Medical Vision Seminar

Luyue Shi
June 30, 2021

Outline:

Structure Boundary Preserving Segmentation for Medical Image with Ambiguous Boundary (CVPR2020)

DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets

(CVPR2021)

(CVPR2020)

Structure Boundary Preserving Segmentation for Medical Image with Ambiguous Boundary

Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, Yong Man Ro
Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

Introduction

Problem

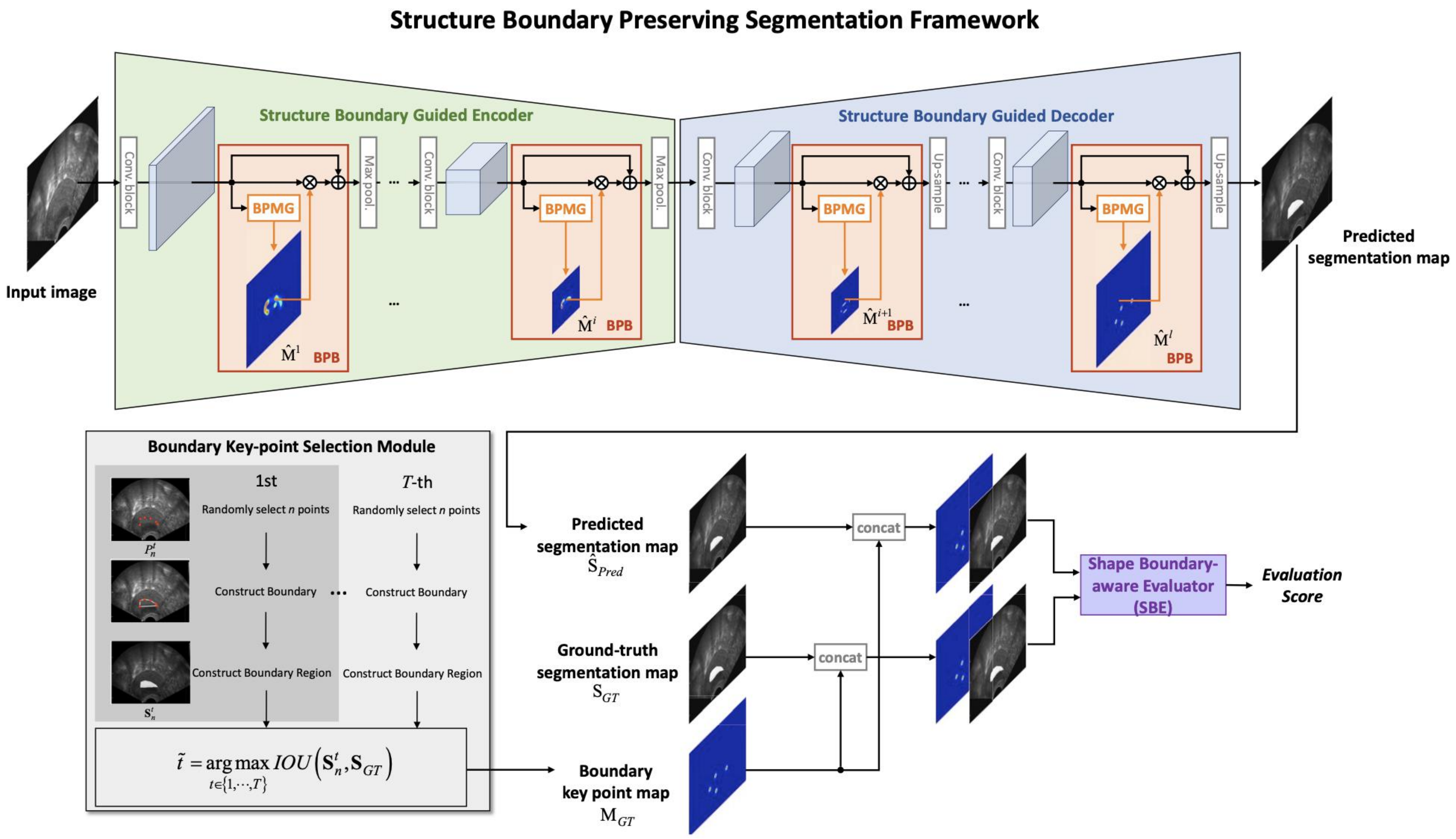
a. Ambiguity of structure boundary in the medical image domain
b. Uncertainty of the segmented region without specialized domain knowledge

Contribution

a. A novel boundary key point selection algorithm that best fit the target region. The selected key points putting on the structure boundary of target region are encoded through the BPB with boundary key point map generator.
b. Employ boundary key point information automatically without the user interaction. To this end, we trained the segmentation network in an adversarial way with SBE. The evaluator gives feedback to segmentation network whether given segmented region coincidences with boundary key points or not.
c. The proposed method can be generalized to different segmentation models. The proposed method improves the prediction performance with statistical significance.

