# ISVAM : Inter-Slice Variation Attention Module for 2.5D Multi-Organ Segmentation in CT


Hyunji Lee
School of Computer Science and Engineering,
Kyungpook National University
Daegu, South Korea
hyunji0483@knu.ac.kr

Jaeseok Jang
School of Computer Science and Engineering,
Kyungpook National University
Daegu, South Korea
supercomeng@knu.ac.kr

Yu Rim Lee
Department of Internal Medicine,
College of Medicine,
Kyungpook National University
Daegu, South Korea
deblue00@naver.com

Soo Young Park
Department of Internal Medicine,
College of Medicine,
Kyungpook National University
Daegu, South Korea
psyoung0419@gmail.com

Won Young Tak
Department of Internal Medicine,
College of Medicine,
Kyungpook National University
Daegu, South Korea
wytak@knu.ac.kr

Soon Ki Jung*
School of Computer Science and Engineering,
Kyungpook National University
Daegu, South Korea
skjung@knu.ac.kr



***Abstract*—Multi-organ segmentation in CT images provides essential information for disease diagnosis, surgical planning, and treatment strategy development. It plays a crucial role in understanding disease progression and deriving optimized treatment plans by analyzing the volume, shape, and spatial relationships of organs. Effectively learning the continuity between adjacent slices and the variability of organ boundaries is a crucial factor in improving segmentation performance. 2D segmentation methods do not reflect this continuity, and 3D segmentation requires a lot of resources. As a result, there are limitations to practical applications. To address this issue, this study proposes the Inter-Slice Variation Attention Module (ISVAM), which can be easily integrated into various U-Net-based models. To overcome the limitations of existing 2D segmentation methods that do not account for continuity between adjacent slices, we utilize Bidirectional ConvLSTM to learn the spatial continuity of adjacent slices. Additionally, a supervised attention mechanism using ground truth differences is introduced to explicitly learn the boundary and positional changes between adjacent slices. The proposed method achieved superior performance compared to the baseline 2D U-Net and 2.5D U-Net on the FLARE22 dataset for most classes, demonstrating particularly significant performance improvements in organs with high inter-slice variability, such as the pancreas and stomach. This study presents a solution to effectively balance the performance and cost of CT image segmentation models by enhancing the efficiency of 2.5D segmentation.**



***Index Terms*—2.5D Segmentation, Multi-Organ Segmentation, Attention Mechanism, CT Sequential Feature**


## I. Introduction

Multi-organ segmentation provides valuable information to healthcare professionals necessary for disease diagnosis, surgical planning, and treatment strategies. By analyzing the volume and spatial relationships of the organs, it can help to understand disease progression and optimize treatment strategies. Since medical image segmentation contributes to medical judgment, achieving high segmentation performance is crucial above all else. In particular, in CT images, the continuity of organ or lesion structures between adjacent slices is prominent, making it important to effectively learn this continuity, as it significantly impacts performance. To effectively learn this continuity, 3D segmentation is widely used, which has the advantage of reflecting the structural continuity and contextual information between slices. However, 3D segmentation models require high computational resources and memory. On the other hand, 2D segmentation uses relatively fewer resources but has the drawback of not being able to learn contextual information, such as continuity between slices or changes in organ structure. This limitation becomes particularly pronounced for organs with significant and complex boundary changes between slices (e.g., pancreas, duodenum), making it difficult to generalize the shape of the organ across different slices, which can lead to a decline in performance.

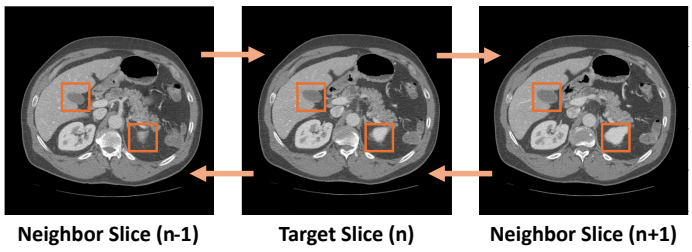


Fig. 1. Example of organ variation between adjacent slices.

