Comparative Analysis of CNN and Transformer Models for Medical Image Segmentation

Taehwan Park
Department of Computer Science
University of Toronto
Toronto, ON, Canada
taehwan.park@mail.utoronto.ca

Jiho Shinn
Department of Computer Science and Engineering
Korea University
Republic of Korea
tlswlgh0801@korea.ac.kr

Seo Won Yi
Department of Computer Science
Department of Statistics
University of Toronto
Toronto, ON, Canada
sean.yi@mail.utoronto.ca

ChatGPT 4
Open AI

Abstract

This study presents a comprehensive analysis of six neural network models—three Convolutional Neural Networks (CNNs) and three Transformers—applied to medical image segmentation using MRI heart images. The models, including U-Net, LRASPP, FCN, MedT, SegFormer, and Swin Transformer, were evaluated under varying conditions such as transfer learning and the use of different loss functions. Our findings highlight the performance trade-offs between these models and their dependency on data augmentation and architectural differences. The study primarily employs the Dice score to gauge model performance and investigates the impact of employing alternative loss functions to address prediction biases. This research provides insights into the strengths and limitations of each model, offering a path forward for enhancing segmentation accuracy in medical imaging.

Introduction

Problem Definition

The creation of neural networks and deep learning models has brought us into a golden era of computer programming. With the surge of various models, it's challenging to review each one and grasp their unique roles. We aim to provide analyses of Convolutional Neural Network (CNN) models and Transformer models designed for image segmentation under various conditions.

We utilized about 2,300 MRI images of the heart and evaluated six different models—three CNNs and three Transformers [1]. We assessed the outcomes of applying transfer learning compared to starting model training from scratch. Additionally, we explored how different loss functions affect the outcomes and assessed the performance differences between CNNs and Transformers.

The models we looked into include the U-Net, Lite Reduced Atrous Spatial Pyramid Pooling (LRASPP), and Fully Convolutional Network (FCN) for CNNs, along with the Medical Transformer (MedT), SegFormer, and Swin Transformer for Transformers.

Preprint. Under review.

We tested various loss functions using the model that performed best out of the six. Moreover, with the same dataset and this top-performing model, we examined the differences between using transfer learning and starting training from the beginning.

Methodology

Prior Work

Automatic segmentation algorithms can broadly be categorized into CNN-based and Transformer-based approaches. Despite the impressive performance of CNNs in image segmentation, studies have revealed certain limitations. CNNs fail to explicitly model long-range dependencies due to the intrinsic locality and weight sharing of receptive fields in convolution operations [2]. They may struggle to generalize well in medical images because variations in parts of the image away from the local pathology are often normal [3].

In response to these challenges, Transformer-based models have emerged as a promising alternative. Transformers have shown remarkable progress in capturing long-range dependencies in medical image analysis. However, current Transformer-based models suffer from limitations such as failure to capture important image features due to a naive tokenization scheme, information loss from considering only single-scale feature representations, and inaccuracies in segmentation label maps without rich semantic contexts and anatomical textures [2].

Recent research has focused on hybrid architectures combining both models or integrating additional ideas such as adversarial training and saliency maps to address these challenges [4]. The significance of our project lies in conducting an empirical study on fundamental backbone architectures to expand the technical understanding and explore potential directions for advancing medical image segmentation technology.

The models used for the studies are all pre-developed and/or pre-trained models that can be easily imported from PyTorch, HuggingFace, and GitHub.

Table 1: Model Sources

Name	Source	Comment (Pre-trained weight or Source address)
U-Net	GitHub	https://github.com/shailensobhee/medical-decathlon
FCN	PyTorch	FCN_ResNet101_Weights
LRASPP	PyTorch	LRASPP_MobileNet_V3_Large_Weights
MedT	GitHub	https://github.com/jeya-maria-jose/Medical-Transformer
SegFormer	Hugging Face	nvidia/mit-b2
Swin Transformer	Hugging Face	microsoft/swinv2-large-patch4-window12-192-22k

The initial draft of the custom decoder for Swin Transformer was obtained from the work from Olga Mindlina which concatenates the feature maps from the Swin Transformer and performs a simple upscaling [5].

