Enhanced TransUNet : Integrating Dual-Attention-and-CutMix for Medical Image Segmentation

Marian Zlateva
mzlateva@ucsc.edu
Department of Computer Science
University of California, Santa Cruz

Marzia Binta Nizam
manizam@ucsc.edu
Department of Computer Science
University of California, Santa Cruz

Abstract

This study enhances the TransUnet model[3] for medical image segmentation, focusing on improving its generalization capabilities. We optimized the baseline TransUnet model by incorporating channel and dual attention mechanisms along with CutMix data augmentation. These strategic modifications significantly enhanced the model's performance, resulting in more precise segmentation of complex anatomical structures. This improvement suggests that advanced attention mechanisms and augmentation techniques are crucial for advancing medical imaging applications. Our code and models are publicly accessible here.

1 Introduction

Medical image segmentation plays a pivotal role in enhancing diagnostic processes by enabling precise segmentation of anatomical structures from medical images. This task is crucial for various clinical applications including, but not limited to, surgical planning and guided interventions. Recent advancements in deep learning have significantly improved segmentation accuracy, particularly through the introduction of traditional CNN models as well as recent transformers models.

The TransUnet[3] model, which integrates the robust feature extraction capabilities of CNNs with the global context encoding of transformers, has set new benchmarks in medical image segmentation. However, despite its success, the model faces challenges in generalization across diverse medical imaging scenarios, which is critical for its application in real-world clinical settings.

In response to these challenges, our study introduces modifications to the baseline TransUnet model to enhance its generalization capabilities. By incorporating dual attention mechanisms [6], we aim to refine the model's sensitivity to relevant features across different scales and conditions. Additionally, the integration of CutMix data augmentation further aids in this by promoting robustness against variations in input data.

These enhancements are designed not only to improve the accuracy of segmentations but also to ensure that the model can effectively adapt to the wide variety of imaging conditions encountered in practical medical settings. The advancements presented in this paper underscore the importance of targeted modifications in deep learning models to meet specific clinical demands.

Our contributions are demonstrated through comprehensive experiments, showing that our enhanced model achieves superior performance in segmenting complex anatomical structures, thus pushing the boundaries of what is achievable with current medical image segmentation technologies. All

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

methodologies and improved model configurations are made publicly available to foster further research and adaptation in the field.

2 Base Model

2.1 What is TransUnet

TransUnet was first introduced in the paper "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation"[3] by Chen et al. It was designed to perform medical image segmentation by combining the capabilities of two preexisting network architectures: U-Net and transformers. U-Net uses a U-shaped architectural design primarily built out of CNNs, which gives the model the capability to identify low-level features, while transformers help the model interpret the global context of the image by treating the image features as sequences.

2.2 Model Selection Overview

In our search for the optimal model for our medical image segmentation task, we evaluated several leading architectures on the Synapse dataset[1], including Swin-UNet[2], HiFormer[5], and TransDeepLab[1], which demonstrated higher mean Dice scores and lower median Hausdorff Distances compared to TransUNet.

Table 1: Comparison of Different Models

Model	Mean Dice ↑	Mean HD95 ↓
TransUNet [3]	0.769	32.870
Swin-UNet [2]	0.796	22.0763
HiFormer (HiFormer-b) [5]	0.784	20.556
TransDeepLab [1]	0.792	24.511

Despite these models showing marginally better segmentation performance, we ultimately selected TransUNet as our base model. This decision was influenced by several practical factors:

- **Efficient Training:** TransUNet requires significantly less training time than its counterparts, which is crucial for rapid prototyping and iterative improvements.
- **Lower Resource Demand:** It is less demanding in terms of GPU memory usage, making it more accessible and practical for use in a wider range of settings.
- **Simpler Architecture:** The architecture of TransUNet is relatively simpler to understand and modify, which is beneficial for ongoing development and experimentation.
- **Hybrid Modality:** Uniquely, TransUNet combines the robust feature extraction capabilities of CNNs with the expansive contextual awareness provided by transformers. This hybrid approach effectively enhances both local and global accuracy in segmentation tasks.

These advantages make TransUNet a more suitable choice for our project's requirements, prioritizing practical deployment and development efficiencies over slight improvements in raw performance metrics.

