OrganSegNet: An Attention-Enhanced Deep Learning Framework for Multi-Organ Semantic Segmentation in CT Images

Amine Banani, NYCU student. Medical Image Processing

***Abstract*—Accurate semantic segmentation of medical images is a critical task in computer-aided diagnosis and treatment planning. In this work, we propose *OrganSegNet*, a novel deep learning framework incorporating attention mechanisms for multi-organ semantic segmentation of CT images. The dataset comprises 892 training and 143 unlabeled testing samples, with segmentation targets including the liver, kidney, spleen, and pancreas. Our architecture leverages encoder-decoder structures augmented with attention blocks to refine feature fusion and enhance organ delineation. To benchmark performance, we compare *OrganSegNet* against DeepLabV3, a state-of-the-art (SOTA) segmentation model.**

Quantitative evaluation is based on recall, precision, and F1-score for each organ. Results demonstrate that *OrganSegNet* achieves competitive performance, particularly in handling small and challenging. These findings suggest that the proposed architecture offers a robust alternative for multi-organ segmentation tasks in medical imaging.

***Index Terms*— Attention mechanisms, convolutional neural networks (CNNs), CT image segmentation, deep learning, multi-organ segmentation.**

I. INTRODUCTION

Medical imaging plays a pivotal role in modern healthcare, aiding in diagnosis, treatment planning, and disease monitoring. Semantic segmentation, which involves pixel-level classification of images, is essential for extracting clinically relevant information from medical scans. In the context of abdominal CT imaging, accurate segmentation of organs like the liver, kidney, spleen, and pancreas is particularly important due to their involvement in critical diseases.

Despite its importance, semantic segmentation of CT images remains challenging due to the variability in organ shapes and sizes, the close proximity of different organs, and the overlapping intensity values in scans. These challenges necessitate robust and precise algorithms to achieve accurate segmentation.

In recent years, deep learning methods, particularly convolutional neural networks (CNNs), have demonstrated remarkable performance in medical image segmentation. Models such as U-Net and DeepLabV3 have become state-of-the-art (SOTA) for their ability to capture both global context and fine-grained details. However, these methods often face limitations when dealing with small organs or complex anatomical boundaries, motivating the development of more specialized approaches.

To address these challenges [4], we propose OrganSegNet, a novel deep learning framework designed for multi-organ segmentation in CT images. The architecture builds on an encoder-decoder structure and integrates attention mechanisms to enhance feature fusion and refine organ boundaries. This design enables the model to focus on the most relevant regions of the image, improving segmentation accuracy.

The contributions of this work are threefold:

1. We introduce OrganSegNet, a new architecture incorporating attention blocks for enhanced feature fusion and precise segmentation.
2. We perform a comprehensive evaluation of OrganSegNet and compare its performance to DeepLabV3, a SOTA segmentation model, using recall, precision, and F1-score as metrics.
3. We apply the proposed framework to multi-category segmentation of abdominal CT images, targeting the liver, kidney, spleen, and pancreas.

II. METHODS

A. State-of-the-Art Model: DeepLabV3

DeepLabV3 [1] is a well-established deep learning model for semantic segmentation that employs atrous (dilated) convolutions to capture multi-scale contextual information. Its Atrous Spatial Pyramid Pooling (ASPP) module is designed to handle objects at various scales by applying parallel atrous convolutions with different rates. Additionally, DeepLabV3

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, “This work was supported in part by the U.S. Department of Commerce under Grant BS123456”.

The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

integrates a pre-trained backbone network, such as ResNet, to enhance feature extraction.

For this task, we used DeepLabV3 with a ResNet-50 backbone. The output feature maps from the ASPP module were processed through a final convolutional layer to produce pixel-wise class predictions. The model was fine-tuned on the training dataset of CT images for five categories: Background, Liver, Kidney, Spleen, and Pancreas.

The loss function used for training was the categorical cross-entropy loss, which measures the discrepancy between predicted and true class distributions. Training was performed using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 8.