Dataset Construction and Splitting Strategy

For our study, we selected the heart dataset from the Medical Segmentation Decathlon. This dataset poses significant challenges due to its relatively small size and the large variability among samples. It specifically comprises 30 mono-modal MRI 3D volumes of the Left Atrium, provided by King's College London, with 20 volumes designated for training and 10 for testing, highlighting the challenges associated with a small training dataset.

Given the limited size of the dataset, which inherently presents a challenge, it also facilitates expedited training times. In our supervised learning setup, only the 20 training volumes, each containing both the input and target images, were utilized. We partitioned these videos into training, validation, and test sets with ratios of 70%, 10%, and 20%, respectively, resulting in 14 videos for training, 2 for validation, and 4 for testing.

Subsequently, we converted these 3D volumes into 2D frames to align with the input format typically required by segmentation models, yielding 1491 frames for training, 370 frames for validation, and 410 frames for testing.

It is noteworthy that a similar methodology was previously applied in the development of the U-Net architecture, which also includes options for data augmentation and frame shuffling. In the U-Net implementation, there is flexibility to distribute frames from a single video across the training, validation, and testing sets. However, we opted against this approach to maintain the integrity of our evaluation process. By ensuring that all frames from a given video were assigned to a single subset (training, validation, or test), we avoided the potential risk of "data leakage," where the model could inadvertently learn specific features from the test set during training. This strict partitioning is essential to ensure a fair assessment of the model's performance and its ability to generalize to new, unseen data.

Data Augmentation

The paucity of extensive training datasets poses a significant challenge in the field of medical image segmentation, where robust model generalization is crucial. To address this issue, data augmentation plays a pivotal role by artificially enhancing the size and diversity of training datasets, thereby improving model robustness against overfitting.

In our study, we adopted a dynamic data augmentation strategy that reflects the approach used in the training of the U-Net architecture. This methodology involves randomly applying horizontal and vertical flips, as well as rotations to the images during the training phase. Specifically, each image frame retrieved by the data loader within the training loop may undergo random transformations, ensuring that the same frame can appear in various orientations and configurations across different epochs.

The augmentation process is implemented as follows: With a probability of 50%, a frame may be flipped horizontally or vertically. The axis of flipping—horizontal or vertical—is determined randomly, promoting variability in the training data. Additionally, with the same probability, a frame may also be rotated by 90, 180, or 270 degrees, selected randomly. This randomness in rotation and flipping is intended to simulate various anatomical orientations and imaging conditions, thus providing a more comprehensive training regime.

To quantitatively assess the impact of these augmentation techniques, we conducted experiments both with and without the application of data augmentations. This comparative analysis aims to elucidate the efficacy of data augmentation in enhancing the segmentation performance and generalization capabilities of the models under consideration.

Introduction of Models

We analyzed six models in total, three of which are CNNs and the other three are Transformer models.

U-Net

U-Net is a specialized architecture for medical image segmentation that first gained attention through the paper "U-Net: Convolutional Networks for Biomedical Image Segmentation" [6]. This model was specifically designed to handle the unique challenges of medical image analysis, including the need for precise localization and detailed contextual understanding of anatomical structures in various types of medical imaging, such as MRI and CT scans. As the name suggests, U-Net consists of a U-shaped structure which is thoughtfully constructed to capture both the context and the localization details required for accurate segmentation.

The U-Net model used for this paper is imported from GitHub page `https://github.com/shailensobhee/medical-decathlon`. The model consists of 0.48M parameters. Minor adjustments were added to fit the heart data specifically.

LRASPP

LRASPP, or Lite Reduced Atrous Spatial Pyramid Pooling, is an efficient neural network architecture tailored for semantic image segmentation tasks, particularly in resource-constrained environments. It was first introduced as part of the broader advancements made in the paper "Searching for MobileNetV3" [7]. This model combines the principles of deep learning with the necessity of operational

efficiency on mobile and edge devices. The primary focus of the model was to reduce the computational cost while maintaining competitive accuracy and performance. This was achieved by utilizing a series of parallel atrous convolutions that operate at different dilation rates.

