Abdominal Medical Image Segmentation with 2D U-Net

1st Mengyao Zhang
School of Physics
Peking University
Beijing, China
mz110@rice.edu

Abstract—Abdominal organ segmentation in CT scans is crucial for clinical diagnosis and treatment planning. While recent advances mostly focus on 3D models and large-scale architectures, these approaches demand substantial computational resources and extensive training data. This work demonstrates that 2D U-Net, a lightweight and efficient architecture, can achieve competitive performance on abdominal organ segmentation tasks. We train our model on a mixed dataset of 113 CT scans from two sources (AbdomenCT-1K and Peking University Third Hospital), implementing careful data preprocessing including HU value clipping and label rearrangement across 12 organ classes. To address class imbalance and missing annotations, we design a weighted cross-entropy loss function that combines class-specific weights with distance-based pixel weights emphasizing organ boundaries, thereby avoiding false negative and false positive errors. Our model achieves an average Dice score of 0.917 on the validation set and comparable performance to state-of-the-art 3D methods on fully annotated organs, while requiring only one hour of training on an A100 GPU and less than one second for inference per CT scan. Although the model faces challenges with missing annotations when generalizing across datasets, it demonstrates the significant potential of 2D U-Net for efficient medical image segmentation in resource-constrained scenarios.

Index Terms—Medical Image Segmentation, U-Net, Deep Learning

I. Background and Motivation

A. Medical Image Segmentation

Medical image segmentation is a crucial task in the field of medical imaging and computer vision (Fig. 1). It involves the process of partitioning medical images into meaningful regions or segments, typically to identify and delineate anatomical structures, tissues, or pathological areas. Accurate and fast segmentation is essential for various clinical applications, including diagnosis, treatment planning, and monitoring of diseases. Among many types of medical images, Computed Tomography(CT) scans are widely used for abdominal organ analysis due to their high resolution and ability to capture detailed anatomical information.

Recently, many deep learning-based methods have been proposed for medical image segmentation, among which many are based on 3D-space segmentation (V-net [1], SegFormer3D [2], SAM-Med3D [3], SegVol [4], etc.), training on the full 3D CT scans with 3D labels. In addition, models based on popular large models like SAM and CLIP are also used for medical image segmentation (VCLIPSeg [5] and SegSAM [6]), usually combined with the former 3D segmentation models.

However, few works focus on 2D image segmentation of abdominal organs in CT slices, which is more lightweight, efficient and simple. The most promising model for 2D medical image segmentation is U-Net [7], which has been widely used in various image segmentation tasks due to its effectiveness and efficiency.

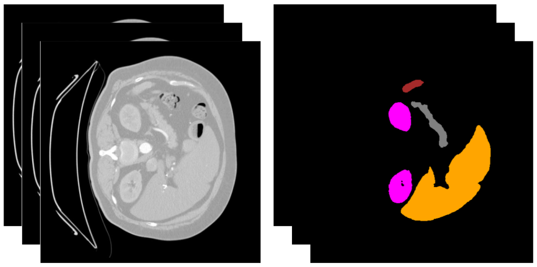


Fig. 1: Demonstration of Medical Image Segmentation. Using data from [8].

B. Motivation for This Work

For 3D models and large models on abdominal organ segmentation, the computational cost is high [9] and the training data requirement is strict (for example, data in separate slices rather than continuum volumes may affect the 3D image processing). Moreover, it has no advantage if we are only interested in segmenting certain slices of a CT scan.

Although sacrificing some generality, 2D U-Net models can be more lightweight and efficient, making them suitable for real-time applications and deployment on resource-constrained devices. Some previous works even show that 2D U-Net outperforms its 3D counterpart on volumetric medical image segmentation task [10]. The comparison between 3D models and 2D U-Net is summarized in Table I.

Therefore, here we try to implement a model based on 2D U-Net, which is simple, easy to train, and fast, while still achieving satisfactory accuracy on abdominal organ segmentation in CT scans. The previous few attempts on

2D U-Net are not extensive and we hope to demonstrate the potential of 2D U-Net in this task.

	Complexity	Accuracy	Training Data Requirement	Generality
3D Models	* *	* *	* *	* *
2D U-Net	*	* *	*	*

TABLE I: Qualitative Comparison Between 3D Models and 2D U-Net on Abdominal Organ Segmentation.

