Computer-Aided Segmentation of Gastrointestinal Structures

Srikaran Boya
srikaran@mit.edu

Joseph Kajon
kajon145@mit.edu

Zeki Yan
zikaiyan@mit.edu

Massachusetts Institute of Technology

Abstract

This paper introduces an innovative approach to computer-aided diagnosis for radiation therapy planning in gastrointestinal (GI) cancer patients. By applying deep learning methodologies on the segmentation of stomach and intestines from MRI scans, we aim to expedite treatment procedures and optimize therapy outcomes. By leveraging anonymized MRIs sourced from the UW-Madison Carbone Cancer Center, we employ advanced models such as UNet, DeepLabv3+, and R-CNNs, exploring multi-task learning through supervised approaches. We found that Efficient Net excels at segmenting the stomach and large bowels, while Mask R-CNN delivers the most balanced high performance across all three organs. This innovative approach has the potential to revolutionize radiation therapy planning, expediting treatment procedures and enabling radiation oncologists to concentrate more on treatment optimization, ultimately enhancing the overall effectiveness of therapies.[1]

Keywords: MRI segmentation, Deep learning, U-Net, Efficient Net, DeepLabv3+, Mask R-CNNs, Multi-task learning

1. Introduction

Gastrointestinal (GI) tract cancer poses a significant global health burden for millions of individuals diagnosed annually. Radiation therapy is the main treatment in combating this malignancy, with about half of diagnosed patients being eligible. The treatment is typically administered over 10-15 minutes per day for 1-6 weeks, and the efficacy relies heavily on detailed treatment planning to focus high radiation doses to tumors while avoiding critical organs like stomachs and intestines. This often requires labor-intensive manual segmentation of vital organs from MRI scans, extending treatment sessions from 15 minutes to an hour. This process not only prolongs treatment durations, taking a toll on patients, but also introduces the potential for human error. With these challenges in mind, this paper presents an innovative approach to streamline radiation therapy planning through the integration of computer-aided diagnosis techniques.

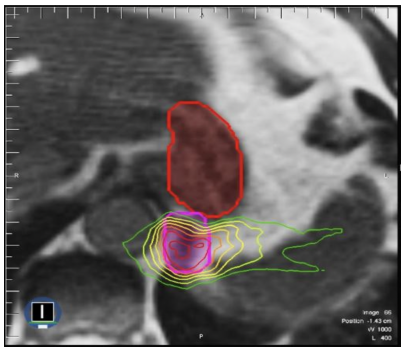

Figure 1. Example of GI Tract with Stomach (Red)

The primary objective of our project is to develop a cutting-edge deep learning model capable of automating the segmentation of the stomach and intestines from MRI scans of GI cancer patients undergoing radiation therapy. By leveraging state-of-the-art models such as UNet, DeepLabv3+, and R-CNNs and employing multi-task learning techniques enhanced by positional embeddings, we seek to revolutionize the treatment planning workflow, enabling radiation oncologists to focus more on treatment optimization rather than manual segmentation tasks.

For this project, our model automatically segments the stomach and intestines on MRI scans, as seen in Figure 1. The model inputs are MRI scans provided by UW-Madison Carbone Cancer, which reflect scans of cancer patients undergoing radiation treatment. Each patient typically undergoes 1-5 MRI scans on separate days throughout their treatment. We use U-Net, DeepLabv3+, R-CNN, and multi-task learning techniques with positional embedding to predict the segmentation.

[1] https://github.com/zikaiyan/Computer-Aided-Diagnosis-GI-Tract-Image-Segmentation

2. Related Work

2.1. U-Net Type Neural Networks

Ronneberger *et al.* [11] introduced a neural network architecture known as U-Net in 2015, originally designed for segmenting electron microscopic images. It was later discovered to be highly effective in segmenting biomedical images as well.