The model used for this paper is imported from PyTorch which follows the architecture introduced in the paper "Searching for MobileNetV3." The model utilizes MobileNetV3-Large backbone and is pre-trained on the subset of the Common Objects in Context (COCO) dataset with VOC labels. It consists of 3M parameters in total.

FCN

The FCN model, or Fully Convolutional Network, is an architecture designed primarily for the task of semantic image segmentation. It was first introduced in the paper "Fully Convolutional Networks for Semantic Segmentation" [8]. Unlike traditional convolutional neural networks that output classification scores, FCNs are capable of outputting spatial maps of classes, making them particularly suited for tasks where understanding the spatial arrangement of objects within an image is crucial.

As the name suggests, the FCN model uses fully convolutional layers, which replace the dense layers typically found in standard CNNs. The model used for this paper is obtained from PyTorch. The model is built on ResNet-101 architecture and is pre-trained with the subset of the COCO dataset with VOC labeling. The model consists of 54M parameters in total.

MedT

MedT, or the Medical Transformer, is specifically engineered for the demands of medical image processing. It was first introduced in the paper titled "Medical Transformer: Gated Axial-Attention for Medical Image Segmentation" [9]. This model innovatively combines the principles of the Transformer architecture with adaptations that address the unique challenges posed by medical imaging datasets, which often include high variability, irregular shapes, and differing textures that are essential for accurate diagnosis and analysis. The Medical Transformer utilizes a gated axial-attention mechanism which enhances its ability to focus on specific anatomical features within the images, significantly improving the precision of segmentation tasks compared to traditional convolutional approaches.

The model that was used in this paper is obtained from Jeya Maria Jose Valanarasu GitHub repository [10]. A minor modification was made to fit the data and the purpose. The model consists of 1.5M parameters.

SegFormer

SegFormer is a neural network architecture designed specifically for semantic segmentation tasks, which effectively addresses the challenges of segmenting fine details and global context in images. Introduced in the paper "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," SegFormer combines the strengths of transformers and convolutional operations to achieve high performance on segmentation tasks [11]. Segformer utilizes a mix transformer encoder as a backbone and uses a lightweight decoder that does not include a self-attention mechanism.

The model that was used in the paper was obtained from the Transformers module from HuggingFace. The model is pre-trained using ImageNet-1k and contains a total of 28M parameters. Both fine-tuning and training from scratch approaches were utilized without changing the structure too much but just adding an appropriate head layer.

Swin Transformer

Swin Transformer was originally introduced in the paper “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" [12]. The Swin Transformer is designed to handle the dense and structured nature of image data better than traditional Transformer models by utilizing a hierarchical structure and shifted window self-attention mechanism.

The model that was used in this paper was obtained from the Transformers module from HuggingFace. The model’s encoder which was pre-trained on ImageNet-21k at resolution 192×192 was utilized while the decoder was custom-designed by implementing channel attention mechanism and up-sampling through pixel shuffle operations. The model’s encoder includes 195M parameters while the decoder includes 132M parameters.

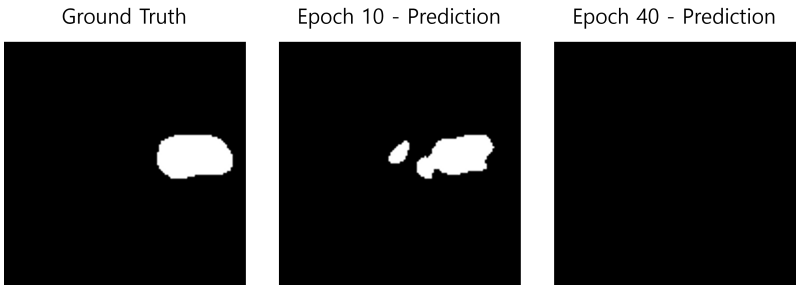

Figure 1: MedT model prediction results for the same target at different training epochs. In epoch 10, the model correctly predicts some foreground regions, though not exactly matching the target. By epoch 40, the model's prediction is entirely background (all black), highlighting the tendency to favor predicting the background as the model is trained over time.