C. Dataset Description

We mix two datasets here for training and evaluation. One is a public dataset called AbdomenCT-1K dataset [8], containing thousands of CT scans with corresponding masks, among which only 41 scans are used for training here. There are four annotated organs: liver(1), kidney(2), spleen(3), pancreas(4), the integer number being the value in the mask files. The other is a private dataset from Peking University Third Hospital(PUTH), containing 72 CT scans with corresponding masks. There are eleven annotated organs: bladder(1), colon(2), left femur head(3), right femur head(4), left kidney(5), right kidney(6), liver(7), rectum(8), smallintestine(9), SpinalCord(10), Stomach(11). A detailed understanding of the two datasets is given in Section II.

CT scans from both datasets are of size (512, 512, N), where N is the number of slices in the scan, varying from different scans.

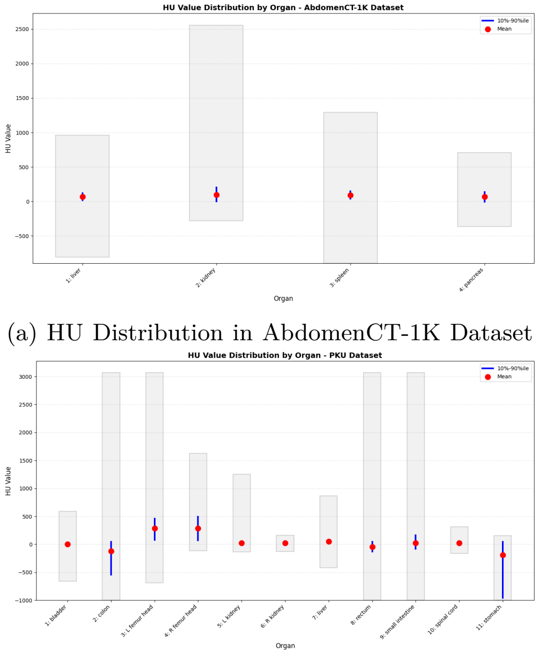
II. Experimentation

A. Data Preprocessing

1) Clipping of HU values: CT scans are stored as .nii.gz files, with pixel values being Hounsfield Units(HU). Different objects show different HU in CT images. The HU values in a CT scan typically range from -1000 to 3000, but most of the values relevant to abdominal organs fall within a narrower range, as can be seen from Fig. 2. Therefore, we clip the HU values to a certain range to reduce the influence of irrelevant tissues and noise, which also allows the model to focus on a more detailed range.

2) Label Rearrangement: Since the two datasets have different organ labels, we rearrange the labels to make them consistent, as shown in Table II.

3) Contrast, Brightness and Rotation: To increase the diversity of training data, we apply random contrast adjustment, brightness adjustment and 90, 180, 270 degree rotations to the training images. Because the physical volume of voxels vary between different CT scans, which means the same organ may appear in different sizes and shapes in different scans, we do not need to explicitly apply scaling or elastic deformation augmentation.



(a) HU Distribution in AbdomenCT-1K Dataset

(b) HU Distribution in PUTH Dataset

Fig. 2: HU Value Distribution in Two Datasets. The red point is average value. The blue line is the 10% and 90% percentile. The gray bar shows the range of HU values(from minimum to maximum).

B. Model Architecture

Based on the original U-Net architecture (Fig. 3a), some adaptations are made to better fit our task (Fig. 3b):

- The output channel is changed to 12, corresponding to the 12 labels after rearrangement.
- Batch Normalization layers are added after each convolutional layer to stabilize training, whose effects are shown in Fig. 4a and Fig. 4b.
- The input image size is (256, 256) to reduce computational cost.

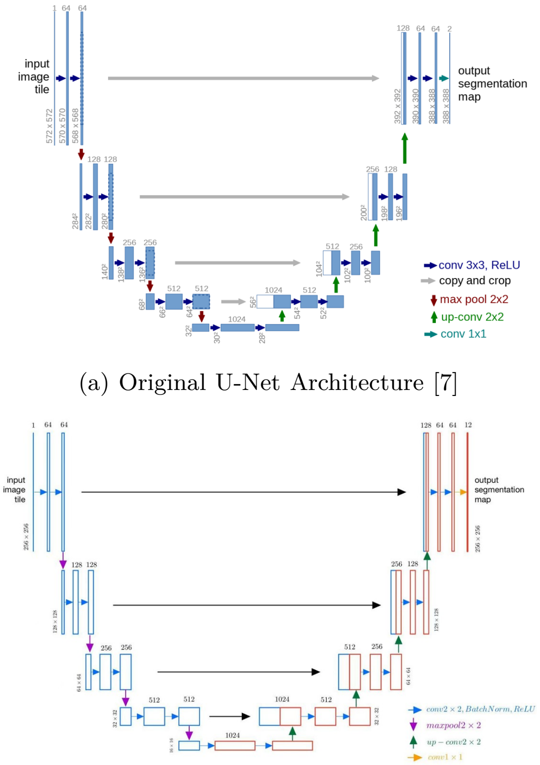
C. Loss Function and Metric

1) Loss Function: To introduce the loss function, we first analyze the possible errors in abdominal organ segmentation. There are several error sources and types interwoven in this task:

- Class imbalance: background pixels dominate the image, while organ pixels are sparse. And some organs occupy even smaller areas than others.
- Missing annotation: some organs are not annotated in certain slices. (For example, in AbdomenCT-1K dataset, only four organs are annotated, while other organs are not labeled at all!)
- False negative: missing or wrong organ pixels, especially on organ boundaries.
- False positive: misclassifying background pixels as organ pixels.

Label	0	1	2	3	4	5	6	7	8	9	10	11
AbdomenCT-1K	BG	-	-	-	kidney	liver	-	-	-	-	spleen	pancreas
PUTH	BG	bladder	colon	femur head	kidney	liver	rectum	small int.	spinal cord	stomach	-	-

TABLE II: Label Rearrangement for Two Datasets. BG stands for background.



(a) Original U-Net Architecture [7]

(b) Modified U-Net Architecture for Abdominal Organ Segmentation

Fig. 3: U-Net Architectures.

The loss function should address these issues properly. Inspired by [7], we designed the Cross-Entropy Loss weighted on both classes and pixel positions. The loss function is defined as:

$$L = \frac{1}{N}\sum_{n,c} -[w_0(c) + w_1(n, y)] \cdot y_c(n) \cdot \log(\frac{\exp(x_c(n))}{\sum_c \exp(x_c(n))}) \tag{1}$$

Here N is the number of pixels in the batch, n is the pixel index. c is the class index, $x_c(n)$ is the output logit for class c at pixel n, $y_c(n)$ is the one-hot encoded ground truth label for class c at pixel n. $w_0(c)$ is the class weight for class c, as a main tool to address the errors mentioned above. The detailed analysis is given in the next paragraph. $w_1(n, y)$ is the pixel weight for pixel n in image y, calculated to emphasize the boundaries between different organs, reducing false negatives and false positives on organ edges.

2) Analysis of the Loss Function: Let's see how the loss function is supposed to address the errors mentioned above.

Initially, we set $w_0(c)$ inversely proportional to the frequency of class c in the training set, in order to address class imbalance. However, we found that this approach overly suppressed the background class, leading to excessive false positives, as shown in Fig. 4c. On the other hand, if we set $w_0(c)$ uniformly, the model tended to miss organ pixels, resulting in false negatives, as shown in Fig. 4d. Therefore, we adjusted $w_0(0)$ empirically to balance false negatives and false positives, while encouraging organ predictions on annotated background, to address missing annotation issues. And $w_0(c), c > 0$ is also empirically adjusted to enhance the accuracy of organs on which the model performs poorly. In conclusion, $w_0(c)$ is the main tool to address the issues mentioned above.

For $w_1(n, y)$, we adjusted the method in [7] to calculate it based on $d(n, y)$: the distance of pixel n to the nearest organ in mask y. The formula being:

$$w_1(n, y) = A \cdot \exp(-\frac{d(n, y)^2}{2\sigma^2}) \tag{2}$$

This way, pixels closer to organs are assigned higher weights, emphasizing organ boundaries more effectively. A comparison is shown in Fig. 5, where A is set to 0.01 and 6 respectively, the latter increasing the Dice score of the example slice from 0.716 to 0.866.

3) Evaluation Metric: For evaluation, we mainly use Dice Score as the metric, defined as:

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{3}$$

where X is the set of predicted organ pixels, and Y is the set of ground truth organ pixels.