To address this issue, a 2.5D segmentation approach has been proposed to compensate for the shortcomings of insuf-


This work was supported by the NRF (RS-2024-00360226) and ITRC program (IITP-2024-RS-2024-00437718), funded by the MSIT, Korea.

ficient continuity learning in 2D segmentation and the high resource consumption in 3D segmentation. This approach is widely used in the fields of CT and medical imaging. 2.5D segmentation uses the target slice along with adjacent slices or multiview slices as input, providing additional contextual information to improve the segmentation performance of the target slice while using fewer resources compared to 3D segmentation.

However, existing 2.5D segmentation approaches [1], [2] mainly focus on merging multiple slices for processing, and there are some studies [3], [4] that apply models such as Convolutional LSTM [5] to utilize the continuous features between slices. However, research that explicitly performs an attention mechanism on the boundary changes of organs between adjacent slices, which is crucial to maximize the efficacy of 2.5D segmentation, is rare.

In this study, we propose the Inter-Slice Variation Attention Module (ISVAM), which can explicitly focus on the structural variations of organs between adjacent slices in 2.5D segmentation. ISVAM is designed based on Bidirectional ConvLSTM to learn the continuity between adjacent slices and introduces a Supervised Attention Mechanism to guide the model to focus on the areas where changes have occurred based on the ground truth differences between adjacent slices and the target slice.

Based on the features that represent gradual change patterns occurring between adjacent slices learned through Bidirectional ConvLSTM, learning an Attention Map using the ground truth differences of adjacent slices allows the model to focus on the information that indicates "this area is where changes occur". This enables the model to efficiently learn that the concentrated areas do not exist in the target slice while being trained alongside the segmentation loss for the target slice. As a result, it reduces over-segmentation errors that occur in organs with complex boundary variations, which are difficult to generalize in 2D segmentation.

In particular, ISVAM achieves high accuracy even in organs with complex boundaries and structures (e.g., pancreas) or small-sized organs, effectively addressing the challenges of generalization in 2D segmentation due to boundary and size changes between slices. Additionally, the ISVAM proposed in this paper is a modular Attention Mechanism rather than a new network architecture, providing scalability and practicality that can be easily integrated into various U-Net [6]-based segmentation networks.

## II. Related Work

2.5D segmentation can address the shortcomings of 2D segmentation, which fails to utilize the volumetric context of CT images, while also alleviating the high computational resource demands required by 3D segmentation.

In particular, research is actively being conducted that provides additional contextual information related to the target slice by integrating the three orthogonal planes of CT images: sagittal, coronal, and axial planes. These studies [7], [8] typically employ 2D CNNs to learn each plane separately and then combine the segmentation results from each plane to enhance performance.

Additionally, methods that utilize consecutive slices as multi-channel inputs to segment the intermediate slice are also widely used. Related studies [1], [2] generally use three slices, including the current slice and adjacent slices, as input to the segmentation network. Some research focuses on designing new architectures for extracting useful information between slices so that the network can better capture the relevant information across slices.

In particular, there are studies [3], [4] that utilize Bidirectional Convolutional LSTM to enable the network to integrate information from both directions (previous and subsequent slices) to learn spatial continuity, thereby improving the overall performance of the segmentation results.

On the other hand, research [9] has also been conducted using Attention modules [10] that emphasize important areas by integrating information from adjacent slices without utilizing RNN-based models. For example, a methodology has been proposed that performs Slice attention to highlight semantic and positional information in 2.5D input data and to compensate for performance degradation caused by anisotropy in MRI.

In this study, we utilize Bidirectional ConvLSTM to learn spatial continuity between adjacent slices while generating channel attention and spatial attention maps based on this. Furthermore, by training these attention maps to be similar to the ground truth differences between adjacent slices, we aim to efficiently learn the variability between adjacent slices based on the target slice, thereby improving segmentation performance.