Introduction of Loss

The Dice score, also known as the Sørensen-Dice coefficient, is a widely used evaluation metric in medical imaging tasks. It measures the similarity between the predicted and ground truth segmentation masks and is calculated as follows:

$$dice\_score = \frac{2\sum p_i g_i}{\sum p_i + \sum g_i}$$

Where p_i and g_i are the predicted and ground truth values for pixel i, respectively. The Dice score ranges from 0 to 1, with higher values indicating greater overlap and better agreement between the predicted and ground truth masks. In our study, we use the Dice score as our primary evaluation metric to assess the performance of our models.

Soft Dice Loss

The soft Dice loss function is derived from the Dice score and is widely used in medical imaging tasks [13, 14]. It is calculated as follows:

$$soft\_dice\_loss = 1 - \frac{2\sum p_i g_i + \epsilon}{\sum p_i + \sum g_i + \epsilon}$$

This loss function is known as the soft Dice loss because we directly use the predicted probabilities instead of thresholding and converting them into a binary mask. By working with the raw probabilities, the soft Dice loss provides a smoother and more differentiable metric that is beneficial for model training. The soft Dice score, calculated as $\frac{2\sum p_i g_i + \epsilon}{\sum p_i + \sum g_i + \epsilon}$, quantifies the similarity between the predicted and ground truth segmentation masks.

The addition of a small constant ϵ prevents division by zero. The soft Dice loss ranges from 0 to 1, with lower values indicating better agreement between the predicted and ground truth masks.

During our empirical studies, we observed a notable bias in the model's predictions when using the soft Dice loss. The model tended to favor predicting the background as it was trained, primarily because most of the images consisted predominantly of background pixels. This phenomenon is illustrated in Figure 1.

To address this issue, we designed three alternate loss functions based on soft Dice loss that aim to mitigate the bias towards the background class and improve model performance as following:

- *FN penalty Dice loss*: This loss function includes a penalty for false negative predictions, addressing the model's tendency to favor background predictions and potentially missing important features in the images.

$$soft\_dice\_loss + mean(g > p) \times penalty\_weight$$

- *FN/FP penalty Dice loss*: This loss function extends the FN penalty Dice loss by also including a penalty for false positive predictions. This adjustment helps balance the model's

sensitivity and specificity.

$$soft\_dice\_loss + mean(g > p) \times penalty\_weight1 + mean(p > g) \times penalty\_weight2$$

- *Focal Dice loss*: Combining the concepts of focal loss and Dice loss, this loss function focuses more on hard-to-classify cases by assigning higher weights to them, aiming to improve overall model performance. By altering the value of γ, one can decide how to penalize lower Dice score cases. Adjusting α allows amplifying or reducing the magnitude of loss value. A higher value of α would ensure escape from the local minimum when dealing with highly imbalanced data which are often the case for image segmentation case.

$$-\alpha(1 - soft\_dice\_score)^{\gamma} \cdot log(soft\_dice\_score + \epsilon)$$

By comparing the performance of these alternate loss functions with the baseline soft Dice loss, we aim to identify a loss function that provides better segmentation performance while reducing the bias toward background predictions.

Results and Discussion

Empirical Studies

For our empirical evaluation, we utilized the soft Dice loss as the primary loss function during the training phase. Model performance was assessed on the test dataset using the hard Dice score, computed to three decimal places. The heart dataset, drawn from the Medical Segmentation Decathlon, served as the basis for our analysis. Throughout the training process, the Adam optimizer was employed, with computational resources provided by Google Colab, including A100, V100, and L4 GPUs.

Table 2: Dice Scores for CNN and Transformer Models on the Heart Dataset. **Note:** 'Data Augmented' indicates whether data augmentation was used. 'Pre-trained' denotes models initialized with weights from related tasks to enhance learning efficiency and accuracy. The 'Dice Score' measures model performance, and 'Parameters (millions)' indicates the complexity of each model.