2.2. DeepLabv3+

DeepLabv3 [3] and DeepLabv3+ [4], introduced by Chen et al. (2017, 2018), enhance semantic segmentation by employing atrous convolution and an encoder-decoder structure, respectively, setting new benchmarks for image segmentation efficiency and precision.

2.3. EfficientNet

EfficientNet, introduced by Tan and Le in 2019 [12], represents a family of convolutional neural networks optimized for accuracy and efficiency across a range of compute scales. Central to its design is the compound scaling method that uniformly scales network width, depth, and resolution, based on a set of fixed scaling coefficients. When adapted as an encoder within segmentation architectures, such as in U-Net or DeepLab frameworks, EfficientNet is found to provide robust feature extraction capabilities in medical imaging.

2.4. R-CNNs

Girshick *et al.* [6] proposed a method employing high-capacity CNNs for image segmentation tasks. He *et al.* [7] further advanced this in 2017 by extending Faster R-CNN to include a branch that predicts an object mask along with the bounding box recognition.

2.5. Multi-Task Learning

Multi-task learning, where a model simultaneously tackles multiple tasks, enhances efficiency and performance by promoting generalization through shared representations [2]. Novosel *et al.* [10] demonstrates that semantic segmentation improves from using a self-supervised multi-task. Bai *et al.* [1] shows that self-supervised learning through position prediction can boost image segmentation for cardiac MR scans.

3. Dataset and Features

3.1. Raw Dataset Details

The data can be accessed at Kaggle competition UW-Madison GI Tract Image Segmentation's repository[2]. The dataset contains 85 cases (patients), each case has MRI scans on different days (normally 3-5 days), and different slices on each day. The segmentation and class information is shown in the "train.csv" file.

[2] https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation/data

3.2. Data Preprocessing

In preparation for model training and evaluation, we executed a series of data preprocessing measures. We developed a decoder and an encoder to facilitate the conversion between run-length encoding (RLE) representation of segmentation masks and binary mask formats, enabling straightforward manipulation and storage of these masks. Extensive metadata was extracted from MRI scan images, including case identifiers, day of scan, slice number, file path, file name, composite id, image dimensions (height and width), and resolution. The dataset was partitioned into training, validation, and testing subsets, consisting of 65%, 15%, and 20%, respectively, stratified based on the number of organs shown in the image. Furthermore, all data was standardized into the Common Objects in Context (COCO) format, a common framework used to benchmark image segmentation tasks [9]. Additionally, we compiled functions to facilitate data generation, image loading, and annotation processing, thereby optimizing the data pipeline and enhancing the efficacy and reproducibility of our segmentation models.

4. Methods

4.1. Encoder - Decoder Models

Encoder - Decoder Models represents a foundational approach in image semantic segmentation. We will employ a small, pre-trained U-Net model for our baseline due to its classical relevance and proven efficiency in similar tasks. Additionally, a larger U-Net model will enhance the prediction accuracy and handle more complex segmentation tasks effectively.

4.1.1 Simple U-Net Model (Baseline)

The Simple U-Net architecture, a streamlined version of the original U-Net and maintains the core symmetrical encoder-decoder structure, which facilitates effective feature extraction and contextual understanding crucial for detailed analysis of medical images. Figure 2 illustrates the foundational U-Net architecture upon which our model is based.

In refining the original design, the Simple U-Net employs dual-layer convolutional units within both the encoder and decoder paths, each enhanced with instance normalization and Leaky ReLU activation to improve non-linear feature learning and stabilization across network layers. The architecture also incorporates efficient down-sampling via max pooling and sophisticated upsampling

through transpose convolutions, integrated with skip connections to preserve and restore spatial and contextual information lost during feature compression. This design not only ensures the production of high-quality segmentation maps with precise boundary delineations but also optimizes computational efficiency, making it suitable for handling large datasets given the constrained resources. The entire model leverages the kaiming normalization technique for weight initialization, significantly improving training dynamics and convergence speed [8].

