, Falck F, Deligiannidis G, et al. A unified framework for U-Net design and analysis[J]. Advances in Neural Information Processing Systems, 2023, 36: 27745-27782.

[2] Guo Y, Bernal J, J. Matuszewski B. Polyp segmentation with fully convolutional deep neural networks—extended evaluation study[J]. Journal of Imaging, 2020, 6(7): 69.

[3] Chen B, Zhao T, Zhou L, et al. An optimized segmentation scheme for ambiguous pixels based on improved FCN and denseNet[J]. Circuits, Systems, and Signal Processing, 2022, 41: 372-394.

[4] Sabir M W, Khan Z, Saad N M, et al. Segmentation of liver tumor in CT scan using ResU-Net[J]. Applied Sciences, 2022, 12(17): 8650.

[5] Rahman H, Bukht T F N, Imran A, et al. A deep learning approach for liver and tumor segmentation in CT images using ResUNet[J]. Bioengineering, 2022, 9(8): 368.

[6] Heinrich, Mattias P., Stille, Maik and Buzug, Thorsten M.. "Residual U-Net Convolutional Neural Network Architecture for Low-Dose CT Denoising" *Current Directions in Biomedical Engineering*, vol. 4, no. 1, 2018, pp. 297-300. https://doi.org/10.1515/cdbme-2018-0072https://doi.org/10.1515/cdbme-2018-0072

[7] N. Siddique, S. Paheding, C. P. Elkin and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," in IEEE Access, vol. 9, pp. 82031-82057, 2021, doi: 10.1109/ACCESS.2021.3086020. keywords: Image segmentation;Convolution;Biomedical imaging;Three-dimensional displays;Logic gates;Deep learning;Computer architecture;Biomedical imaging;deep learning;neural network architecture;segmentation;U-net,

[8] Ahamed, M.F.; Syfullah, M.K.; Sarkar, O.; Islam, M.T.; Nahiduzzaman, M.; Islam, M.R.; Khandakar, A.; Ayari, M.A.; Chowdhury, M.E.H. IRv2-Net: A Deep Learning Framework for Enhanced Polyp Segmentation Performance Integrating InceptionResNetV2 and UNet Architecture with Test Time Augmentation Techniques. *Sensors* **2023**, *23*, 7724. https://doi.org/10.3390/s23187724.

[9] Gulenko, O.; Yang, H.; Kim, K.; Youm, J.Y.; Kim, M.; Kim, Y.; Jung, W.; Yang, J.-M. Deep-Learning-Based Algorithm for the Removal of Electromagnetic Interference Noise in Photoacoustic Endoscopic Image Processing. *Sensors* **2022**, *22*, 3961. https://doi.org/10.3390/s22103961

[10] Li, C.; Chen, W.; Tan, Y. Point-Sampling Method Based on 3D U-Net Architecture to Reduce the Influence of False Positive and Solve Boundary Blur Problem in 3D CT Image Segmentation. *Appl. Sci.* **2020**, *10*, 6838. https://doi.org/10.3390/app10196838

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.