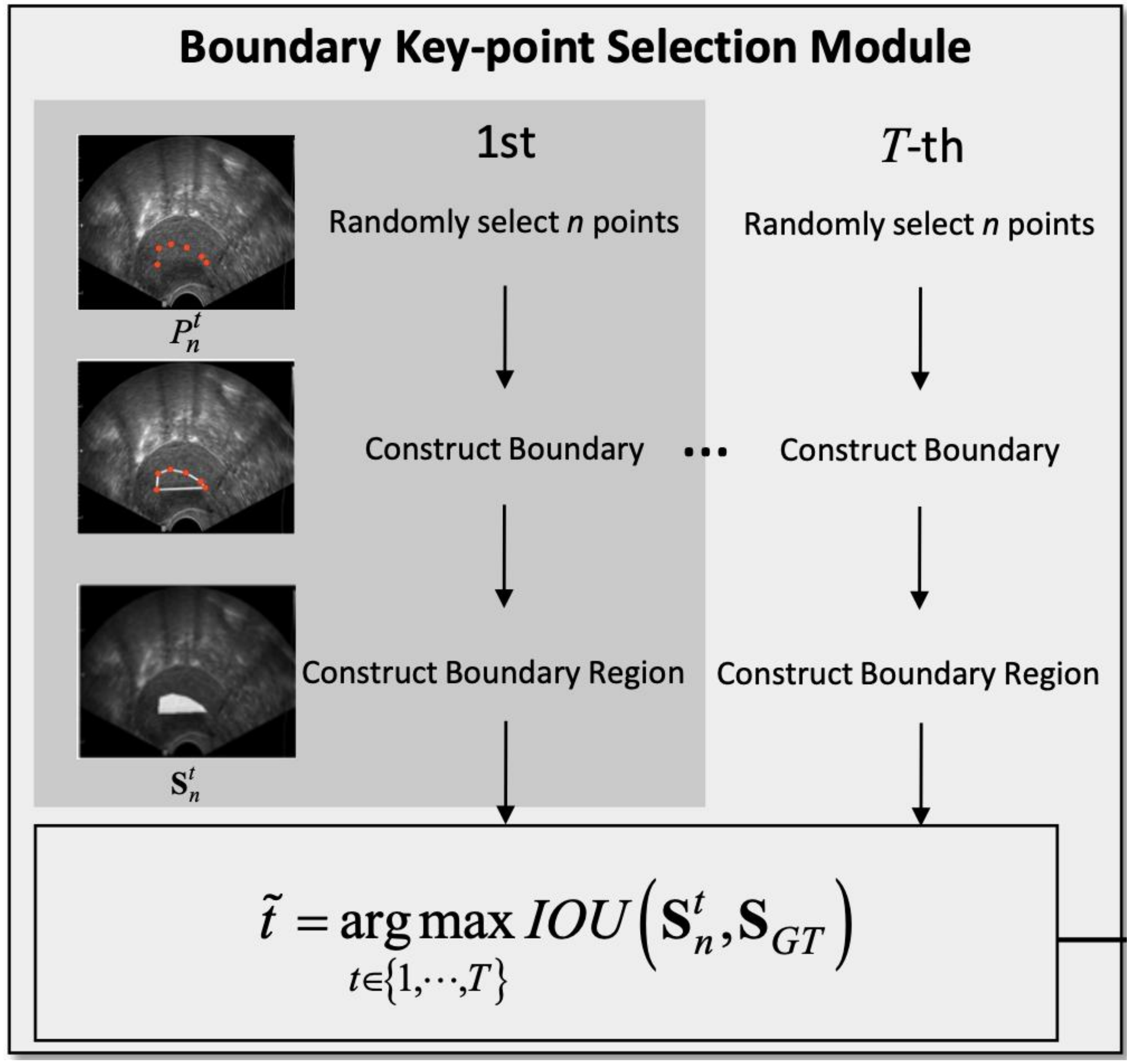
Methodology

- Overview



Methodology

- Boundary Key Point Selection Algorithm



Algorithm 1: Boundary key point selection algorithm

Input: Total number of iterations T, number of boundary key points n, ground truth segmentation map $\mathbf{S}_{GT}$