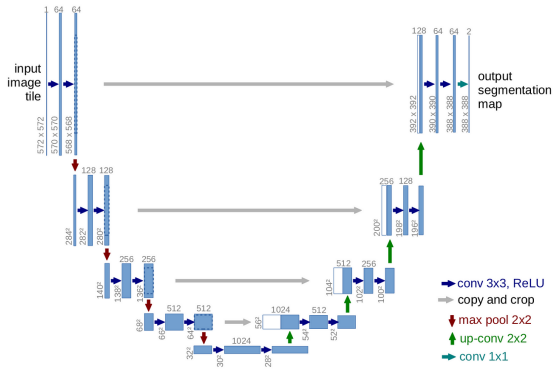

Figure 2. U-Net Architecture from Original Paper

4.1.2 Large U-Net Model (Efficient Net)

In our approach, we incorporate the EfficientNet-B6 as the encoder within a U-Net architecture, leveraging its advanced capabilities for deep feature extraction. Utilizing the U-Net configuration, the model employs efficientnet-b6, pre-trained on the ImageNet dataset [5], ensuring a rich initial feature space conducive to complex segmentation tasks. The network output is activated by a sigmoid function, facilitating the generation of class probabilities suitable for multi-class segmentation.

4.1.3 DeepLabv3+ Model

In this approach, we also utilize the DeepLabV3+ model, integrated with a ResNeXt101_32x8d encoder to harness its sophisticated segmentation capabilities [13]. The model benefits from the ResNeXt architecture, known for its aggregated residual transformations that provide enhanced model capacity and efficiency. This configuration leverages weights pre-trained on the ImageNet dataset, offering a strong foundational understanding of diverse visual features necessary for accurate and detailed segmentation. The output layer utilizes a sigmoid activation function to produce a probability map for each class, enabling precise pixel-wise classification.

4.2. Mask R-CNN Model

Mask R-CNN enhances the capabilities of Faster R-CNN by integrating a branch dedicated to predicting segmentation masks for each Region of Interest (RoI). This branch operates concurrently with the branches for classification and bounding box regression. Mask R-CNN's dual functionality ensures high accuracy and maintains processing speed, making it an ideal choice for our project.

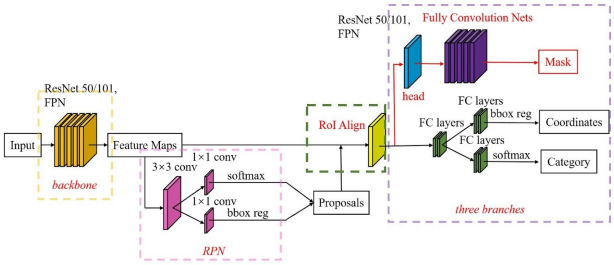

Figure 3. Mask R-CNN Architecture

4.3. Multi-task Learning: Position Learning

To enhance the baseline model, we integrated multi-task learning by leveraging positional learning alongside U-Net architecture. This is designed to extract richer contextual information from the images, specifically the slice numbers of the MRI scans.

Multi-task learning entails performing two tasks simultaneously. In this case: (1) semantic segmentation of medical images to identify the three organs of interest (stomach, large bowel, and small bowel) through U-Net, and (2) regression of the position of each cross-sectional scan within the image. This joint task allows the model to capture both fine-grained spatial details and global positional information from the input images. As shown in Figure 4, the U-Net takes the image and shares the encoder weights with the two task-specific decoders.