## III. Method

ISVAM is designed based on Bidirectional ConvLSTM to effectively reflect the structural variations of the target slice while learning the continuity between adjacent slices. In particular, it introduces a Supervised Attention Module [11] based on the ground truth differences between adjacent slices and the target slice, as can be seen in Fig. 2, allowing for explicit learning of boundary changes or positional changes between adjacent slices and the target slice.

ISVAM is designed based on the U-Net family of models [12], [13] and is inserted into the skip connection between the encoder and decoder, as shown in Fig. 3. This allows the high-resolution feature maps extracted by the encoder to not only be passed directly to the decoder but also to effectively convey changes between adjacent slices.

The target slice and two adjacent slices are stacked along the channel axis in the shape of [b, 3, h, w] and input into the encoder. The encoder independently extracts features from the three slices, processing them in parallel using group convolution [14] to perform independent convolution operations for each slice group, rather than sequentially. The feature maps extracted from the three slices pass through the ISVAM via skip connection, where ISVAM learns the structural continuity and changes between the target slice and the adjacent slices.

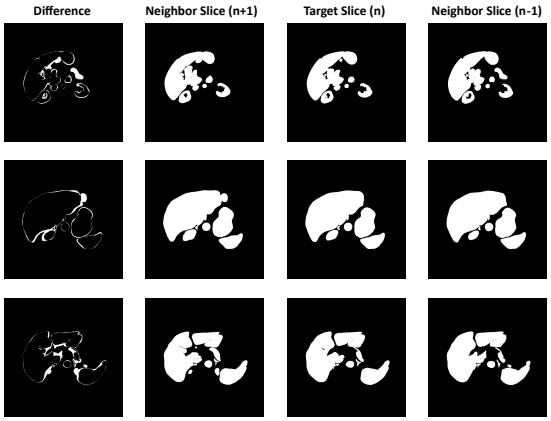


Fig. 2. Example of ground truth difference between adjacent slices.

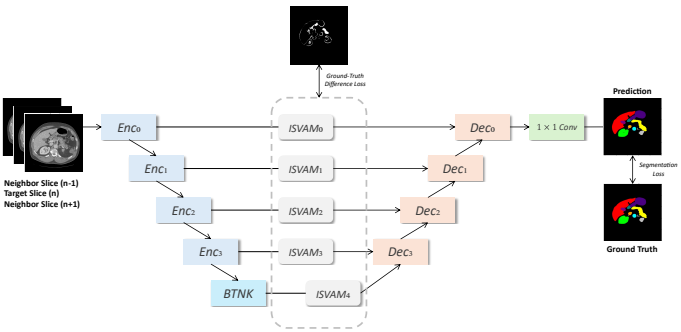


Fig. 3. Example of applying ISVAM to the skip connection of U-Net.

Ultimately, the enhanced feature map through ISVAM is passed to the decoder. ISVAM has three main components as follows.

### A. Bidirectional ConvLSTM

As shown in Fig. 4, the feature maps of the adjacent slices and the target slice extracted from the encoder layer are used as inputs for the Bidirectional ConvLSTM. The Bidirectional ConvLSTM is an extended structure of ConvLSTM that learns sequential information between slices in both directions, allowing it to reflect information from both previous and subsequent slices for the target slice. The Forward Layer processes the data in order and learns information from the previous slice, while the Backward Layer processes the data in reverse order to learn information from the subsequent slice. The Hidden States obtained from both directions are combined to generate the final output for each sequence. In this study, the Variant Attended Feature Map, which applies Channel Attention and Spatial Attention to the Hidden State of the target slice an intermediate point reflecting the sequential features of the previous and subsequent slices is passed to the decoder.