Output: Boundary key Points $\tilde{P}$

Initialize $IOU_{best} = 0$

for $t = 1, 2, \cdots, T$ **do**

  Randomly select N points

  $P_n^t \leftarrow \{(x_1^t, y_1^t), (x_2^t, y_2^t), \cdots, (x_n^t, y_n^t)\}$

  $\mathbf{S}_n^t \leftarrow c(P_n^t)$

  $IOU_t \leftarrow IOU(\mathbf{S}_n^t, \mathbf{S}_{GT})$

  if $IOU_t > IOU_{best}$ **then**

    $IOU_{best} \leftarrow IOU_t$

    $\tilde{P} \leftarrow P_n^t$

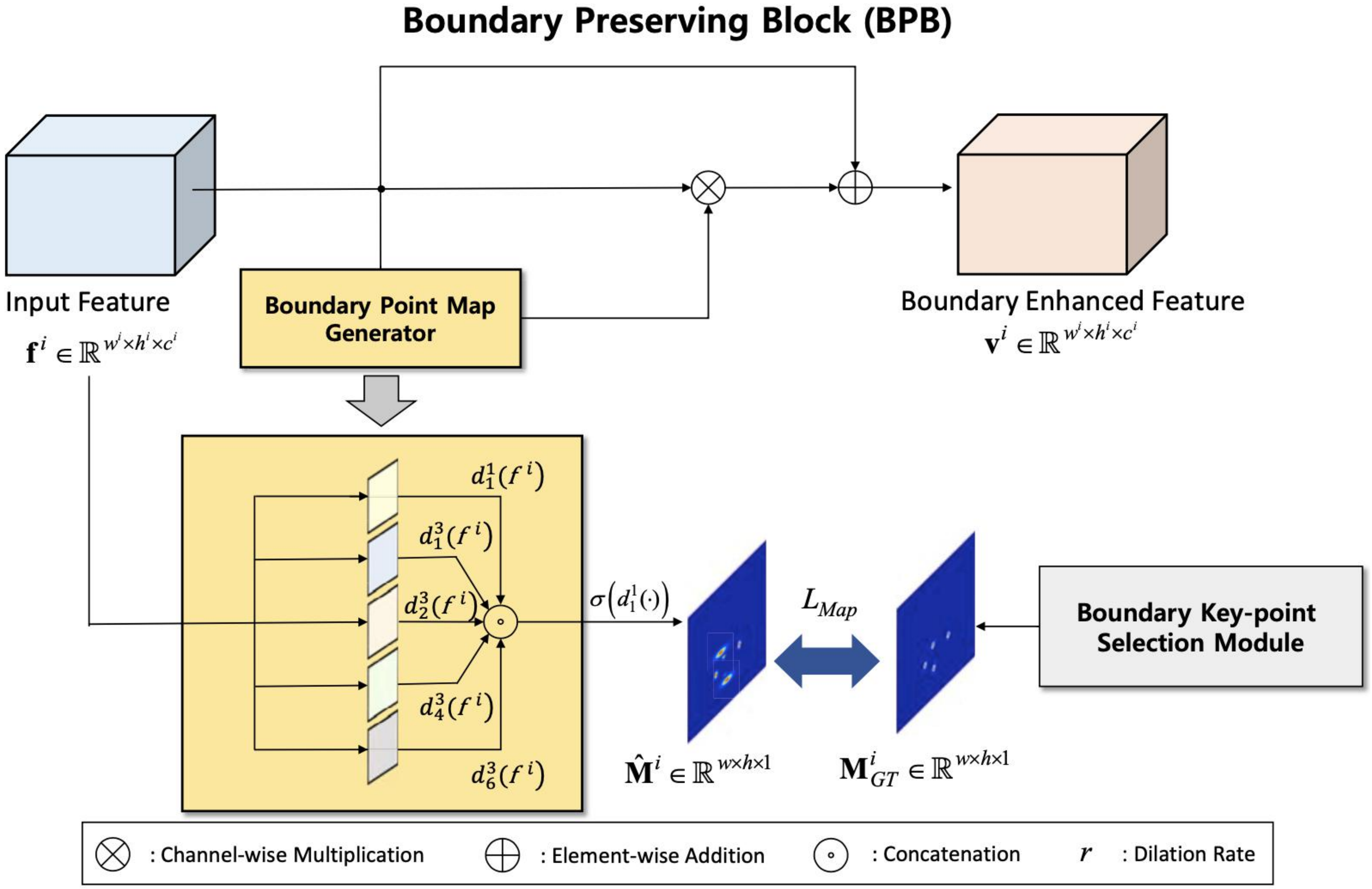
  end

end

Return: $\tilde{P}$

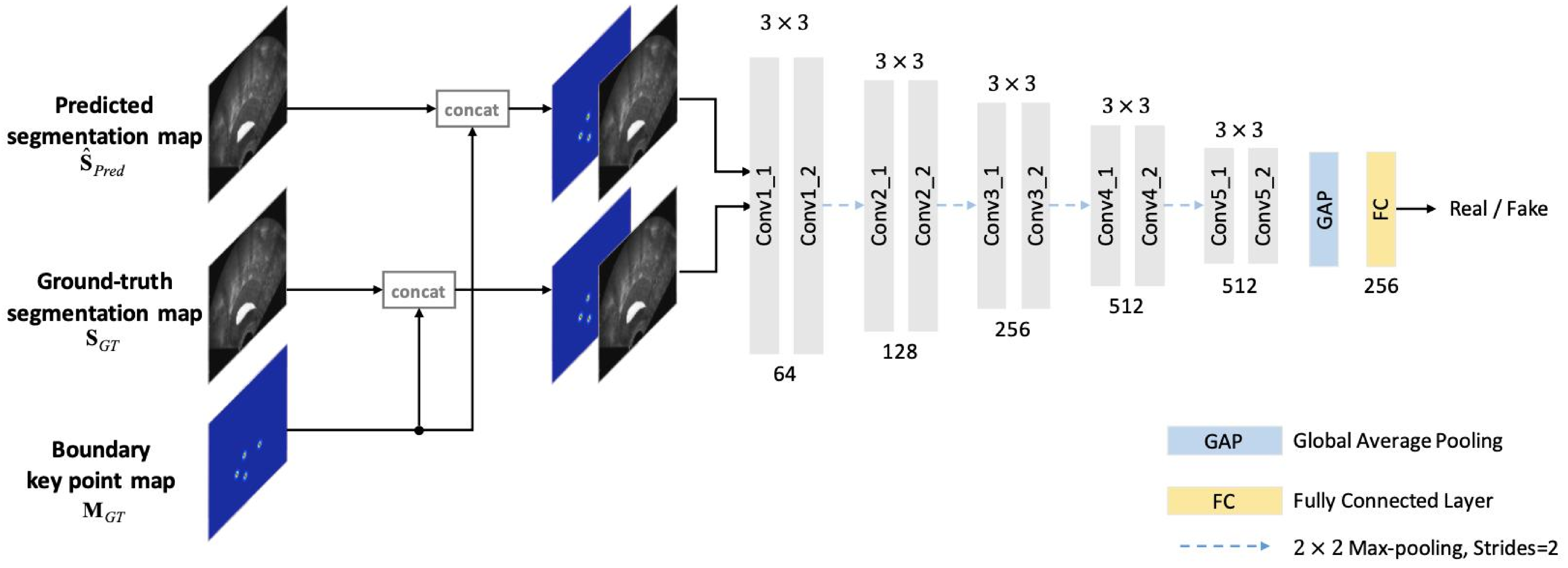
Methodology

- Boundary Preserving Block (BPB)



Methodology

- Shape Boundary-aware Evaluator (SBE)



$$
\begin{aligned}
L_{SBE} = & -\log\left(D\left(\mathrm{S}_{GT};\mathrm{M}_{GT}\right)\right) \\
& -\log\left(1 - D(\hat{\mathrm{S}}_{Pred};\mathrm{M}_{GT})\right)
\end{aligned}
$$

Methodology

- Loss for Segmentation Network

$$L_{Seg} = -\,\mathrm{S}_{GT} \cdot \log\left(\hat{\mathrm{S}}_{Pred}\right)$$
$$\qquad - (1 - \mathrm{S}_{GT}) \cdot \log\left(1 - \hat{\mathrm{S}}_{Pred}\right)$$

$$L_{BA} = -\log\left(D\left(\hat{\mathrm{S}}_{Pred}; \mathrm{M}_{\mathrm{GT}}\right)\right)$$

$$L_{Map}^{i} = -\mathrm{M}_{GT}^{i} \cdot \log \hat{\mathrm{M}}^{i} - (1 - \mathrm{M}_{GT}^{i}) \cdot \log(1 - \hat{\mathrm{M}}^{i})$$

$$L_{Total} = L_{Seg} + L_{BA} + \sum_{i=1}^{l} L_{Map}^{i}$$

Experiments

- Dataset

A. PH2+ISBI 2016 Skin Lesion Challenge dataset (public):

200 dermoscopic images (Testing) + 900 skin lesion images (Training)

B. Transvaginal Ultrasound (TVUS) dataset (private):

3,360 transvaginal ultrasound images and the corresponding endometrium segmentation maps (5-fold cross-validation)

Experiments

- Quantitative Evaluation

Table 1. Dice and Jaccard coefficient comparison of our approach and six different approaches on PH2 + ISBI 2016 Challenge dataset.