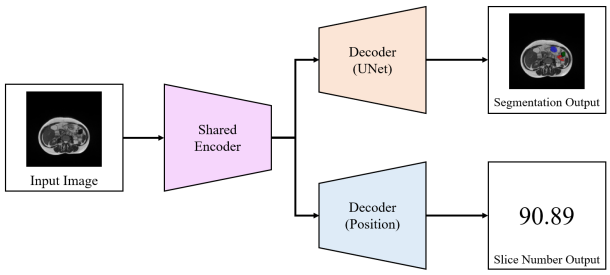

Figure 4. Multi-Task Learning with Position Architecture

Given that the cross-sectional scans in the medical images data vary in position along the axis, we encode the positional information of each scan using positional learning. The loss function for the regression task is Mean-Squared

Error as the loss is non-negative and sensitive to large numbers. It is defined below where S_i is the true value and $\hat{S}_i$ is the predicted value:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(S_i - \hat{S}_i\right)^2$$

This is used with a weighted loss function that reflects a combined loss functions of the two tasks.

5. Experiments

5.1. Evaluation Metric

5.1.1 Dice Loss

The Dice loss function is universally applied across all of our methods to evaluate the similarity between two binary masks. Consider two binary masks, M_1 and M_2, which correspond to the set of pixels each mask covers. The Dice coefficient for these masks is calculated as follows:

$$C = \frac{2 \times |M_1 \cap M_2|}{|M_1| + |M_2|}$$

where $|\cdot|$, denotes the number of pixels in the set. The coefficient C ranges between 0 and 1, where a Dice coefficient of 1 indicates perfect overlap between the masks and 0 indicates no overlap.

For M_1 as the predicted mask and M_2 as the true mask, the Dice loss is defined as:

$$L_{\text{Dice}} = 1 - C$$

Our objective is to minimize this Dice loss, L_{Dice}, to improve the accuracy of the predicted mask in matching the true mask.

5.1.2 Intersection over Union (IoU)

Intersection over Union (IoU) is a metric used to evaluate the accuracy of an object detector on a particular dataset.

$$IoU = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|} = \frac{|M_1 \cap M_2|}{|M_1| + |M_2| - |M_1 \cap M_2|}$$

IoU provides a value between 0 and 1, where 0 means no overlap and 1 means perfect overlap. A higher IoU score indicates a more accurate model. In practice, a threshold (like 0.5) is often used to decide whether predictions are correct.

For evaluating models, IoU is favored over Dice Loss. Since Mask R-CNN is not fit for Dice Loss, so it will naturally have higher loss and lead to unfair comparisons between models.

5.2. Experiment Setup

To ensure the consistency and repeatability of our experiments, we run them all on the Google Colab platform with an A100 GPU. We set up our experiment for each model with the following combination of hyper-parameters. We also show each model's running time and the number of epochs we choose to terminate.

- **Simple U-Net — Efficient Net — DeepLabV3+** : {Batch Size: 32, Image Resize: 128 X 128, Learning Rate: 0.001, Learning Rate factor: 0.2}, takes slightly under 2 minutes to run 1 epoch, terminate at epoch 9.
- **Mask R-CNN**: {Batch Size: 2, Image Resize: 512 X 512, ResNet Mean: (0.485, 0.456, 0.406), ResNet SD: (0.229, 0.224, 0.225), Momentum: 0.9, Learning Rate: 0.001, Weight Decay: 0.0005}, takes 15 minutes to run 1 epoch, terminate at epoch 9.
- **Multi-task U-Net**: {Batch Size: 32, Image Resize: 128 X 128, Learning Rate: 0.001, Learning Rate factor: 0.2}, takes slightly under two minutes to run 1 epoch, terminate at epoch 5.

5.3. Results and Analysis

Table 1 presents the evaluation metrics of five models based on their performance of the test dataset., indicating that the baseline Simple U-Net model achieves the lowest Dice Loss and records satisfactory IoU scores across all three organs. The EfficientNet model demonstrates superior performance in segmenting the stomach and large bowels, whereas Mask R-CNN significantly outperforms the other models in terms of the IoU score for the small bowel. The implementation of multi-task learning appears to improve the IoU score only for the small bowel, which may be attributed to the low prediction accuracy of the auxiliary task. The efficient net may inadequately preserve fine-grained spatial information, exacerbating the challenge of accurately delineating the fragmented segments of small bowel. As a result, this limitation might contribute to blurry or imprecise predictions for its boundaries, ultimately resulting in lower IoU score. In summary, considering all performance metrics, Mask R-CNN emerges as the most balanced model, exhibiting high performance across each evaluated metric.