3 Data

For our experiment, we used the Synapse multi-organ segmentation dataset, comprising 30 abdominal CT scans. These scans encompass a total of 3,779 axial contrast-enhanced clinical CT images. The voxel spatial resolution for each volume varies within ranges from 0.54 mm to 0.98 mm in-plane and from 2.5 mm to 5.0 mm inter-slice.

[1] `https://www.synapse.org/#!Synapse:syn3193805/wiki/217789`

The dataset annotates several organs across each scan, including the aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach. Alongside these organs, a 'none' label is included, bringing the total to nine target classes. We accessed the preprocessed version of this dataset from the original paper[3].

The test dataset is formatted in HDF5 (h5), containing 12 black and white volumetric images along with their corresponding labels. Each sample in this dataset represents a 3D image, as the evaluation metrics are calculated for each volume individually. To ensure consistency in model evaluation, these volumes were transformed to match the shape of the training data. This involved resizing both the images and the labels to 224×224 pixels per slice, maintaining uniformity across the dataset.

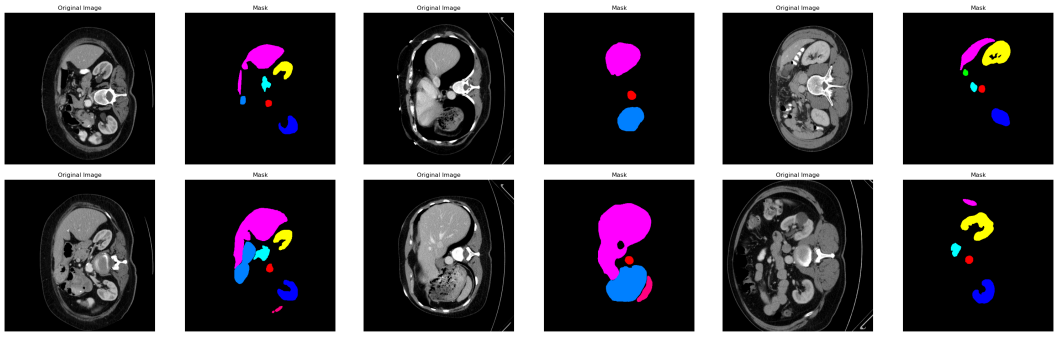

Figure 1: Visualization of Synapse Training Dataset

4 Experiment

For our experiments, we first replicated the TransUnet model following the exact setup from the original study [3]. We selected the R50-ViT-B_16 configuration, which was highlighted as the most effective model in the paper. All parameters, such as the model architecture and training settings, were kept consistent with the original specifications. This initial model served as the baseline for all subsequent experiments reported in this document. All trainings were conducted on a single RTX 3080 GPU, with the entire process taking approximately 2 hours to complete.

Upon successfully replicating the R50-ViT-B_16 configuration of the TransUnet model, we proceeded to evaluate its performance using the standard metrics from the original study. The model achieved a mean Dice coefficient of 76.97% and a Hausdorff distance (hd95) of 32.87%, demonstrating strong agreement with the findings reported in [3].

Table 2: Results for model trained with original parameters

Model	mean_dice	mean_hd95	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
TransUNet	0.769	32.870	0.868	0.596	0.814	0.740	0.945	0.541	0.873	0.778

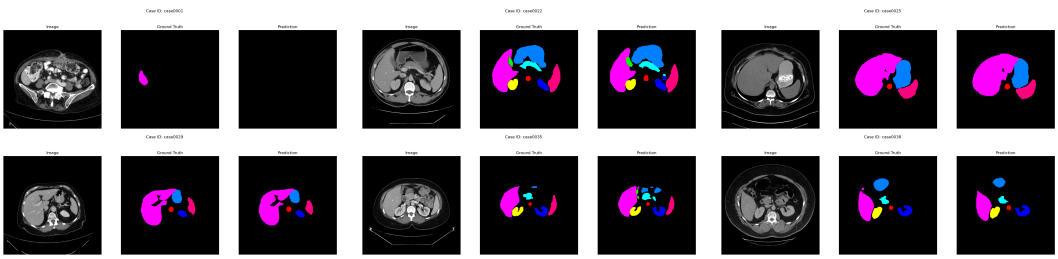

Figure 2: Visualization of Ground Truth vs Base Model Predictions

4.1 Enhancing TransUNet with Channel Attention Module

To enhance TransUNet's capability in medical image segmentation, we integrated a Channel Attention Module (CAM) into the existing R50-ViT-B_16 architecture. This module is specifically designed to amplify the significance of channel-wise features within the network, making it highly effective for detailed and nuanced tasks like organ segmentation.