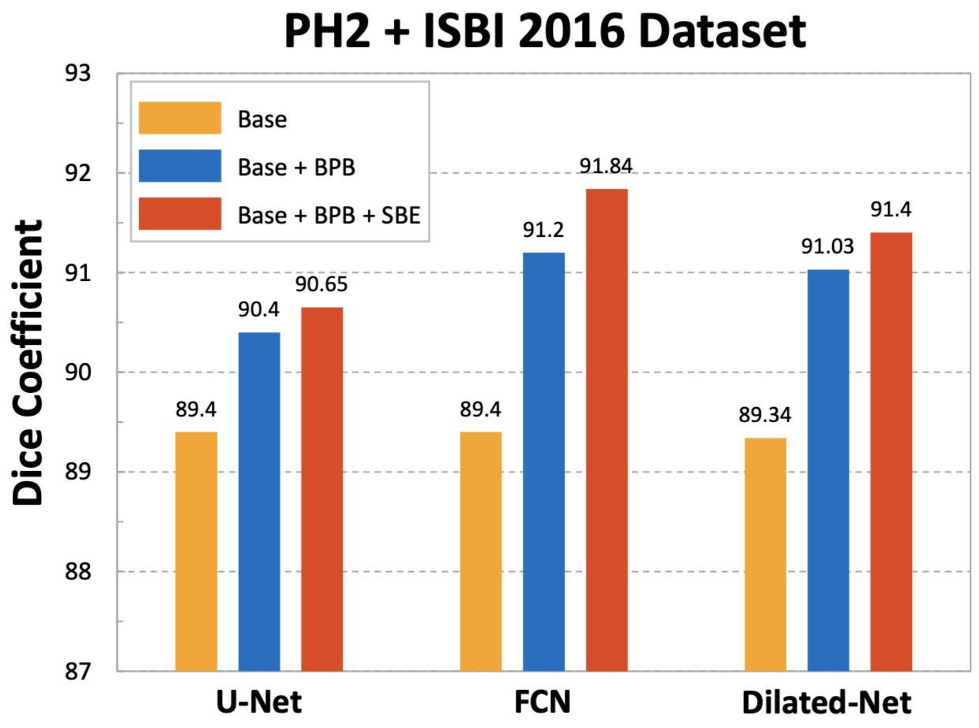
Method	Dice Coefficient	Jaccard Coefficient
SCDRR [4]	86.00	76.00
JCLMM [23]	82.85	-
MSCA [2]	81.57	72.33
SSLS [1]	78.38	68.16
FCN [15]	89.40	82.15
Bi et al. (2017) [3]	90.66	83.99
FCN+BPB+SBE (Our method)	**91.84**	**84.30**

Table 2. Dice and Jaccard coefficient comparison of our approach and conventional segmentation network on TVUS dataset.

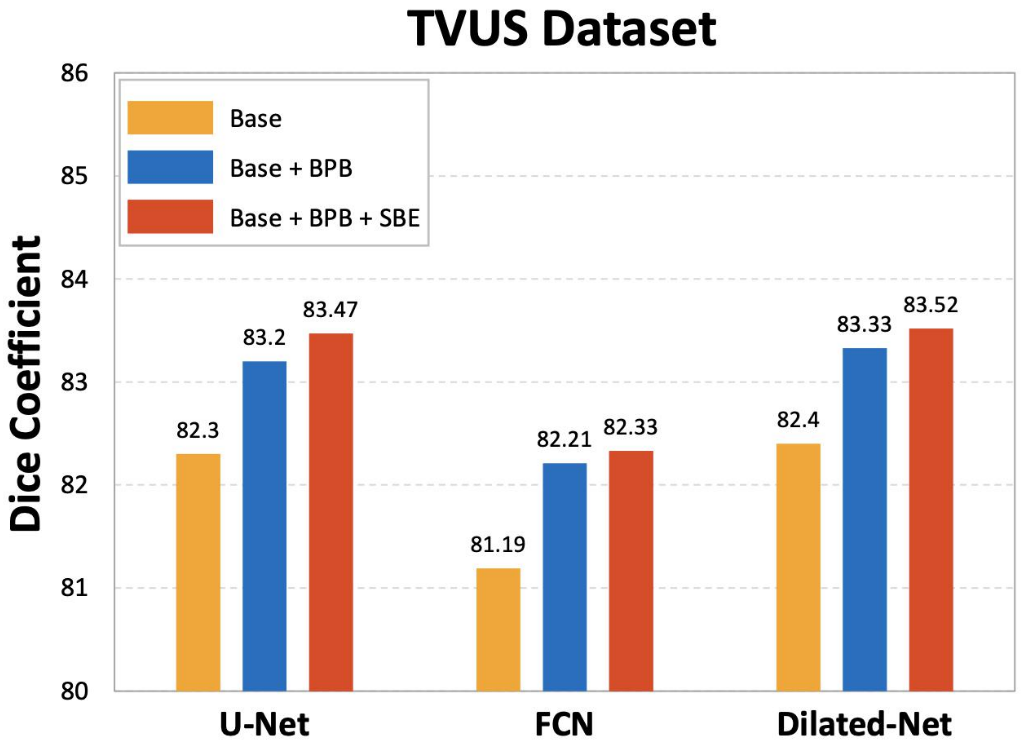
Method	Dice Coefficient	Jaccard Coefficient
U-Net [21]	82.30	70.38
FCN [15]	81.19	69.12
Dilated-Net [31]	82.40	70.36
Park et al. (2019) [18]	82.67	70.46
Dilated-Net+BPB+SBE (Our method)	**83.52**	**71.58**

Experiments

- Ablation Studies



(a)



(b)

Table 3. Statistical significance analysis of performance improvements by paired t-test on PH2+ISBI 2016 dataset.

Baseline Network	Mean difference ± Standard Error	95% CI	*p*-value
U-Net [21]	1.22±0.21	[0.51, 1.35]	p<0.0001
FCN [15]	2.17±0.28	[1.41, 2.72]	p<0.0001
Dilated-Net [31]	2.03±0.25	[1.34, 2.46]	p<0.0001

Table 4. Statistical significance analysis of performance improvements by paired t-test on TVUS dataset.