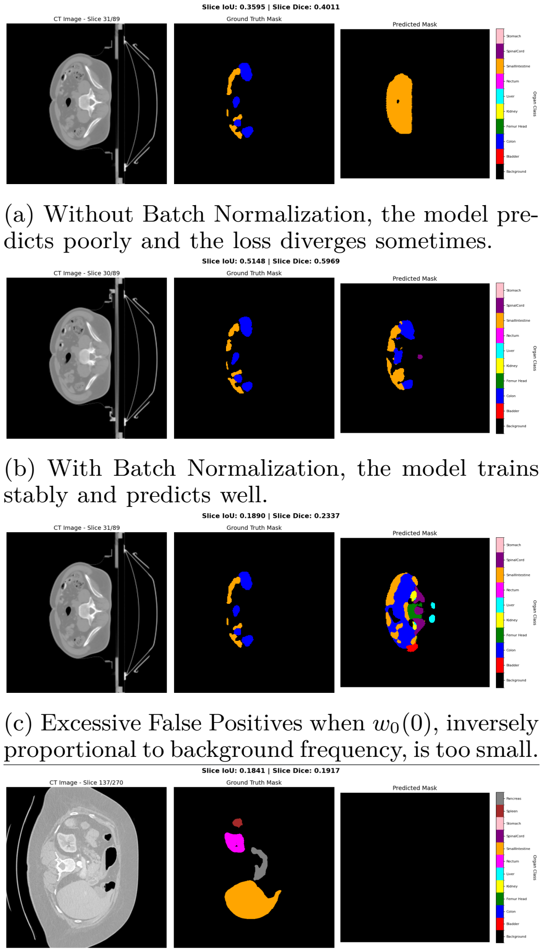
We also include Intersection over Union(IoU) as a secondary metric, defined as:

$$IoU = \frac{|X \cap Y|}{|X \cup Y|} \tag{4}$$

D. Special Techniques

To further save computational resources, we applied the following techniques:

- Mixed Precision Training: Using autocast in PyTorch to automatically choose the appropriate precision for different operations, reducing memory usage and speeding up training.
- Gradient Accumulation: Accumulating gradients over multiple mini-batches before performing a weight update, effectively increasing the batch size without requiring more memory.
- Gradient Checkpointing: Saving memory by re-computing certain intermediate activations during



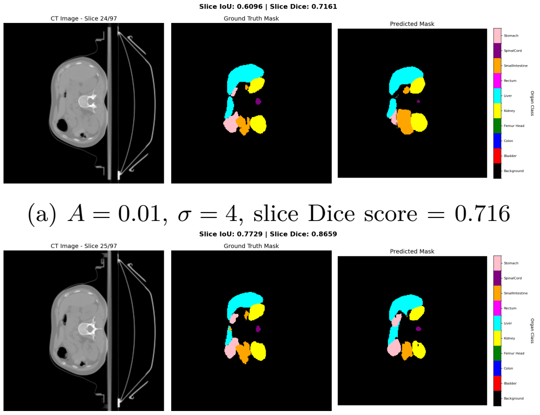
(a) Without Batch Normalization, the model predicts poorly and the loss diverges sometimes.

(b) With Batch Normalization, the model trains stably and predicts well.

(c) Excessive False Positives when $w_0(0)$, inversely proportional to background frequency, is too small.

(d) Excessive False Negatives when $w_0(c)$ is set similarly across all classes.

Fig. 4: Examples of Failure Cases due to Improper Settings.



(a) $A = 0.01$, $\sigma = 4$, slice Dice score = 0.716

(b) $A = 6$, $\sigma = 4$, slice Dice score = 0.866

Fig. 5: Comparison of $w_1(n, y)$ Settings. Considering $w_1(n, y)$ emphasizes organ boundaries more effectively, improving segmentation accuracy.

the backward pass instead of storing them all during the forward pass, which sacrifices some computation time for reduced memory usage. This method is not applied in the final training due to incompatibility with mixed precision training.

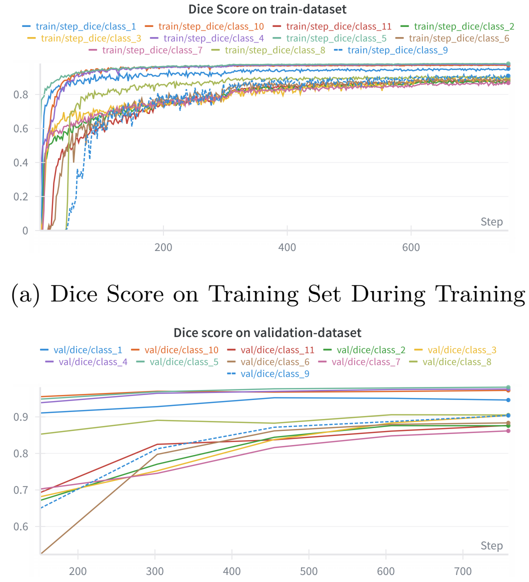
III. Results

A. Training Results

After numerous experiments and hyperparameter tuning(including batch size, learning rate, weight decay, loss weights, model architecture, clipping windows, loss function forms etc.), the final model achieves the Dice score of over 0.85 on all organs and an average Dice score of 0.917 on the validation set, as shown in Tabel III. It also reaches similar accuracy on the test set, reaching the same level as previous works on AbdomenCT-1K dataset for annotated organs(liver, kidney, spleen, pancreas).