How the Channel Attention Module works

The Channel Attention Module[4] functions by analyzing the interdependencies between channels of convolutional features. It computes a set of attention weights for each channel by using global average pooling to reduce spatial dimensions to a single value per channel. These weights are then used to scale the channels, amplifying features that are more relevant and suppressing the less important ones. This selective focus on informative features allows the model to adapt more dynamically to the complexities of medical images.

For our experiment, we integrated this by modifying the convolutional blocks of the R50-ViT-B_16 model to include the CAM. This was achieved by inserting the channel attention mechanisms post the convolutional operations, prior to activation, following the mechanism mentioned in the DA-TransUnet paper[6]. The CAM assesses the importance of each channel through learned weights, which adaptively adjust during training to enhance feature extraction and prioritization.

Upon integrating the Channel Attention Module (CAM) into our TransUnet model, we observed a noticeable improvement in the segmentation performance. To quantify the impact, we compared the enhanced model against the baseline R50-ViT-B_16 configuration on the Dice coefficient and Housdroff distance.

Table 3 presents the detailed score of our experiment with the Channel Attention Module. The incorporation of CAM resulted in an increase in the Dice coefficient from 76.97% in the baseline model to 9.7% in the enhanced model. The incorporation of CAM reduced the median HD95 to 23.562.

Table 3: Comparison of results between the original model and CAM integrated model

Model	mean_dice	mean_hd95	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
TransUNet	0.769	32.870	0.868	0.596	0.814	0.740	0.945	0.541	0.873	0.778
TransUNet+CA	0.797	23.562	0.864	0.645	0.828	0.763	0.948	0.616	0.903	0.812

The increased mean dice score post-CAM integration suggests that the model is now better at identifying and segmenting the organs accurately, as the dice score directly reflects the overlap between the predicted segmentation and the ground truth. Moreover, most organs show significant improvements in Dice scores after incorporating the Channel Attention Module, particularly the Pancreas and Stomach, which have historically been challenging due to their variable shape and proximity to other organs.

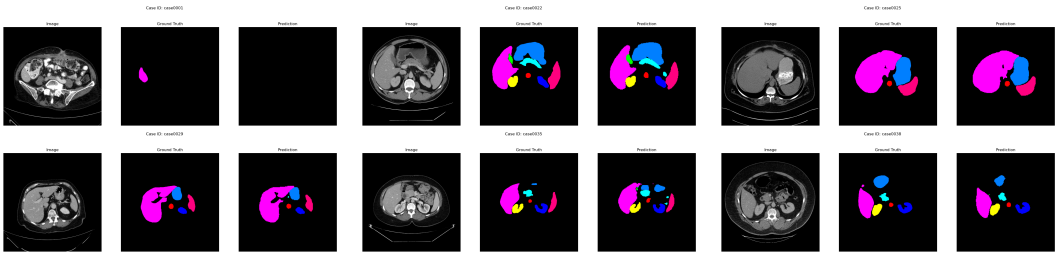

Figure 3: Visualization of Ground Truth vs Base Model(+Channel Attention) Predictions

4.2 Enhancing TransUNet with Dual Attention Mechanisms

To further enhance the segmentation capabilities of the TransUnet model, we introduced Dual Attention Mechanisms, specifically combining Channel Attention and Pyramid Attention.

What is Dual Attention with Channel and Pyramid Attention?

- **Channel Attention** focuses on identifying and emphasizing informative features across the channel dimensions of the feature maps. It modulates the feature responses channel-wise, enhancing the representational power of the network.
- **Pyramid Attention** involves processing the input through a series of convolutional layers at different scales, facilitating a multi-scale analysis of the input features. This type of attention helps the model capture both local and global contextual information, crucial for accurately delineating complex anatomical structures.

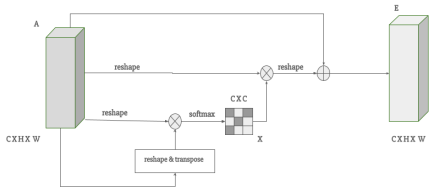

Figure 4: Architecture of Channel Attention Module [6]

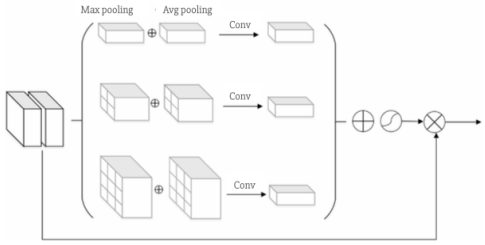

Figure 5: Architecture of Pyramid Attention Module [7]

Why Dual Attention with Channel and Pyramid?