### B. Channel Attention Module

Channel Attention [10] emphasizes important features in the channel dimension of the feature map. To summarize the channel information of the feature map to which attention will

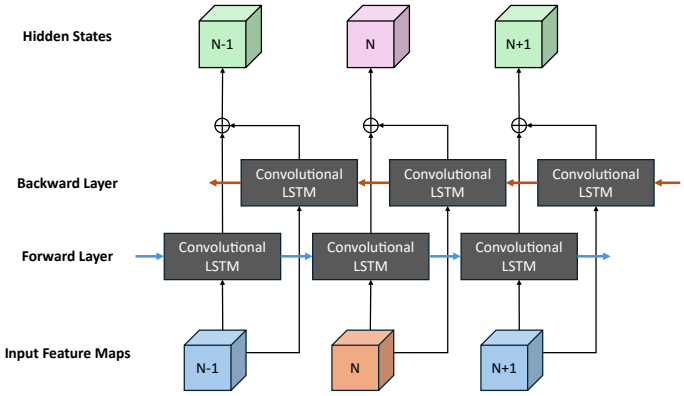


Fig. 4. How Bidirectional ConvLSTM Works.

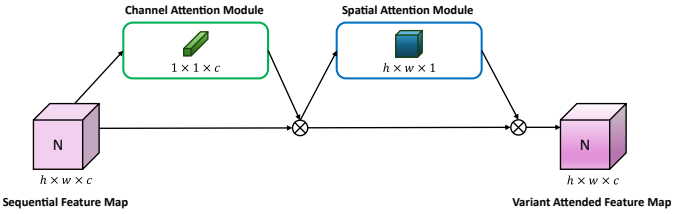


Fig. 5. Example of applying channel attention and spatial attention to the hidden state of the target slice in a bi-directional ConvLSTM.

be applied, Max Pooling and Average Pooling are performed along the height and width axes, compressing the global context of each channel. The two pooling results calculated in this way provide complementary information, which is then input into a Multi-Layer Perceptron (MLP) for a nonlinear transformation process.

The MLP consists of two Fully Connected Layers, where the first layer performs channel reduction, and the second layer performs channel expansion. After that, the values obtained from Max Pooling and Average Pooling are summed, and a Sigmoid function is applied to generate weights in the range of [0, 1]. These weights are ultimately used as the Channel Attention Map, reflecting the relative importance of each channel.

In this study, we designed the model's training to focus on the channels that best reflect the ground truth differences between the target slice and the adjacent slices, as shown in Fig. 7. The multi-scale Channel Attention Map generated at

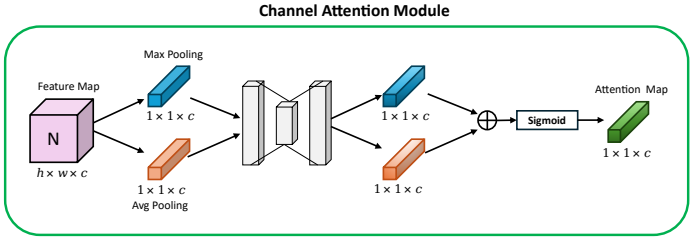


Fig. 6. Channel Attention Mechanism.

each layer is learned based on the downsampled ground truth differences corresponding to the resolution of that layer. The ground truth differences between adjacent slices are calculated through the following formula:

$$\text{Diff}_{prev}(i,j) = |GT_n(i,j) - GT_{n-1}(i,j)|, \quad (1)$$

$$\text{Diff}_{next}(i,j) = |GT_{n+1}(i,j) - GT_n(i,j)|, \quad (2)$$

$$\text{Diff}_{GT}(i,j) = \max(\text{Diff}_{prev}(i,j), \text{Diff}_{next}(i,j), \quad (3)$$

where $GT_n(i,j)$ represents the ground truth value at pixel $(i,j)$ in the slice $n$ and $\max(\cdot,\cdot)$ selects the larger value between its two inputs.

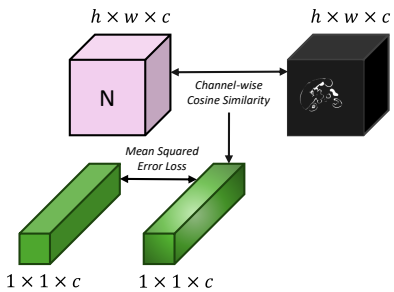


Fig. 7. Supervised Channel Attention Map.