Apart from being accurate, the training time is about an hour on A100 GPU, and the inference time per CT scan is within 1 second, which is significantly faster than previous works, fully demonstrating the simplicity and efficiency of 2D U-Net here.

During the training, the Dice score on training and validation datasets is shown in Fig. 6a and Fig. 6b respectively. We observe that the accuracy on certain organs improves abruptly at some training stages, which may be due to the competition of loss contributions from different organs.



(a) Dice Score on Training Set During Training

(b) Dice Score on Validation Set During Training

Fig. 6: Dice Score During Training Process.

B. Challenges and Limitations

In the Table III, it seems that our model perfectly outperforms previous works. However, we should notice

Organ Lable	1	2	3	4	5	6	7	8	9	10	11	Avg.
Ours Val Dice	0.946	0.876	0.904	0.976	0.981	0.883	0.862	0.905	0.904	0.972	0.877	0.917
Ours Val IoU	0.898	0.779	0.825	0.954	0.963	0.791	0.757	0.823	0.824	0.946	0.780	0.849
Others Test Dice	-	-	-	0.849-0.96	0.955-0.982	-	-	-	-	0.926-0.985	0.800-0.901	-
Ours Test Dice	-	-	-	0.956	0.978	-	-	-	-	0.956	0.779	-

TABLE III: Scores on Validation and Test Sets. Others' Dice Scores are from [11]. Here evaluation of our model are only considering annotated organs, so it may overestimate the performance. The label-organ correspondence is shown in Table II.

that the evaluation of our model here only considers annotated organs. For example, when the model inferences CT scans from AbdomenCT-1K dataset, it rarely predicts organs other than liver, kidney, spleen and pancreas (Fig. 7), even though these other organs do exist in the images and are labeled in PUTH dataset. Ideally, after trained on mixed datasets, the model should be able to segment all organs annotated in both datasets, but it fails to do so. This is where the above-mentioned missing annotation issue shows up, which most of the current literatures are trying to solve.

Therefore, although our model achieves high accuracy on annotated organs and is more efficient, implying the potential of 2D U-Net, it sacrifices some generality compared to previous works. Luckily, when the model inferences on test datasets, it is less affected by the missing annotation issue, as shown in Fig. 8.

IV. Discussion and Conclusion

As mentioned above, with proper data preprocessing, model architecture design, loss function formulation, and training techniques, our model based on 2D U-Net is memory-saving, fast, and easy to train, while achieving significantly high accuracy on fully annotated organs during training, validation and testing. This demonstrates the potential of 2D U-Net in abdominal organ segmentation, especially in scenarios where computational resources are limited or real-time processing is required.

However, the model faces challenges in generality due to missing annotations in the training data. In other words, the model is not "bold" enough to predict the existence of organs on pixels which are labeled as background in the training dataset. In fact, our model is less affected by the missing annotation issue when it inferences on test datasets, compared to working on training and validation dataset. An example is shown in Fig. 8, where the a zero-shot CT scan from AbdomenCT-1K is inputted to the trained model, and the model does segment organs other than the four annotated ones, but the segmentation is still not very satisfactory.

The missing annotation issue is a common challenge in medical image segmentation, especially when mixing datasets with different annotation scopes. We'd planned to build a network that separately handles each organ, which may alleviate this issue, but due to time constraints, we leave it for future work.