The rationale for incorporating both Channel and Pyramid Attention stems from the need for precise segmentation in medical imaging, which often involves complex and overlapping structures. The combination offers:

- **Enhanced Depth and Detail Discernment through Channel Attention**, ensuring that the most significant features across the channels are highlighted and utilized effectively.
- **Improved Scale and Context Awareness through Pyramid Attention**, allowing the model to recognize and adapt to anatomical features at various resolutions and scales.

For implementing the Pyramid Attention Module, we followed the DA block architecture of the DA-TransUnet[6], where instead of the Positional Attention Mechanism, we used the Pyramid Attention Mechanism. Keeping the Channel Attention Module as described in the previous section, we integrated the Pyramid Attention Module at key stages within the decoder of the TransUnet.

As the upsampled features from deeper layers ascend through the decoder, they are merged with the corresponding features from the skip connections. At these junctures, Pyramid Attention is applied to

harmonize and enhance the feature sets, integrating multi-scale contextual information. This method allows the model to effectively utilize both local and broader contextual details, crucial for accurate segmentation across varying organ sizes and spatial complexities.

Upon integrating the dual attention block into the TransUnet model, we conducted a series of evaluations to assess the impact on segmentation performance across various organs in the abdominal CT scans. These evaluations utilized the same metrics as previously discussed, allowing for a direct comparison with the baseline model's performance.

Table 4: Comparison of results between the base model and model trained with Dual Attention

Model	mean_dice	mean_hd95	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
TransUNet	0.769	32.870	0.868	0.596	0.814	0.740	**0.945**	0.541	0.873	0.778
TransUNet+DA	**0.803**	**23.734**	**0.884**	**0.600**	**0.829**	**0.816**	0.943	**0.648**	**0.911**	**0.797**

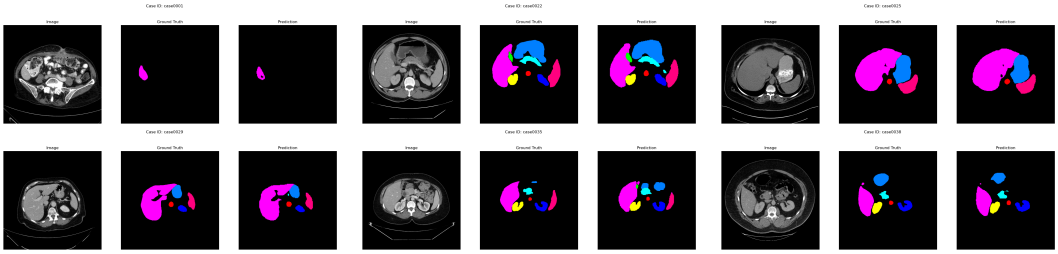

Figure 6: Visualization of Ground Truth vs Base Model(+Dual Attention) Predictions

4.3 Incorporating Advanced Data Augmentation Techniques

To further enhance the training robustness and generalizability of the TransUnet model, we incorporated CutMix, a state-of-the-art data augmentation technique, along with the Dual Attention Mechanism. CutMix helps improve model performance by encouraging the model to learn more robust and generalized features, which is especially useful in the context of medical image segmentation.

We implemented the CutMix data augmentation technique in the following way:

- **Segment Selection**: Segments from one image are randomly cut and pasted onto another, with their respective labels.
- **Size and Proportion**: The size of the segments is varied randomly, covering between 20% to 60% of the image area, ensuring diverse learning scenarios.
- **Frequency of Application**: CutMix was applied to 33% of the images in each training batch to maintain a balance between original and augmented data. For the remaining images in each batch, standard augmentation techniques such as flipping and rotating were applied as described in the original paper, ensuring comprehensive variability in the training data.