Baseline Network	Mean difference ± Standard Error	95% CI	*p*-value
U-Net [21]	0.92±0.13	[0.66, 1.17]	p<0.0001
FCN [15]	1.62±0.28	[1.07, 2.17]	p<0.0001
Dilated-Net [31]	0.90±0.11	[0.62, 1.1]	p<0.0001

Experiments

- Effect of Multiple BPBs

Table 5. Performance changes along with the number of BPBs on U-Net.

Method	Dice Coefficient
Encoder(front)	82.15
Decoder(end)	82.43
Center (1)	82.47
Center (3)	82.66
Center (6)	**83.20**

- Encoder (front): A BPB in the first layer of U-Net encoder.
- Decoder(end): A BPB in the last layer of U-Net decoder.
- Center (1): A BPB in the center of U-Net
- Center (3): 3 BPBs after 8 convolution layers
- Center (6): 6 BPBs after 4 convolution layers.

Experiments

- Qualitative Evaluation

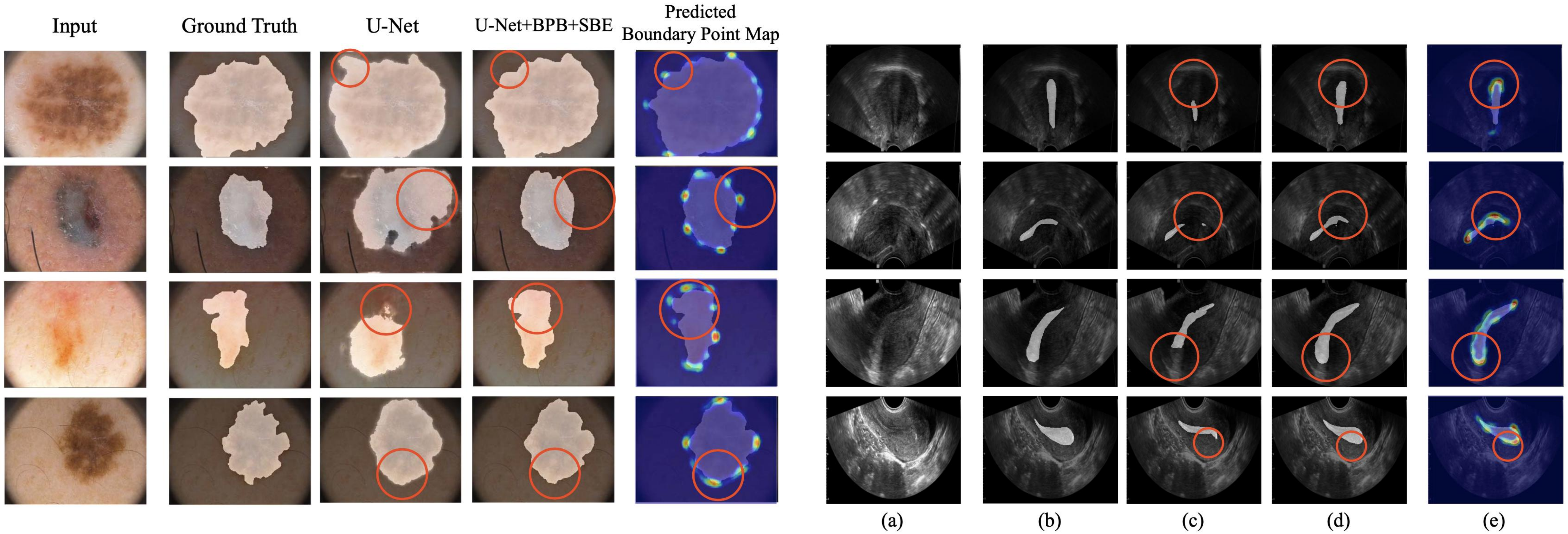


Figure 5. Segmentation results comparison of U-Net and U-Net + BPB +SBE method on PH2+ISBI 2016 (1-3 rows) and TVUS (4-6 rows). (a) is the original images, (b) is the ground truth segmentation images, (c) is the results of the U-Net, (d) is the U- Net+BPB+SBE and (e) is the visualization results of the generated key point map.