Table 1. Model results

Model	DiceLoss	IoU-S	IoU-LB	IoU-SB
Simple U-Net	**0.125**	0.506	0.578	0.259
Efficient Net	0.295	**0.728**	**0.733**	0.000
DeepLabv3+	0.337	0.609	0.606	0.372
Mask R-CNN	0.246	0.654	0.611	**0.530**
Multi-task U-Net	0.194	0.180	0.314	0.355

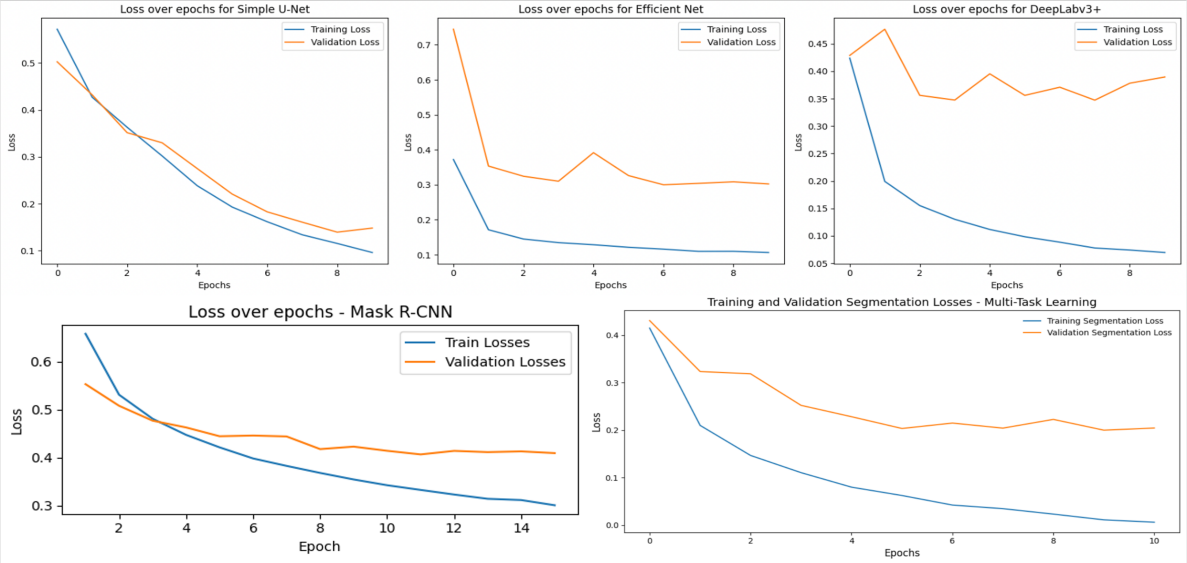

Figure 5. Loss vs Epochs for Segmentation Models

5.4. Discussions and Error Analysis

Training and loss analysis of segmentation models, including Simple U-Net, Efficient Net, and DeepLabv3+, reveals insights into their performance dynamics as depicted in Figure 5. For Simple U-Net, the training loss steadily decreases over epochs, indicating effective optimization and learning of features. However, the validation loss tends to fluctuate, suggesting potential overfitting or instability in performance on unseen data. In contrast, Efficient Net demonstrates a consistent decrease in both training and validation losses, indicative of robust optimization and generalization. DeepLabv3+, while exhibiting a decreasing trend in both losses, shows higher fluctuations in validation loss, potentially indicating sensitivity to variations in the validation dataset.