Table 5: Comparison of results between the base model and model trained with DA+Cutmix

Model	mean_dice	mean_hd95	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
TransUNet	0.769	32.870	0.868	0.596	0.814	0.740	0.945	0.541	0.873	0.778
TransUnet(ours)	**0.823**	**19.743**	**0.882**	**0.631**	**0.860**	**0.830**	**0.946**	**0.693**	**0.907**	**0.833**

The integration of Dual Attention mechanisms and CutMix data augmentation significantly enhanced the TransUnet model's segmentation capabilities. The overall Mean Dice score increased by approximately 6.9% to 0.822, while the median Hausdorff Distance (HD95) decreased by about 39.9% to 19.743 mm, indicating more precise organ boundary delineation. Notably, organs such as the gallbladder, both kidneys, and the pancreas showed marked improvements in segmentation accuracy, with the

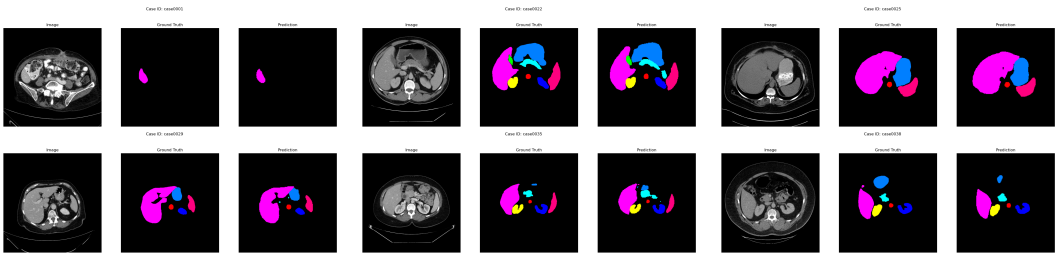

Figure 7: Visualization of Ground Truth vs Base Model(+Dual Attention Cutmix) Predictions

pancreas, spleen, and stomach achieving higher Dice scores. These enhancements demonstrate the model's improved performance in accurately segmenting complex anatomical structures.

5 Ablation Study

When improving the TransUNet model, iteratively trying out different modifications was essential to determine what worked and what didn't. In the figure below we have an overview of the specific modifications we made to the model and the effects these modifications had on the model's overall capabilities.

Initially we tried several different methods of augmenting the input data. Adding a random horizontal flipping of all the data, resulted in a mild improvement compared to the baseline, however this improvement was not substantial enough to consider it remarkable in the context of our project. We also tried only flipping part of the data, however this resulted in results that were worse than the baseline.

Another method we tried for data augmentation was adding different types of filters to the input data. The intent for these filters was to hopefully reduce the noise in the raw imaging data since it was quite noisy and possibly difficult for the convolutional layers to train on. In our experiments, both gaussian and median filters were tested as they were the standard for reducing noise. Unfortunately, adding these filters also removed essential data, causing the model to predict the correct segmentation less accurately.

Finally, we tried some other methods that focused on tuning the model itself. Out of our attempts: cyclic learning rate, upsampling, and adding a new convolutional layer, it appears that upsampling was the most effective. Unfortunately, it only proved to create a mild improvement for the model. Overall, none of these methods were effective enough to vastly improve the model, and as a result, we decided to exclude them from the final version.

Table 6: Comparison of results between different methods

mean_dice	mean_hd95	method
0.769	32.870	Baseline
0.769	31.986	Horizontal flip all data
0.779	22.887	Horizontal flip train + original test
0.760	36.953	Add Gaussian Smooth to all data
0.764	32.020	Median smooth on all data
0.763	40.754	Adam Optimizer
0.768	31.070	Cyclic learning rate
0.778	29.466	Upsampling = nearest
0.769	32.870	Add new convulational layer

6 Discussion

Throughout this study, our primary objective was to establish a robust baseline model for medical image segmentation and subsequently enhance its generalization capabilities through various strategic modifications. The TransUNet model served as our foundation, chosen for its balance between performance, efficiency, and ease of use.