Experiments

- Define the Number of Key Points

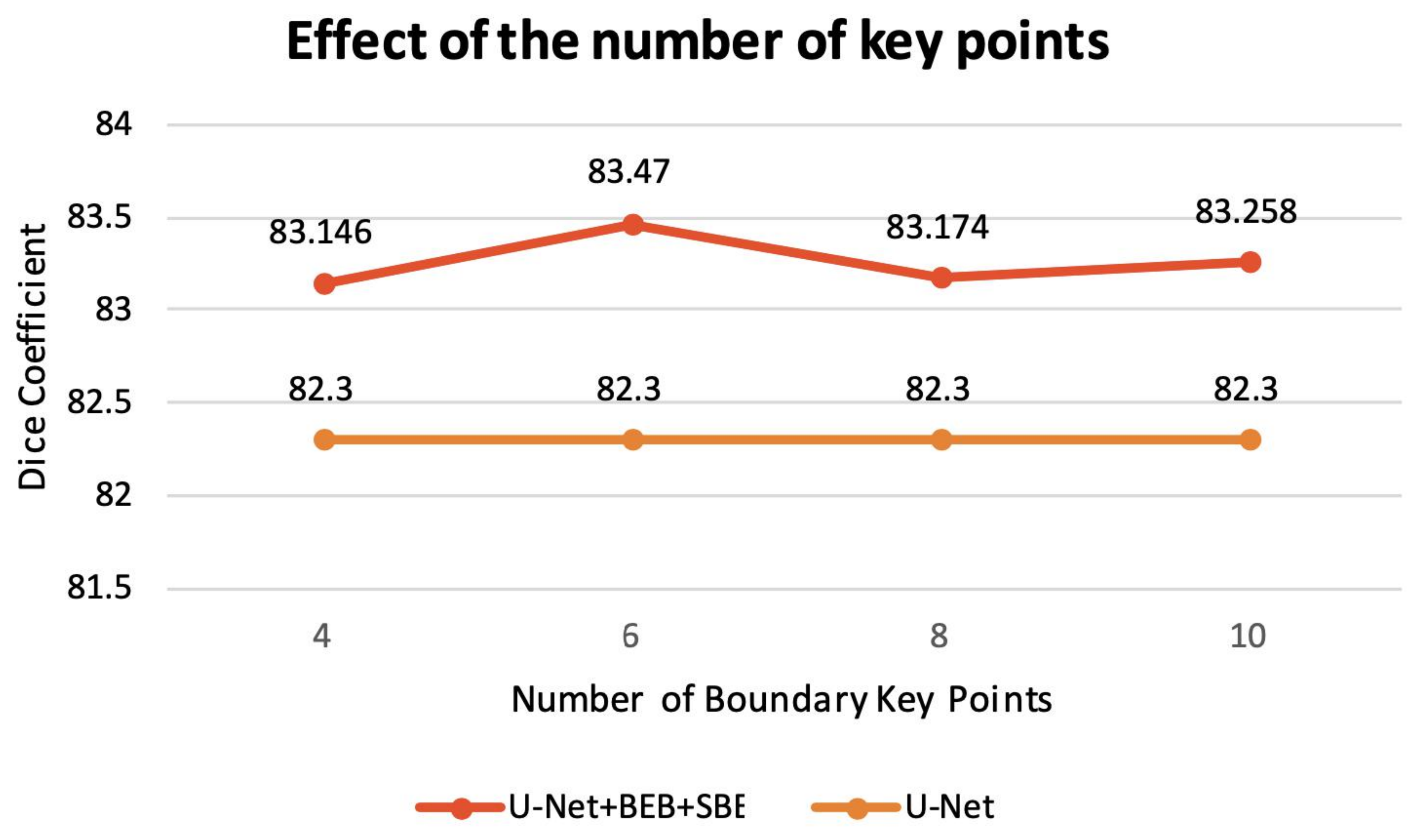


Figure 6. Performance evaluation in accordance with the number of boundary key points on TVUS dataset. It shows the comparison results between U-Net and U-Net + BPB + SBE.

(CVPR2021)

DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets

Jianpeng Zhang*1,2, Yutong Xie∗ 1,2, Yong Xia1 , and Chunhua Shen2

1 School of Computer Science and Engineering, Northwestern Polytechnical University, China

2 The University of Adelaide, Australia

Introduction

Partially labeling issue:

Most benchmark datasets were collected for the segmentation of only one type of organs and/or tumors, and all task-irrelevant organs and tumors were annotated as the background. How to learn the representation of multiple organs and tumors under the supervision of these partially annotated images.