Model Name	Data Augmented	Dice Score	Pre-trained	Parameters (millions)
U-Net	True	0.422	False	0.48
LRASPP	False	0.482	True	3
LRASPP	True	0.500	True	3
FCN-ResNet101	False	0.679	True	54
FCN-ResNet101	True	0.644	True	54
MedT	True	0.375	False	1.5
SegFormer_mi2	False	0.792	True	28
SegFormer_mi2	True	0.695	True	28
Swin Transformer	True	0.200	False	132

The results, as detailed in Table 2, illustrate significant variability in model performance based on the presence or absence of data augmentation. Notably, the SegFormer_mi2 model exhibited the highest Dice scores in both scenarios, achieving particularly robust performance without data augmentation (Dice Score = 0.792). In contrast, the LRASPP model showed modest performance, with scores adversely impacted in the absence of data augmentation.

Initially, our study aimed to directly compare CNNs with Transformers, but it became evident that factors like data augmentation and pre-training were pivotal, influencing the outcomes significantly. Thus, our analysis pivoted to comparing the top-performing models within each category, FCN-ResNet101 for CNNs and SegFormer_mi2 for Transformers, to provide a clearer picture of their respective capabilities and limitations within the scope of medical image segmentation. This approach highlights the nuanced role of model architecture alongside training strategies in determining performance efficacy.

Transfer Learning vs Full Learning: Segformer vs FCN

In our research, we juxtaposed the strategies of full training and transfer learning across two advanced model architectures: FCN-ResNet101 and Segformer_mi2. Inspired by methodologies from seminal works such as BERT and BEIT, our approach was to explore the viability of adapting pre-trained models, developed initially for RGB image segmentation, to medical imaging tasks. This exploration was motivated by the desire to harness models trained in non-medical contexts for medical segmentation, thereby potentially expanding their applicability and effectiveness.

Given the computational constraints posed by Google Colab, such as session durations, our initial full training attempts with the U-Net and MedT models, well-known in biomedical image segmentation, highlighted the challenges of extensive training times. Consequently, we excluded U-Net and MedT from pre-training consideration, as they are typically tailored specifically for medical image segmentation and do not have pre-trained versions on RGB datasets. This exclusion allowed us to focus on models that could potentially be adapted from non-medical to medical segmentation tasks.

We selected the FCN-ResNet101 and Segformer_mi2 models based on their parameter scales—54 million for FCN-ResNet101 and 28 million for Segformer_mi2. During the experiments, only the classifier layers of FCN and the decoder head of Segformer were unfrozen to facilitate fine-tuning.

Our empirical findings revealed distinctive behaviors between these models under different training paradigms. The Segformer consistently outperformed the FCN in scenarios involving transfer learning, achieving a Dice Score of 0.792 without data augmentation, compared to FCN's 0.679. This suggests that Segformer's architecture possibly offers better feature transferability from non-medical to medical segmentation tasks.

Conversely, when subjected to full training, the Segformer's performance markedly deteriorated, particularly without data augmentation, where it scored only 0.3774. This was attributed to challenges in model convergence, necessitating a significantly reduced learning rate of 0.0001 to prevent instability. In contrast, the FCN displayed a noteworthy improvement under full training without data augmentation, elevating its Dice Score to 0.7495.

Table 3: Impact of Fine-Tuning on Model Performance for Heart Image Segmentation

Model Name	Data Augmented	Fine-Tuned	Dice Score	Parameters (millions)
FCN-ResNet101	True	True	0.644	54
SegFormer_mi2	True	True	0.695	28
FCN-ResNet101	False	True	0.679	54
SegFormer_mi2	False	True	0.792	28
FCN-ResNet101	True	False	0.628	54
SegFormer_mi2	True	False	0.3774	28
FCN-ResNet101	False	False	0.7495	54
SegFormer_mi2	False	False	0.3774	28

This comparative analysis underscores the nuanced demands of full versus transfer learning in medical image segmentation, particularly highlighting the enhanced stability and efficacy of transfer learning for Transformer-based models like the Segformer. On the other hand, CNN models like the FCN demonstrated more robust full-learning capabilities compared to Transformer-based models.