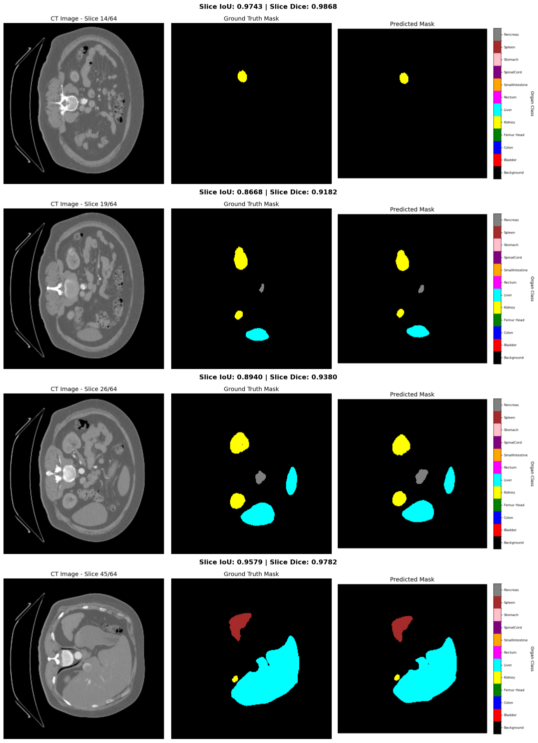


Fig. 7: Great Segmentation on Fully Annotated Organs. Poor segmentation of unannotated organs in AbdomenCT-1K training dataset. The middle column is the mask, where unannotated organs(among bladder, colon, femur head, rectum, small int., spinal cord, and stomach) are labeled as background. These unannotated organs are inputted to the model along slices from PUTH dataset, and should be predicted in slices from AbdomentCT-1K dataset as well, but they are mostly missed.

V. Code Availability

All codes are available at github repository: https://github.com/smilezzm/Abdominal-organ-segmentation

VI. Group Members Roles

All work is done by the author Mengyao Zhang.

(As an assignment, I'm afraid to drag down the whole team due to my limited time and dedication during final season, so I chose to work alone.)

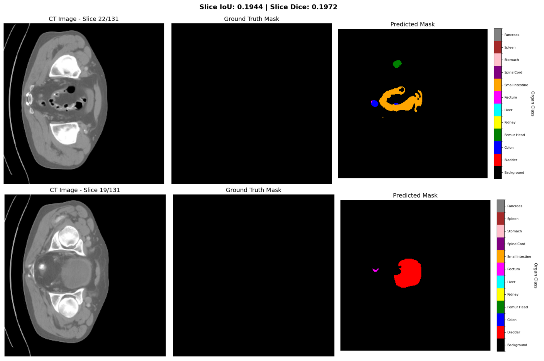


Fig. 8: Inference on AbdomenCT-1K Test Scan. The model does try to segment organs other than liver, kidney, spleen and pancreas, even though they are largely unannotated in the training dataset. The middle column is the mask, where unannotated organs are labeled as background, even though they do exist in the image.

VII. Acknowledgment

The author would like to thank TA Zishen Li for her helpful suggestions on the poster presentation. The author would also like to thank Medical Physics Workshop at School of PHysics, Peking University for providing the clinical data from PUTH.

References

[1] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” 2016. [Online]. Available: https://arxiv.org/abs/1606.04797

[2] S. Perera, P. Navard, and A. Yilmaz, “Segformer3d: an efficient transformer for 3d medical image segmentation,” 2024. [Online]. Available: https://arxiv.org/abs/2404.10156

[3] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen, B. Fu, S. Zhang, J. He, and Y. Qiao, “Sam-med3d: Towards general-purpose segmentation models for volumetric medical images,” 2024. [Online]. Available: https://arxiv.org/abs/2310.15161

[4] Y. Du, F. Bai, T. Huang, and B. Zhao, “Segvol: Universal and interactive volumetric medical image segmentation,” arXiv preprint arXiv:2311.13385, 2023.

[5] L. Li, S. Lian, Z. Luo, B. Wang, and S. Li, “Vclipseg: Voxel-wise clip-enhanced model for semi-supervised medical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024, pp. 692–701.

[6] S. Huang, H. Liang, Q. Wang, C. Zhong, Z. Zhou, and M. Shi, “Seg-sam: Semantic-guided sam for unified medical image segmentation,” 2024. [Online]. Available: https://arxiv.org/abs/2412.12660

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[8] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, “Abdomenct-1k: Is abdominal organ segmentation a solved problem?” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6695–6714, 2022.

[9] N. Shen, Z. Wang, J. Li, H. Gao, W. Lu, P. Hu, and L. Feng, “Multi-organ segmentation network for abdominal ct images based on spatial attention and deformable convolution,” Expert Systems with Applications, vol. 211, p. 118625, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422016748

[10] N. Zettler and A. Mastmeyer, “Comparison of 2d vs. 3d u-net organ segmentation in abdominal 3d ct images,” 2021. [Online]. Available: https://arxiv.org/abs/2107.04062

[11] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, “Abdomenct-1k: Is abdominal organ segmentation a solved problem?” 2021. [Online]. Available: https://arxiv.org/abs/2010.14808