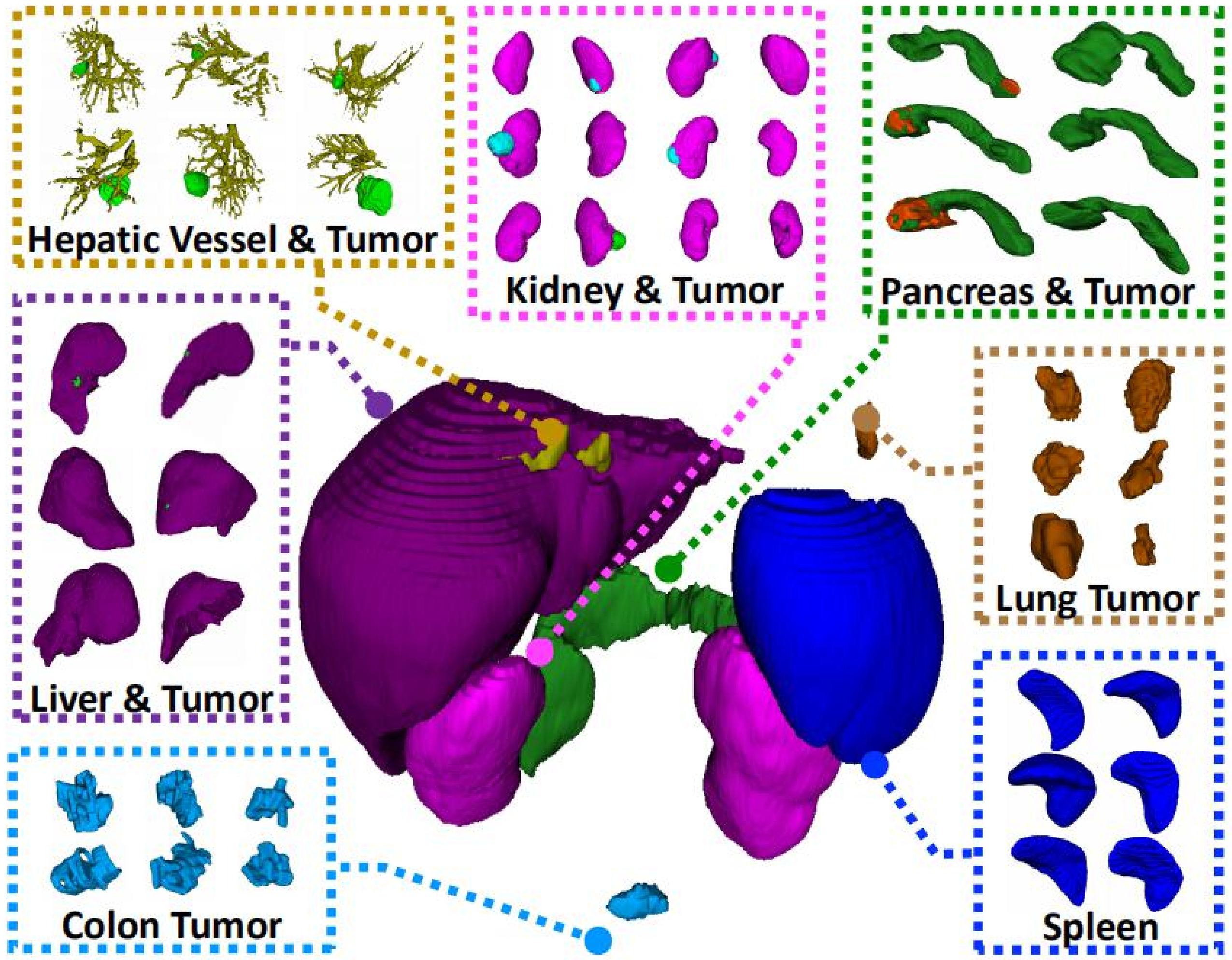


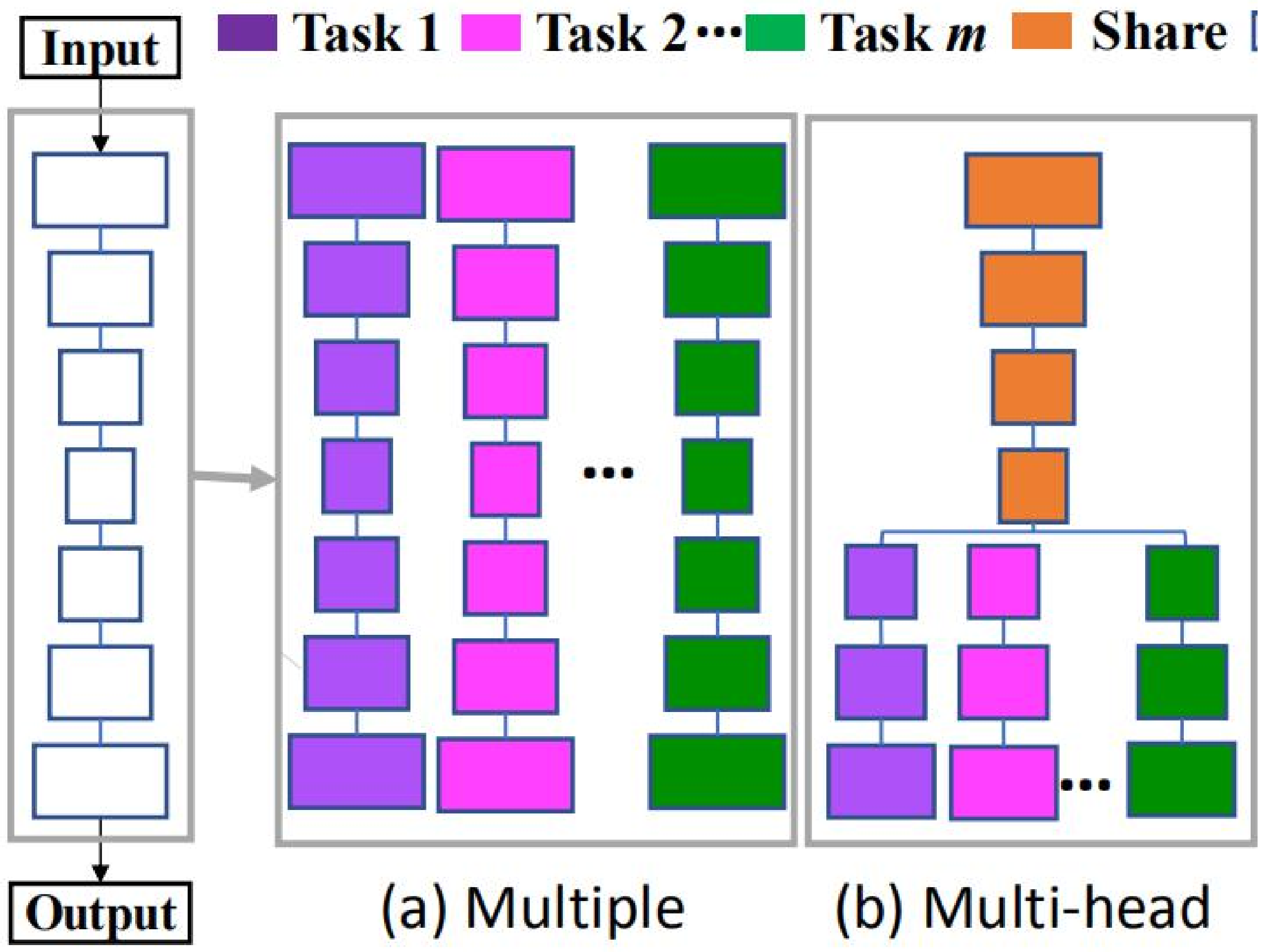
Figure 1 – Illustration of partially labeled multi-organ and tumor segmentation. This task aims to segment multiple organs and tumors using a network trained on several partially labeled datasets, each of which is originally specialized for the segmentation of a particular abdominal organ and/or related tumors. For instance, the first dataset only has annotations of the liver and liver tumors, and the second dataset only provides annotations of kidneys and kidney tumors. Here each color represents a partially labeled dataset.

Introduction

Related Works:

a) Multiple networks: increases the computational complexity dramatically.

b) Multi-head networks: In the training stage, when each partially labeled data is fed to the network, only one head is updated and others are frozen. The inferences made by other heads are unnecessary and wasteful. Besides, the inflexible multi-head architecture is not easy to extend to a newly labeled task.