Pick the best model, compare multiple loss functions

We evaluated the performance of different loss functions using our best model, SegFormer, on the heart dataset. This model was trained with transfer learning and without data augmentation. The results for the Dice score across train, validation, and test sets are in Table 4.

Overall, the FN/FP penalty Dice loss appears to be the most promising loss function for this dataset, providing the highest test performance.

To see how different methods for calculating loss affect another model and dataset, we tested the U-Net model on a brain tumor MRI dataset [1]. This brain tumor dataset is much bigger than the heart dataset, giving us more data to use for training, validating, and testing. In detail, the brain tumor dataset has 411 video files for training, 36 for validation, and 37 for testing, making a total of 75,020

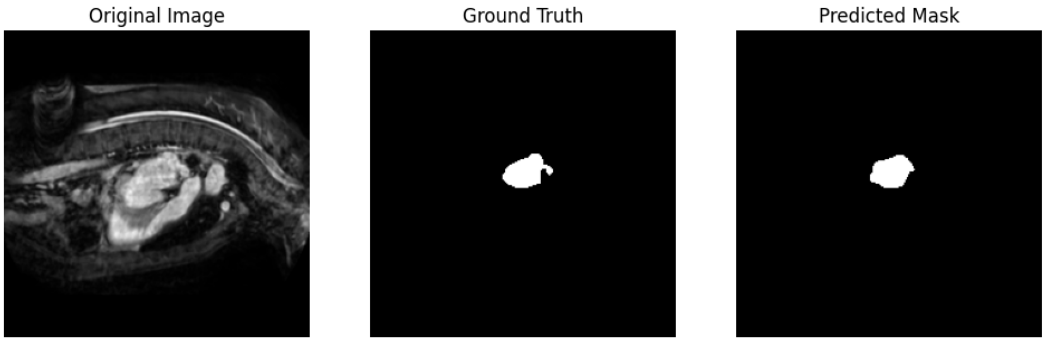

Figure 2: Ground truth and prediction by the Segformer trained with FN/FP penalty Dice loss, achieving the highest test Dice score.

Table 4: SegFormer: Heart, Data Augmented=False, Transfer Learning=True

Loss	Train Dice Score	Validation Dice Score	Test Dice Score
Soft Dice (Baseline)	0.8488	0.8470	0.7915
FN Penalty Dice	0.8634	0.8513	0.7744
FN/FP Penalty Dice	0.8512	0.8452	0.8022
Focal Dice	0.8276	0.8252	0.7748

frame images. We fully trained the U-Net model and enhanced the training with data augmentation techniques. The results for the Dice score across the three datasets (i.e., training, validating, and testing datasets) are listed in Table 5.

Table 5: U-Net: Brain Tumour, Data Augmented=True, Transfer Learning=False

Loss	Train Dice Score	Validation Dice Score	Test Dice Score
Soft Dice (Baseline)	0.6316	0.7246	0.6898
FN Penalty Dice	0.6811	0.7454	0.7254
FN/FP Penalty Dice	0.6606	0.7308	0.4730
Focal Dice	0.5459	0.4889	0.4767

While FN penalty Dice loss demonstrated improvements over soft Dice loss in the brain tumour dataset, FN/FP penalty Dice and focal Dice losses struggled with generalization. The soft Dice loss performed consistently well across both datasets and models.

The evaluation of different loss functions across two distinct datasets and models reveals varying performances based on the choice of loss function. The FN/FP penalty Dice loss showed the highest potential in the heart dataset, suggesting the benefit of considering both types of errors. In contrast, soft Dice loss performed consistently well across both datasets, particularly in test scenarios. Given the variation in performance, the choice of loss function should be tailored to the specific task, dataset, and model being used. Future research could explore further refinements to loss functions to achieve improved balance and performance across different settings.

Limitation

Due to hardware limitations, our experiments were constrained to smaller datasets, particularly when working with transformer models. This restriction may have limited our ability to fully explore the performance of our models on larger datasets or with more complex architectures. We tried to restrict the training time to be maximum 3 hours which may not be enough if we are not utilizing the transfer learning technique. The primary dataset used in our study is the heart dataset, which is relatively small in size ($\sim$ 1,500 training data). The limited number of training, validation, and testing samples may affect the robustness and reliability of our model's performance. Our results may not generalize well to larger datasets or other medical imaging tasks.