The ground truth difference between adjacent slices is generated in the form of $h \times w \times 1$, and the cosine similarity is calculated between each channel of the hidden state (feature map) of the Bidirectional ConvLSTM for the target slice. This allows us to evaluate how well each channel reflects the ground truth difference. Subsequently, using min-max normalization, the Cosine Similarity results are normalized to the range of [0, 1], and we trained the model to make the channel attention map and the cosine similarity distribution similar using Mean Squared Error (MSE) Loss. This process can be represented by the following equation:

$$\text{CosSim}(F_c, \text{Diff}_{\text{GT}}) = \frac{F_c \cdot \text{Diff}_{\text{GT}}}{\|F_c\| \|\text{Diff}_{\text{GT}}\|}, \quad (4)$$

$$\text{CosSim}_{\text{norm}} = \frac{\text{CosSim} - \min(\text{CosSim})}{\max(\text{CosSim}) - \min(\text{CosSim})}, \quad (5)$$

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} \|\text{Channel Attention Map} - \text{CosSim}_{\text{norm}}\|^2, \quad (6)$$

where $F_c$ represents the feature map for the $c$-th channel generated by the Bidirectional ConvLSTM and the $n$ in mse loss represents the total number of channels.

Through this, we guided the Channel Attention Map to assign higher weights to important channels that reflect the differences from the ground truth.

### C. Spatial Attention Module

Spatial Attention [10] is used to emphasize spatially important locations in the feature map. To achieve this, Average Pooling and Max Pooling operations are performed along the channel dimension of the feature map to summarize the importance of spatial information. The average and maximum values calculated for each spatial location are then combined along the channel axis to form a map of size $h \times w \times 2$. Subsequently, the combined map passes through a $3 \times 3$ convolution layer to learn the correlations of spatial information, and then a Sigmoid function is applied to generate an attention map in the range of [0, 1]. Finally, this attention map is multiplied by the input feature map to assign higher weights to more important spatial areas, guiding the model to focus on crucial location information.

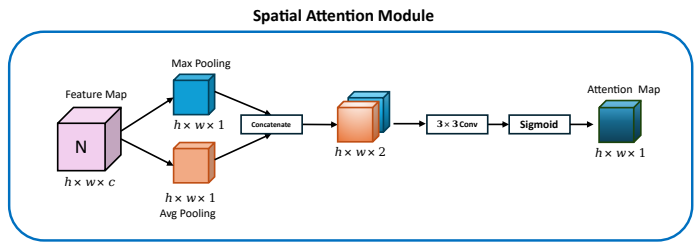


Fig. 8. Spatial Attention Mechanism.

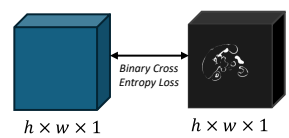


Fig. 9. Supervised Spatial Attention Map.

In this study, we used Binary Cross Entropy Loss as shown in Fig. 9. to train the model so that the difference in ground truth between adjacent slices becomes similar to the Spatial Attention Map. This approach encourages the model to focus on areas with significant spatial changes in the target slice compared to the adjacent slices during training. This process can be represented by the following equation:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{h \cdot w} \sum_{i=1}^{h} \sum_{j=1}^{w} \Big[ \text{Diff}_{\text{GT}}(i,j) \cdot \log(S(i,j)) + \big(1 - \text{Diff}_{\text{GT}}(i,j)\big) \cdot \log\big(1 - S(i,j)\big) \Big], \quad (7)$$

where $S(i,j)$ denotes the spatial attention map value at pixel $(i,j)$.

Ultimately, the segmentation model is trained by combining two losses for learning the attention map with a loss that combines the dice loss and cross-entropy loss for the target slice. The final formula for the loss used is as follows:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum P \cdot G}{\sum P^2 + \sum G^2}, \quad (9)$$

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum \Big[ G \cdot \log(P) + (1 - G) \cdot \log(1 - P) \Big], \quad (10)$$

$$\mathcal{L}_{\text{Seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}, \quad (11)$$

$$\mathcal{L}_{\text{Attention}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{BCE}}, \quad (12)$$

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Seg}} + \mathcal{L}_{\text{Attention}}, \tag{13}$$

where P represents the predicted value, G is the ground truth.