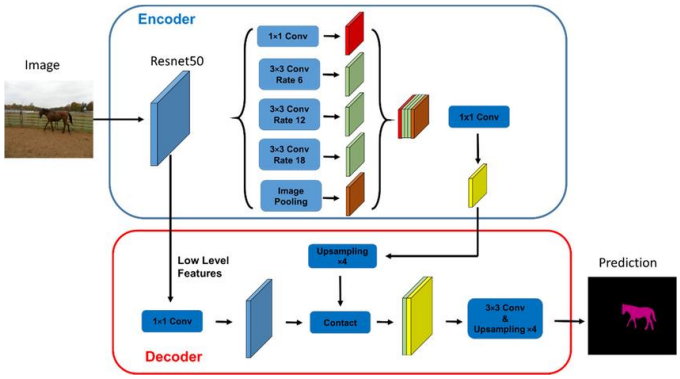

Fig. 1. The network structure combining Deeplabv3 and ResNet50.

B. Proposed Model: OrganSegNet:

The OrganSegNet architecture builds upon the encoder-decoder design, incorporating attention blocks to enhance feature fusion and segmentation accuracy. Below, we describe the components of the architecture in detail:

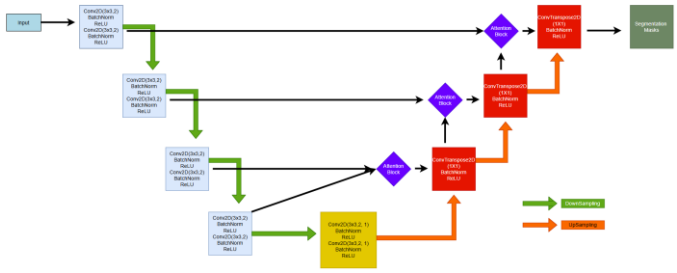

Fig. 1. OrganSegNet architecture

1) Encoder:

The encoder [2] consists of four convolutional blocks, each doubling the number of feature channels. Each block applies two convolutional layers, followed by batch normalization and a ReLU activation function:

$$y = ReLU(BatchNorm(W * x + b))$$

where W and b are the weights and bias of the convolutional filter, x is the input feature map, and $*$ denotes the convolution operation. Downsampling is achieved using max-pooling layers:

$$y_{\{downsampled\}} = max_{\{window\ size\}}(y)$$

2) Bottleneck:

To capture features with a broader receptive field, the bottleneck employs dilated convolutions. The dilation rate d controls the spacing between kernel elements, enabling the model to incorporate multi-scale context:

$$y = \sum_{i=1}^{k} W_i \cdot x_{i+d} + b$$

where k is the kernel size, and $d > 1$ increases the receptive field without additional computational cost. This ensures detailed context preservation critical for accurate segmentation.

3) Attention Mechanisms:

Attention blocks [4][5] enhance feature fusion by focusing on relevant spatial regions in the feature maps. Each attention block computes a spatial weight map ψ_f using gating signals g and input feature maps x:

$$\theta_x = W_\theta * x, \quad \phi_g = W_\phi * g,$$
$$f = Relu(\theta_x + \phi_g), \quad \psi_f = Sigmoid(W_\psi * f)$$
$$y_{attention} = x.\psi_f,$$

Where W_θ, W_ϕ and W_ψ are learnable parameters of the attention block. This mechanism allows the decoder to emphasize regions most relevant for organ segmentation. A diagram illustrating the attention mechanism is provided in **Fig. 2.**

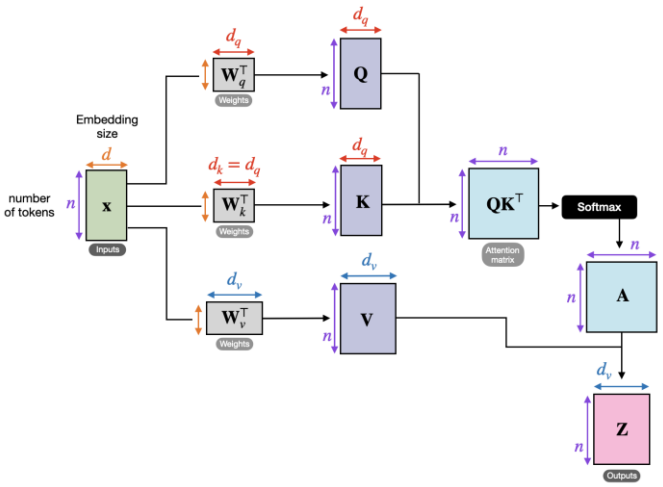

Fig. 3. Visualization of self attention mechanism.