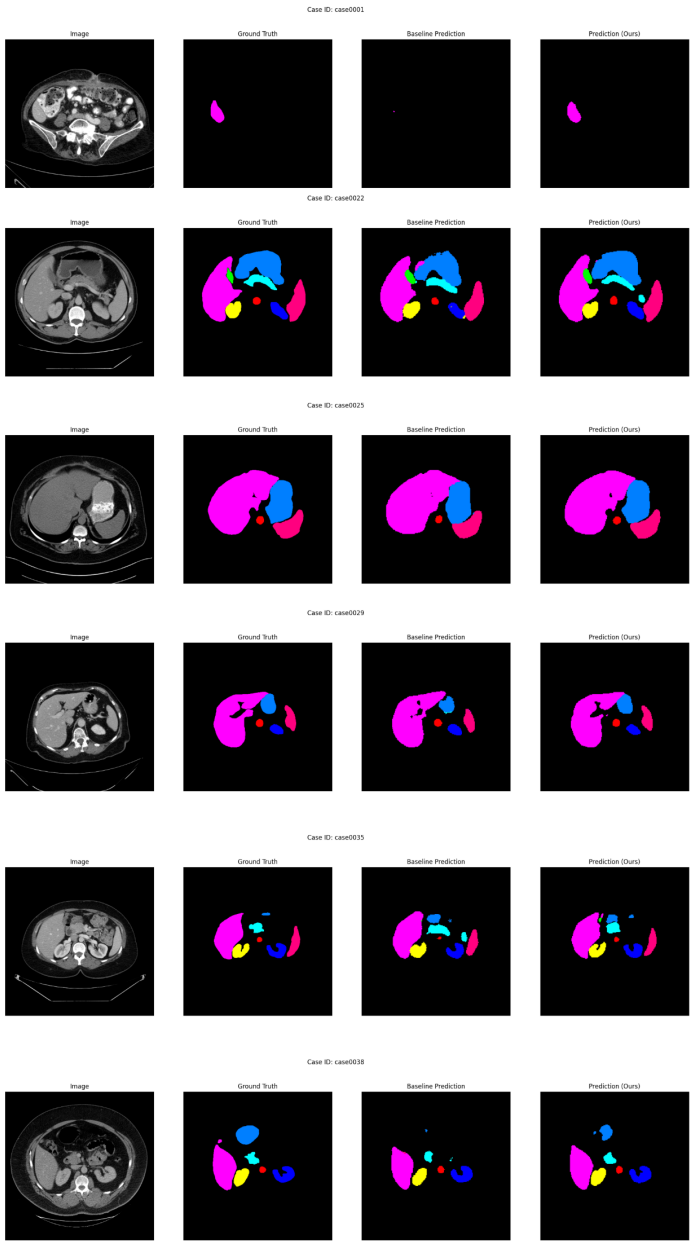

Figure 8: Segmentation results of TransUNet and Our model on the Synapse dataset.

Our incremental enhancements began with the integration of a channel attention mechanism, which improved the Mean Dice score from 0.769728 to 0.798 and reduced the median Hausdorff Distance from 32.870214 mm to 23.562. Building on this, the introduction of dual attention mechanisms further pushed the Mean Dice score to 0.804 and maintained the Hausdorff Distance at a comparable level of 23.734.

The most significant improvements were observed after implementing the dual attention mechanisms alongside CutMix data augmentation. This combination not only elevated the Mean Dice score to 0.822 but also markedly reduced the median Hausdorff Distance to 19.743. Notably, the model demonstrated substantial gains in the accurate segmentation of challenging organs such as the pancreas and kidneys, with notable improvements across all organs.

7 Conclusion

This study successfully enhanced the TransUNet model, significantly improving its generalization for medical image segmentation. By integrating dual attention mechanisms, along with CutMix data augmentation, we achieved notable gains in segmentation precision and accuracy, particularly for complex anatomical structures. These strategic modifications have proven critical in advancing the model's performance, offering promising directions for future research in the application of deep learning to diverse medical imaging tasks.

References

[1] Reza Azad et al. "Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation". In: *International Workshop on PRedictive Intelligence In MEdicine*. Springer. 2022, pp. 91–102.

[2] Hu Cao et al. "Swin-unet: Unet-like pure transformer for medical image segmentation". In: *European conference on computer vision*. Springer. 2022, pp. 205–218.

[3] Jieneng Chen et al. "Transunet: Transformers make strong encoders for medical image segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[4] Jun Fu et al. "Dual attention network for scene segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3146–3154.

[5] Moein Heidari et al. "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2023, pp. 6202–6212.

[6] Guanqun Sun et al. "DA-TransUNet: integrating spatial and channel dual attention with transformer U-net for medical image segmentation". In: *Frontiers in Bioengineering and Biotechnology* 12 (2024), p. 1398237.

[7] Fuli Yu et al. "A multi-class COVID-19 segmentation network with pyramid attention and edge loss in CT images". In: *IET Image Processing* 15.11 (2021), pp. 2604–2613.