### D. Implementation Details

In this study, all experiments were conducted on a single NVIDIA H100 (80GB) GPU. To effectively visualize the differentiation of tissues in CT images, Hounsfield Unit (HU) values were clipped to the range of $[-260, 340]$ and then normalized to the range of [0, 1]. Both the baseline model and the proposed method were implemented based on the U-Net architecture. The Adam optimizer was used, with an initial learning rate set to $1X10^{-4}$. The learning rate scheduler employed was CosineAnnealingLR [16], which allowed for a high initial learning rate that gradually decreased to promote stable training. The scheduler's period was set to 100 epochs, and the main training was conducted for a total of 100 epochs. Additionally, to prevent overfitting, the Early Stopping technique was applied, and training was halted early if the validation loss did not improve for more than 15 epochs.

## IV. Experimental Result

### A. Dataset

Experiments were conducted using the FLARE22 (FAST and LOW resource abdominal organ segmentation challenge) [15] dataset. The FLARE22 dataset consists of 3D CT scan data for abdominal organ segmentation, including abdominal CT data from 50 patients collected from various institutions. Each scan is provided as abdominal cross-sectional images with a resolution of $512 \times 512$, and it includes labels for major abdominal organs such as the liver, kidneys, spleen, pancreas, and aorta. Notably, this dataset offers detailed labeling for 13 organs, making it suitable for evaluating and comparing multi-organ segmentation performance. Additionally, out of a total of 4,794 slices, 3,846 were used for the training set and 948 for the validation set.

### B. Evaluation Metrics

To evaluate the performance of the model, we used the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU), which are frequently used in medical image segmentation, It is defined as follows:

$$\text{DSC} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \tag{14}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{15}$$

where $TP$, $FP$, and $FN$ are the number of true positives, false positives, and false negatives, respectively. These measures result in an output between 0 and 1, with closer to 1 indicating better performance.

### C. Quantitative Results

Tables I and II present the results of evaluating model performance based on the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) using the FLARE22 dataset. The experiments were conducted with seven different settings, including the conventional 2D U-Net and the 2.5D U-Net that uses adjacent slices as multi-channel input. Additionally, methods were included that applied only bidirectional ConvLSTM, used the conventional attention module without supervised attention based on ground truth differences between adjacent slices, applied only supervised attention without bidirectional ConvLSTM, and used only the conventional attention module, as well as the proposed ISVAM.

As a result, the proposed method demonstrated the best performance in terms of average DSC and IoU, showing improvements over the baseline in most of the 13 classes of the FLARE22 dataset. Notably, performance improvements were particularly pronounced in organs with high variability between slices, such as the pancreas and stomach. Significant performance improvements were also observed in relatively stable and consistent organs, such as the spleen and left kidney. Furthermore, the key contributions of our study, which are the continuity learning between adjacent slices using bidirectional ConvLSTM and the supervised attention utilizing ground truth differences between adjacent slices, both showed

TABLE I
PERFORMANCE COMPARISON (DSC)