4) Decoder:

The decoder reconstructs high-resolution segmentation maps by combining features from the encoder and the bottleneck through skip connections. Upsampling is performed using transposed convolutions:

$$y_{upsampled} = W_{transpose} * x,$$

where $W_{transpose}$ represents the transposed convolution filter. Concatenation layers integrate information from encoder features, ensuring both global and local context are retained.

5) Final layer:

The final layer applies a 1×1 convolution to reduce the number of feature channels to five, corresponding to the segmentation categories (Background, Liver, Kidney, Spleen, and Pancreas):

$$y_{final} = W_{final} * x + b$$

III. Experiments

1) Dataset Description:

The dataset used for this study consists of **892 labeled training images** and **143 unlabeled testing images**, with pixel-level annotations for five categories: Background, Liver, Kidney, Spleen, and Pancreas. Each image has a resolution of 224×224, which ensures a manageable input size for deep learning models while retaining sufficient anatomical detail.

To standardize the input data and improve model performance, all images were normalized using the following parameters:

- **Mean**: 0.485
- **Variance**: 0.225

This normalization ensures that the pixel intensity values are scaled consistently, making training more stable and effective.

2) Evaluation Metrics:

The performance of both *OrganSegNet* and DeepLabV3 was assessed using the following metrics, computed for each organ (excluding the background):

- **Recall** (R): Measures the model's ability to identify all pixels belonging to a target organ.

$$R = \frac{TP}{TP + FN}$$

where TP is the number of true positive pixels, and FN is the number of false negative pixels.

- **Precision** (P): Assesses the proportion of correctly identified organ pixels relative to all pixels predicted as belonging to that organ.

$$P = \frac{TP}{TP + FP}$$

where FP is the number of false positive pixels.

- **F1-Score** ($F1$): The harmonic mean of recall and precision, providing a balanced measure of the model's segmentation performance.

$$F_1 = 2.\frac{P\,.R}{P + R}$$

These metrics ensure a comprehensive evaluation of segmentation accuracy for each organ category. Pixels corresponding to each organ were treated as positive samples, while all other pixels were considered negative samples.

3) Performance Comparison:

The quantitative results for *OrganSegNet* and DeepLabV3 are summarized in **Table I**. The table includes recall, precision, and F1-scores for each organ

Organ	Model	Precision (P)	Recall (R)	F1-Score (F_1)
Liver	DeepLabV3	0.979213	0.983409	0.981286
	OrganSegNet	**0.998897**	**0.998488**	**0.998692**
Kidney	DeepLabV3	0.945063	0.965652	0.955228
	OrganSegNet	**0.986330**	**0.990620**	**0.988471**
Spleen	DeepLabV3	0.891595	0.894262	0.892877
	OrganSegNet	**0.975454**	**0.982471**	**0.978950**
Pancreas	DeepLabV3	0.857226	0.888000	0.870641
	OrganSegNet	**0.972135**	**0.990156**	**0.981063**

Table. 1. Performance metrics for OrganSegNet and DeepLabV3.

4) Observations:

OrganSegNet outperformed DeepLabV3 across all organ categories in precision, recall, and F1-score, demonstrating its robustness and effectiveness in multi-organ segmentation tasks.

Liver:

- Both models achieved high scores due to the liver's large size and clear boundaries.
- *OrganSegNet* showed slight improvements, with an F1-score of 0.9987 compared to DeepLabV3's 0.9813.

Kidney:

- *OrganSegNet* significantly improved precision

(0.9863 vs. 0.9451) and F1-score (0.9885 vs. 0.9552), likely due to its attention mechanism, which enhances boundary delineation.

Spleen:

- The spleen, which is moderately sized and has variable boundaries, saw a noticeable improvement with *OrganSegNet* (F1-score: 0.9790 vs. 0.8929).

Pancreas:

- The most challenging organ, the pancreas, demonstrated the largest gain.
- *OrganSegNet* achieved a significant F1-score improvement (0.9811 vs. 0.8706), owing to its ability to focus on small and irregular features through the attention mechanism.