Mask R-CNN model uses its own loss function, which is a combination of *Classification Loss*, *Bounding Box Regression Loss*, and *Mask Loss*, rather than *Dice Loss*. Therefore, its training loss value cannot be compared with other models. Mask R-CNN graph in Figure 5 shows that the training loss is at first higher than validation loss but is reducing at a fast speed and is less than validation loss after the third epoch and continues to go down. The validation loss, however, is relatively stable after epoch 9. Thus, we choose epoch 9's model as the final model.

Implementing multi-task learning has a negative effect on the baseline, significantly decreasing IoU metrics. Predicting the slice number proved to be difficult, and the relatively high MSE seems to have hurt the model more than aid it. Also, this additional task for positional learning may encourage the model to focus on global features of the image and therefore not put as much emphasis on the fine-grained details that are important for accurate segmentation. To compare loss, the visualization below just highlights dice loss. Note as with other loss metrics, the validation loss plateaus after epoch 5, so epoch 5 is chosen as the final model.

6. Conclusion and Future Work

While advanced models and multi-task learning did not consistently outperform the baseline, some showed notable improvements in specific metrics, leading to slight enhancements in test set performance for certain classes. To enhance our techniques further, we propose three key directions for future refinement: enhancing data augmentation to bolster model robustness, implementing ensemble methods to leverage the strengths of individual models, and focusing on improving the accuracy of position embedding in the multi-task U-Net for potential performance gains.

7. Contributions and Acknowledgements

The team collaborated closely, starting by collectively choosing a topic and sourcing a dataset from the internet. Srikaran handled data cleaning, preprocessing, and transformation, while Zeki and Joe conducted a literature review to identify suitable models and gather implementation resources for semantic segmentation. Srikaran implemented a simple U-Net model as a baseline and enhanced it with Efficient U-Net and DeepLabv3+ models. Zeki worked on the Mask R-CNN model, model visualization, and metrics alignment. Joe integrated multi-task learning into the basic models and created the presentation slides. All three members collaborated to complete the final report.

References

[1] Wenjia Bai and Dinggang Shen. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, Cham, 2019. Springer. 2

[2] Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. 2

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 2

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014. 2

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2

[10] Jelena Novosel, Prashanth Viswanath, and Bruno Arsenali. Boosting semantic segmentation with multi-task self-supervised learning for autonomous driving applications. In *Proc. of NeurIPS-Workshops*, volume 3, 2019. 2

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[12] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6105–6114, 2019. 2

[13] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

Appendices

Figure 6 shows the Dice Loss of both training set and validation set for five models over training epochs.

Figure 7 shows Intersection over Union (IoU) Score of all three organs based on test set data for five models.

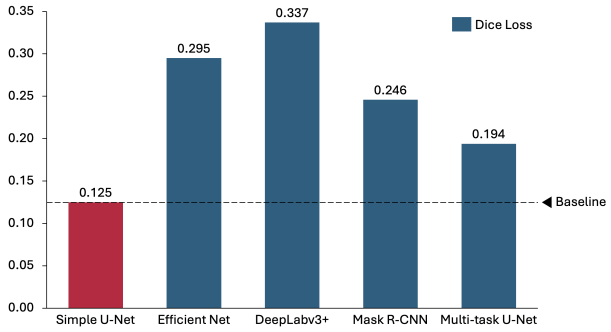

Figure 6. Dice loss of models

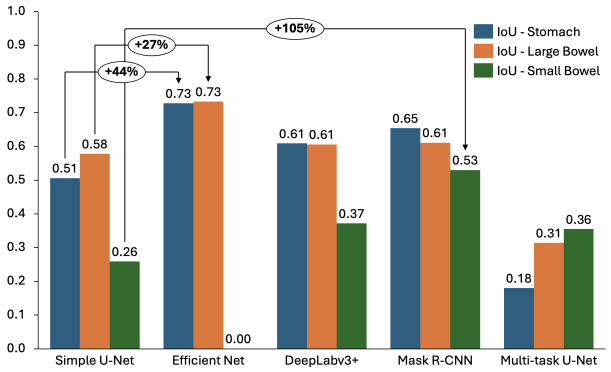

Figure 7. IoU Score of models

Figure 8 shows the visualization of prediction for five prediction models based on randomly selected four slice examples (2 from training set and 2 from test set).