Our study primarily focused on modifying loss functions based on the soft Dice loss. This approach was guided by the risk minimization principle, which suggests that the loss function used during training should match the loss function used for evaluation during testing [15]. While this approach addressed some biases in the soft Dice loss, it may not offer significant advantages over the original Dice loss in terms of performance. In our study, we also experimented with combining cross-entropy loss and Dice loss; however, this combination did not yield improved performance. Further research is needed to identify alternative loss functions that could provide better outcomes for medical image segmentation.

Conclusion

The study concludes that Transformer models, particularly the SegFormer, demonstrated superior adaptability and performance in medical image segmentation tasks, particularly when fine-tuned from pre-trained states. However, CNNs showed resilience and effectiveness, particularly when trained from scratch. The exploration of different loss functions revealed that while traditional soft Dice loss generally performs consistently, alternative formulations like the FN/FP penalty Dice loss can offer improved results by addressing both false positives and negatives, which is crucial for medical diagnostics. This work underscores the importance of model and loss function selection based on specific dataset characteristics and segmentation requirements. Future research should further explore hybrid models and advanced loss functions to optimize performance across diverse medical imaging tasks.

Appendix

Individual Contribution

Taehwan Park

- Responsible for Data Preparation, Augmentation and implementing LRASPP, FCN, SegFormer
- Developed FN/FP penalty Dice Loss
- Provided in-depth analysis on the results and findings
- Led the group by providing extensive knowledge in transfer learning and fine-tuning

Jiho Shinn

- Responsible for implementation of MedT and Unet.
- Developed FN penalty Dice Loss
- Analyzed results from the usage of different loss functions
- Provided intellectual findings regarding model development and research

Seo Won Yi

- Responsible for implementing Swin Transformer with custom decoder
- Developed Focal Dice Loss
- Introduced models and relative information
- Provided critical discussion on handling the data and model selection

References

[1] Medical Decathlon. (2018). *Medical Segmentation Decathlon* [Data set]. Retrieved from `http://medicaldecathlon.com`

[2] You, C., Zhao, R., Liu, F., Dong, S., Chinchali, S., Topcu, U., Staib, L., & Duncan, J. S. (2022). Class-Aware Adversarial Transformers for Medical Image Segmentation. *arXiv*. `Phttps://arxiv.org/abs/2201.10737`

[3] Pawlowski, N., Bhooshan, S., Ballas, N., Ciompi, F., Glocker, B., & Drozdzal, M. (2020). Needles in Haystacks: On Classifying Tiny Objects in Large Images. *arXiv*. https://arxiv.org/abs/1908.06037

[4] Fontanella, A., Antoniou, A., Li, W., Wardlaw, J., Mair, G., Trucco, E., & Storkey, A. (2023). ACAT: Adversarial Counterfactual Attention for Classification and Detection in Medical Imaging. *arXiv*. https://arxiv.org/abs/2303.15421

[5] Mindlina, O. (2024). Vision Transformer for Semantic Segmentation on Medical Images: Practical Uses and Experiments. *Medium*. Retrieved from https://medium.com/@olga.mindlina/vision-transformer-for-semantic-segmentation-on-medical-images-practical-uses-and-experiments-a2e3939d9870

[6] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*. https://arxiv.org/abs/1505.04597

[7] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. *arXiv*. https://arxiv.org/abs/1905.02244

[8] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *arXiv*. https://arxiv.org/abs/1411.4038

[9] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. *arXiv*. https://arxiv.org/abs/2102.10662

[10] *GitHub repository for Medical Transformer*. Retrieved from https://github.com/jeya-maria-jose/Medical-Transformer

[11] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv*. https://arxiv.org/abs/2105.15203

[12] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv*. https://arxiv.org/abs/2103.14030

[13] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. *LNCS*, 10553, 240–248. https://doi.org/10.1007/978-3-319-67558-9_28

[14] Milletari, F., Navab, N., Ahmadi, S.A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. *IEEE*.

[15] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.