Precision vs. Recall:

- *OrganSegNet* consistently maintained a better balance between precision and recall, resulting in higher F1-scores.
- The attention blocks likely contributed to fewer false positives (improving precision) and better identification of true positives (enhancing recall).

Conclusion:

These results highlight the superiority of *OrganSegNet* over DeepLabV3, particularly for smaller and more complex organs like the pancreas and spleen. The integration of attention mechanisms is key to achieving these improvements.

5) Qualitative analysis:

To visually assess the performance of the two models, we present segmentation outputs for five test images from the dataset. **Figure 3** shows results produced by *OrganSegNet*, and **Figure 4** shows results from DeepLabV3.

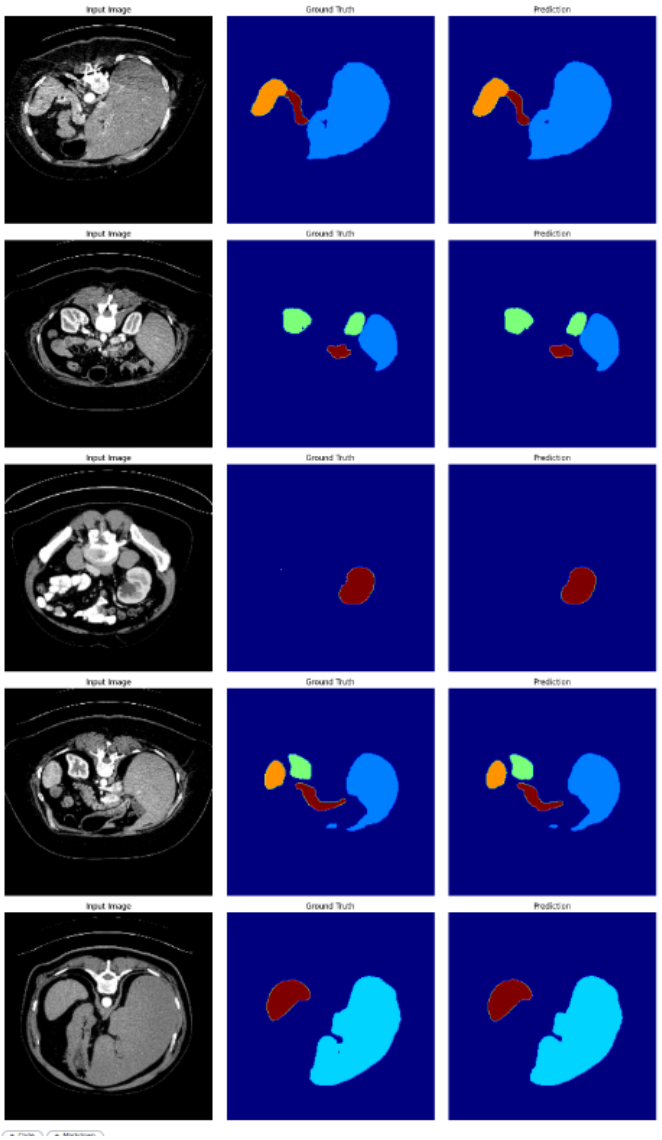

Fig. 4. Example segmentation results from *OrganSegNet* on five test images. (middle: Ground truth) (right: Prediction)