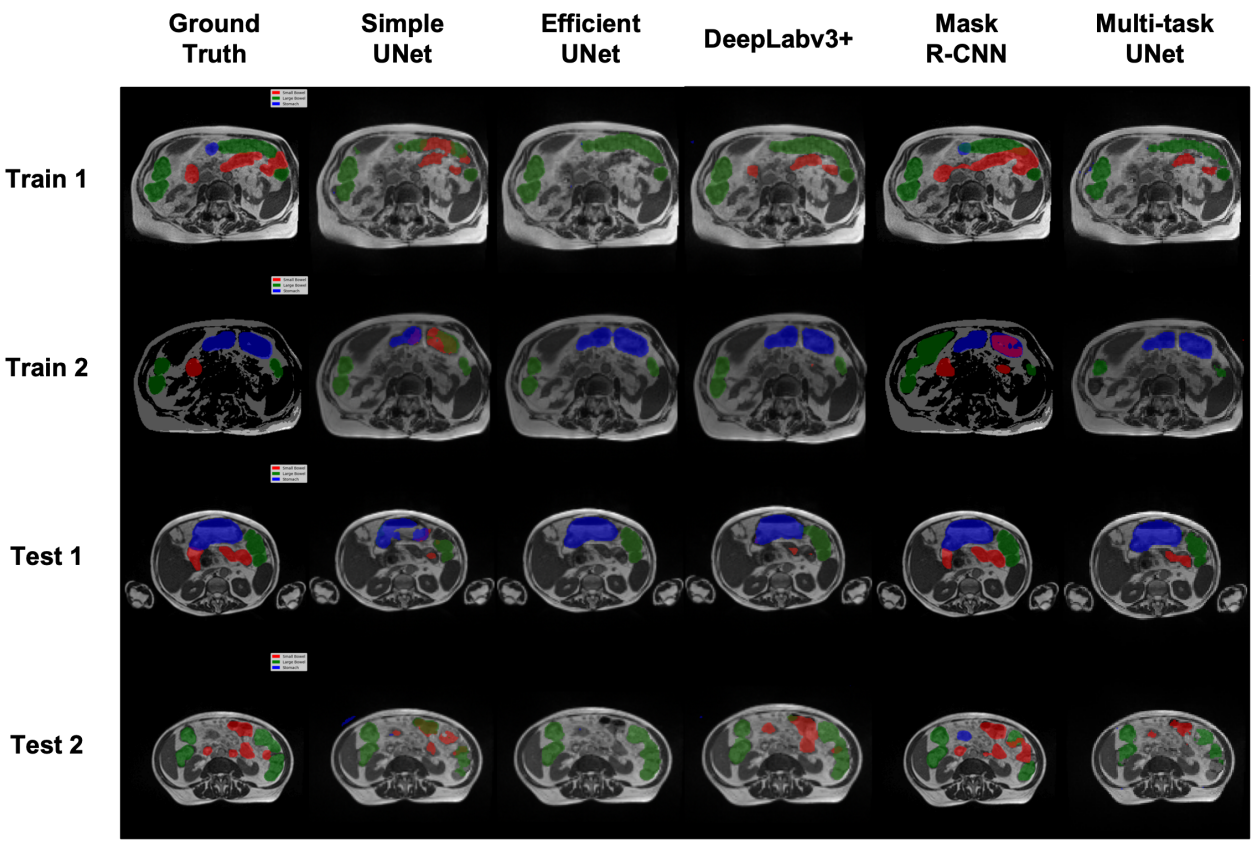

Figure 8. Visualization of Predictions for Segmentation Models (**Red**: **Small Bowel**, **Green**: **Large Bowel**, **Blue**: **Stomach**)