| | Liver | RK | Spleen | Pancreas | Aorta | IVC | RAG |
|---|---|---|---|---|---|---|---|
| 2D U-Net (baseline) | 93.02 | 92.04 | 91.54 | 84.69 | 94.10 | 90.17 | 87.21 |
| 2.5D U-Net (baseline) | 93.77 | **96.62** | 94.92 | 81.98 | 93.56 | 86.60 | 85.29 |
| Bi-ConvLSTM | 94.76 | 92.40 | 93.00 | 82.76 | 93.53 | 88.70 | 89.74 |
| Bi-ConvLSTM + CBAM [10] | **95.34** | 93.05 | 94.51 | 82.47 | 93.99 | 89.41 | 90.80 |
| supervised attention | 92.76 | 94.62 | 94.07 | 85.51 | **95.42** | 89.56 | **93.45** |
| CBAM [10] | 92.22 | 92.78 | 93.31 | 85.56 | 94.63 | 89.92 | 87.36 |
| ISVAM (ours) | 93.84 | 95.71 | **95.72** | **85.85** | 94.46 | **91.17** | 91.52 |
| | LAG | Gallbladder | Esophagus | Stomach | Duodenum | LK | AVG |
| 2D U-Net (baseline) | 91.64 | 94.06 | 93.19 | 90.30 | 82.78 | 91.54 | 90.48 |
| 2.5D U-Net (baseline) | 91.32 | 93.68 | 92.67 | 87.21 | 76.58 | 92.53 | 89.75 |
| Bi-ConvLSTM | 88.62 | **96.26** | 94.86 | 90.69 | **86.40** | 93.22 | 91.15 |
| Bi-ConvLSTM + CBAM [10] | 92.57 | 87.31 | 94.56 | 91.84 | 82.00 | 92.97 | 90.83 |
| supervised attention | **93.50** | 94.67 | 94.51 | 92.38 | 83.50 | 93.84 | 92.14 |
| CBAM [10] | 92.54 | 94.28 | **95.41** | 91.85 | 85.84 | 92.13 | 91.37 |
| ISVAM (ours) | 92.12 | 92.73 | 95.08 | **93.00** | 83.51 | **96.13** | **92.37** |

TABLE II
PERFORMANCE COMPARISON (IOU)

| | Liver | RK | Spleen | Pancreas | Aorta | IVC | RAG |
|---|---|---|---|---|---|---|---|
| 2D U-Net (baseline) | 91.72 | 90.98 | 90.80 | 80.79 | 91.58 | 86.67 | 85.70 |
| 2.5D U-Net (baseline) | 92.57 | **95.63** | 94.27 | 78.26 | 91.02 | 82.74 | 83.61 |
| Bi-ConvLSTM | 93.21 | 91.23 | 92.30 | 78.89 | 91.16 | 84.92 | 88.23 |
| Bi-ConvLSTM + CBAM [10] | **93.75** | 91.84 | 93.87 | 78.36 | 91.58 | 86.11 | 89.35 |
| supervised attention | 91.29 | 93.56 | 93.31 | 81.77 | **92.86** | 86.16 | **91.85** |
| CBAM [10] | 90.75 | 91.72 | 92.54 | 81.82 | 92.28 | 86.53 | 85.81 |
| ISVAM (ours) | 92.29 | 94.70 | **95.15** | **82.08** | 92.04 | **87.75** | 90.01 |
| | LAG | Gallbladder | Esophagus | Stomach | Duodenum | LK | AVG |
| 2D U-Net (baseline) | 90.05 | 92.97 | 91.31 | 87.88 | 78.62 | 90.43 | 88.42 |
| 2.5D U-Net (baseline) | 89.76 | 92.54 | 90.50 | 84.94 | 72.66 | 91.37 | 87.68 |
| Bi-ConvLSTM | 87.06 | **95.04** | 92.96 | 88.24 | **82.23** | 91.98 | 89.21 |
| Bi-ConvLSTM + CBAM [10] | 91.07 | 86.12 | 92.48 | 89.73 | 77.63 | 91.74 | 88.74 |
| supervised attention | **91.98** | 93.63 | 92.73 | 90.13 | 79.53 | 92.54 | 90.10 |
| CBAM [10] | 91.03 | 93.20 | **93.59** | 89.90 | 82.00 | 90.94 | 89.39 |
| ISVAM (ours) | 90.54 | 91.52 | 93.28 | **90.63** | 79.39 | **94.99** | **90.34** |

superior performance compared to the baseline when applied individually, and the best performance was achieved when both were applied together.