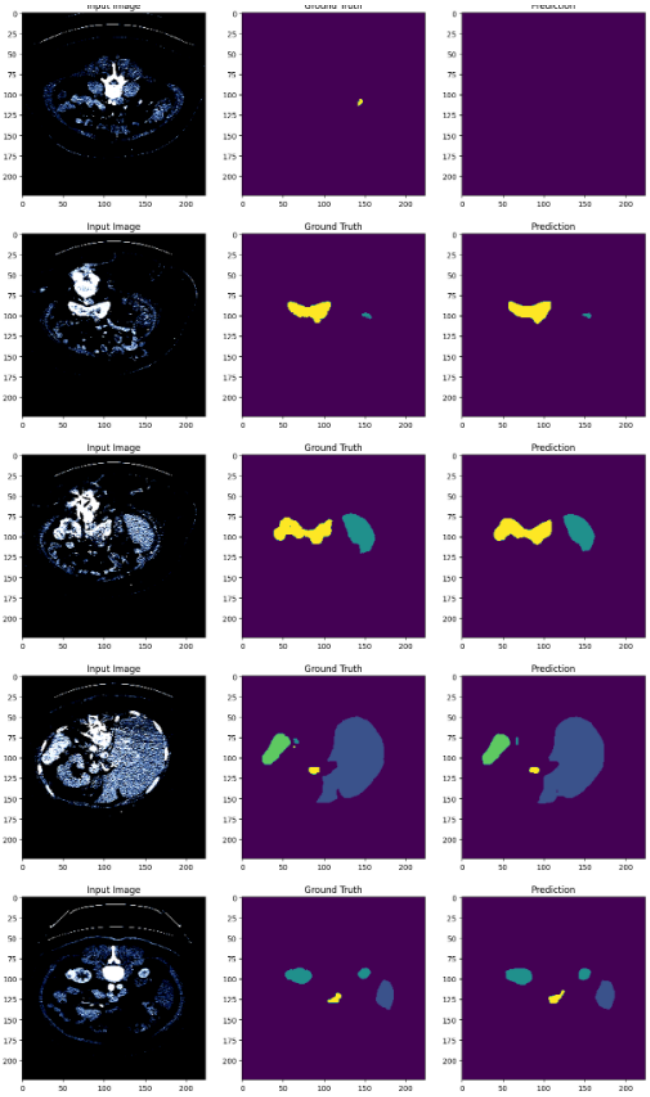

Fig. 6. Example segmentation results from *DeepLabV3* on five test images. (middle: Ground truth) (right: Prediction)

IV. Discussion

A. Model Strengths

The results demonstrate that *OrganSegNet* consistently outperforms DeepLabV3 across all metrics, particularly for smaller and more challenging organs like the pancreas and spleen. This improvement can be attributed to the integration of attention blocks, which enhance feature fusion and enable the model to focus on relevant regions within the images.

The use of an encoder-decoder structure with skip connections further aids in preserving spatial details, ensuring accurate segmentation boundaries. Additionally, the attention mechanism effectively mitigates the impact of overlapping organ intensities and complex shapes, providing a robust solution for multi-organ segmentation.

B. Comparison with SOTA

While DeepLabV3 is a strong baseline, it lacks specialized mechanisms like attention blocks, which limits its ability to handle small or overlapping anatomical structures. *OrganSegNet* addresses these limitations, achieving higher precision and recall without significantly increasing computational complexity.

However, DeepLabV3's performance on larger organs such as the liver remains competitive, indicating that its ASPP module effectively captures large-scale contextual features.

C. Implications for Medical Imaging

The enhanced segmentation performance of *OrganSegNet* suggests its potential for real-world applications in medical imaging. Accurate multi-organ segmentation can facilitate tasks such as organ volume estimation, disease localization, and preoperative planning. Moreover, the model's generalizability makes it a promising candidate for other medical imaging modalities.

V. Conclusion

In this study, we proposed *OrganSegNet*, a novel deep learning framework for multi-organ semantic segmentation in CT images. The architecture incorporates attention blocks within an encoder-decoder structure, enabling effective feature fusion and accurate segmentation of challenging anatomical regions.

Through quantitative and qualitative evaluation, *OrganSegNet* demonstrated superior performance compared to DeepLabV3, achieving higher precision, recall, and F1-scores across all organ categories. The integration of attention mechanisms proved particularly beneficial for segmenting smaller and more complex organs such as the pancreas and spleen.

The findings highlight the potential of *OrganSegNet* to enhance clinical applications, including organ volume estimation and disease localization. Future work will focus on addressing the model's limitations, such as sensitivity to noisy data, and exploring its generalizability to other imaging modalities.

References

[1] [1] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[2] [2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention (MICCAI)*, 2015, pp. 234–241.

[3] [3] O. Oktay, J. Schlemper, L. L. Folgoc, et al., “Attention U-Net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
[4] [4] G. Litjens, T. Kooi, B. E. Bejnordi, et al., “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
[5] [5] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.