### D. Qualitative Results

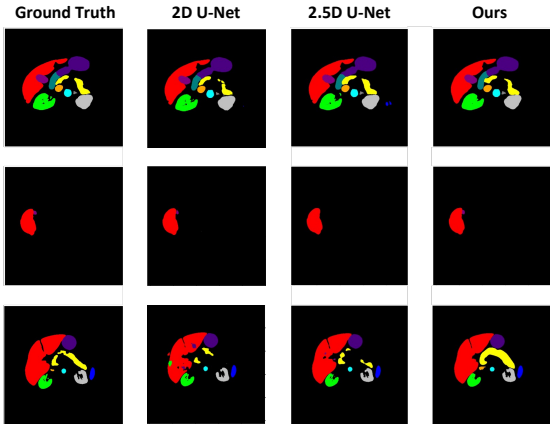


Fig. 10. Visual Result on FLARE22.

Fig. 10 shows the visual results of the experimental outcomes on the FLARE22 dataset. In comparison, the proposed ISVAM-based model demonstrates better segmentation performance by accurately learning the boundaries, especially in organs with high variability between slices, such as the pancreas. Additionally, when compared to the ground truth, it can be confirmed that the proposed method mitigates over-segmentation errors compared to the vanila 2D U-Net or 2.5D U-Net models.

## V. Conclusion and Future Work

In this study, we proposed ISVAM, which utilizes the continuity between adjacent slices and the differences in ground truth, effectively enhancing the performance of 2.5D multi organ segmentation. The combination of Bidirectional ConvLSTM and a supervised attention mechanism is designed to focus on boundary changes and positional differences between slices, which particularly contributed to improving performance in organs with complex boundary changes, such as the pancreas. Experimental results showed that our method outperformed existing approaches in terms of average DSC and IoU on the FLARE22 dataset, demonstrating improved segmentation performance on target slices.

Future work should address the limitation of representing the ground truth differences of adjacent slices in binary form without distinguishing between classes, in order to improve the differentiation of changes between classes. Additionally, we aim to design a mechanism that can reflect directional changes, such as the increase or decrease in the size of specific organs, in the ground truth differences to achieve a more refined variation attention module. Through this, we expect to enhance the scalability with various medical imaging datasets and maintain high performance even in more complex organ structures.

## Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00360226) and also supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center)support program(IITP-2024-RS-2024-00437718) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

## References

[1] Zhou, Hanyue, et al. "Intracranial vessel wall segmentation for atherosclerotic plaque quantification." 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021.

[2] Li, Lei, et al. "Learning consistency-and discrepancy-context for 2D organ segmentation." Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing, 2021.

[3] Zhang, Yao, et al. "SequentialSegNet: combination with sequential feature for multi-organ segmentation." 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018.

[4] Chen, Jianxu, et al. "Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation." Advances in neural information processing systems 29 (2016).

[5] Shi, Xingjian, et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems 28 (2015).

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.

[7] Zhang, Yue, et al. "A deep learning framework for pancreas segmentation with multi-atlas registration and 3D level-set." Medical Image Analysis 68 (2021): 101884.

[8] Li, Hao, et al. "A model-driven stack-based fully convolutional network for pancreas segmentation." 2020 5th International Conference on Communication, Image and Signal Processing (CCISP). IEEE, 2020.

[9] Hung, Alex Ling Yu, et al. "CSAM: A 2.5 D Cross-Slice Attention Module for Anisotropic Volumetric Medical Image Segmentation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.

[10] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV). 2018.

[11] Fan, Yingruo, Victor OK Li, and Jacqueline CK Lam. "Facial expression recognition with

[12] deeply-supervised attention network." IEEE transactions on affective computing 13.2 (2020): 1057-1071. Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing, 2018.

[13] Huang, Huimin, et al. "Unet 3+: A full-scale connected unet for medical image segmentation." ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2020.

[14] Cohen, Taco, and Max Welling. "Group equivariant convolutional networks." International conference on machine learning. PMLR, 2016.

[15] FLARE Challenge. "FLARE22 Dataset." 2022. Grand Challenge, https://flare22.grand-challenge.org/.

[16] Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts." arXiv preprint arXiv:1608.03983 (2016).