(a) Multiple networks

(b) Multi-head network

Introduction

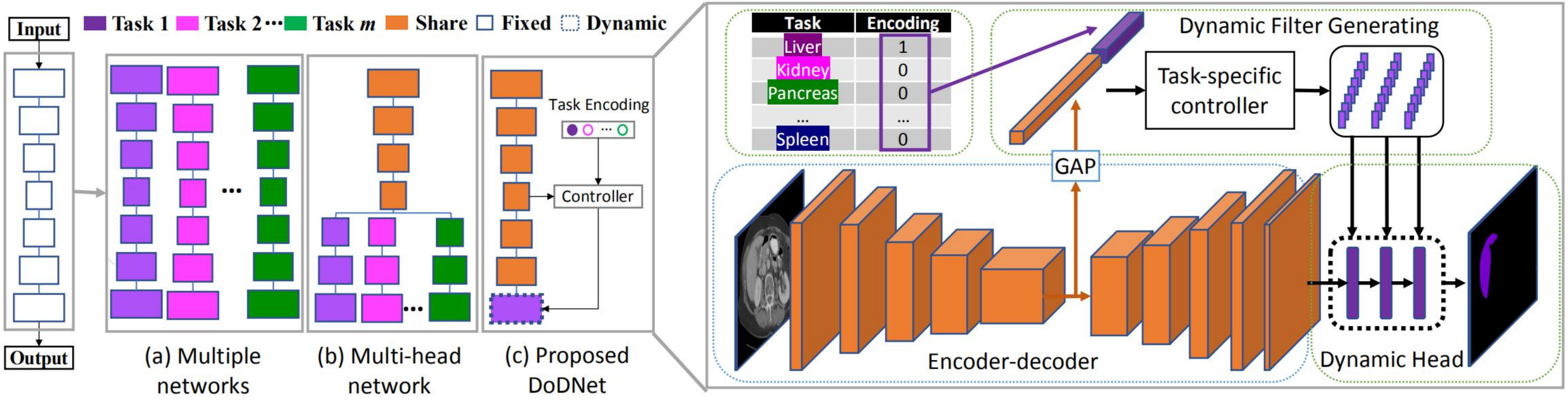
Contribution

a) We attempt to address the partially labeling issue from a new perspective, *i.e.*, proposing a single network that has a dynamic segmentation head to segment multiple organs and tumors as done by multiple networks or a multi-head network.

b) Different from the traditional segmentation head which is fixed after training, the dynamic segmentation head in our model is adaptive to the input and assigned task, leading to much improved efficiency and flexibility.

c) The proposed DoDNet pre-trained on partially labeled datasets can be transferred to downstream annotation limited segmentation tasks, and hence is beneficial for the medical community where only limited annotations are available for 3D image segmentation.

Method



- Shared encoder-decoder
- Task encoding module
- Dynamic filter generation module

$$\boldsymbol{\omega}_{ij} = \varphi(\mathrm{GAP}(\mathbf{F}_{ij})||\mathbf{T}_{ij};\boldsymbol{\theta}_{\varphi})$$

- Dynamic segmentation head

Conv layer	#Weights	#Bias
1	8×8	8
2	8×8	8
3	8×2	2
Totoal	162	

$$\mathbf{P}_{ij} = ((\mathbf{M}_{ij} * \boldsymbol{\omega}_{ij1}) * \boldsymbol{\omega}_{ij2}) * \boldsymbol{\omega}_{ij3} \quad \in \mathbb{R}^{2\times D\times W\times H}$$

Experiments

Dataset

- MOTS: 1155 3D abdominal CT scans (920 scans for training and 235 for test), composed of seven partially labeled sub-datasets, involving seven organ and tumor segmentation tasks.

Partial-label task	Annotations		# Images	
	Organ	Tumor	Training	Test
#1 Liver	✓	✓	104	27
#2 Kidney	✓	✓	168	42
#3 Hepatic Vessel	✓	✓	242	61
#4 Pancreas	✓	✓	224	57
#5 Colon	×	✓	100	26
#6 Lung	×	✓	50	13
#7 Spleen	✓	×	32	9
Total	-	-	920	235

- BCV: 50 abdominal CT scans, 30 scans for training and 20 for test, Each training scan is paired with voxel-wise annotations of 13 organs

Experiments

Ablation Study

Table 3 – Comparison of dynamic head with different depth (#layers), varying from 2 to 4.

Depth	Average Dice	Average HD
2	71.30	**25.72**
3	**71.67**	25.86
4	71.63	26.07

Table 4 – Comparison of dynamic head with different width (#channels), varying from 4 to 8.

Width	Average Dice	Average HD
4	69.79	30.40
8	**71.67**	**25.86**
16	71.45	26.31

Table 5 – Comparison of the effectiveness of different conditions (image feature, task encoding) during the dynamic filter generation.

Image feat.	Task enc.	Average Dice	Average HD
✓	✓	**71.67**	**25.86**
×	✓	71.26	29.38
✓	×	51.80	79.94

Experiments

Comparison with SOTA

Methods	Task 1: Liver				Task 2: Kidney				Task 3: Hepatic Vessel			
	Dice		HD		Dice		HD		Dice		HD	
	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor
Multi-Nets	96.61	61.65	4.25	41.16	96.52	74.89	1.79	11.19	63.04	72.19	13.73	50.70
TAL [9]	96.18	60.82	5.99	38.87	95.95	75.87	1.98	15.36	61.90	72.68	13.86	43.57
Multi-Head [3]	96.75	64.08	3.67	45.68	96.60	79.16	4.69	13.28	59.49	69.64	19.28	79.66
Cond-NO	69.38	47.38	37.79	109.65	93.32	70.40	8.68	24.37	42.27	69.86	93.35	70.34
Cond-Input [2]	96.68	65.26	6.21	47.61	96.82	78.41	1.32	10.10	62.17	73.17	13.61	43.32
Cond-Dec [6]	95.27	63.86	5.49	36.04	95.07	79.27	7.21	8.02	61.29	72.46	14.05	65.57
DoDNet	96.87	65.47	3.35	36.75	96.52	77.59	2.11	8.91	62.42	73.39	13.49	53.56

Methods	Task 4: Pancreas				Task 5: Colon		Task 6: Lung		Task 7: Spleen		Average score	
	Dice		HD		Dice	HD	Dice	HD	Dice	HD	Dice↑	HD↓
	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Tumor	Tumor	Organ	Organ		
Multi-Nets	82.53	58.36	9.23	26.13	34.33	103.91	54.51	53.68	93.76	2.65	71.67	28.95
TAL [9]	81.35	59.15	9.02	21.07	48.08	66.42	61.85	39.92	93.01	3.10	73.35	23.56
Multi-Head [3]	83.49	61.22	6.40	18.66	50.89	59.00	64.75	34.22	94.01	3.86	74.55	26.22
Cond-NO	65.31	46.24	36.06	76.26	42.55	76.14	57.67	102.92	59.68	38.11	60.37	61.24
Cond-Input [2]	82.53	61.20	8.09	31.53	51.43	44.18	60.29	58.02	93.51	4.32	74.68	24.39
Cond-Dec [6]	77.24	55.69	17.60	48.47	51.80	63.67	57.68	53.27	90.14	6.52	72.71	29.63
DoDNet	82.64	60.45	7.88	15.51	51.55	58.89	71.25	10.37	93.91	3.67	75.64	19.50

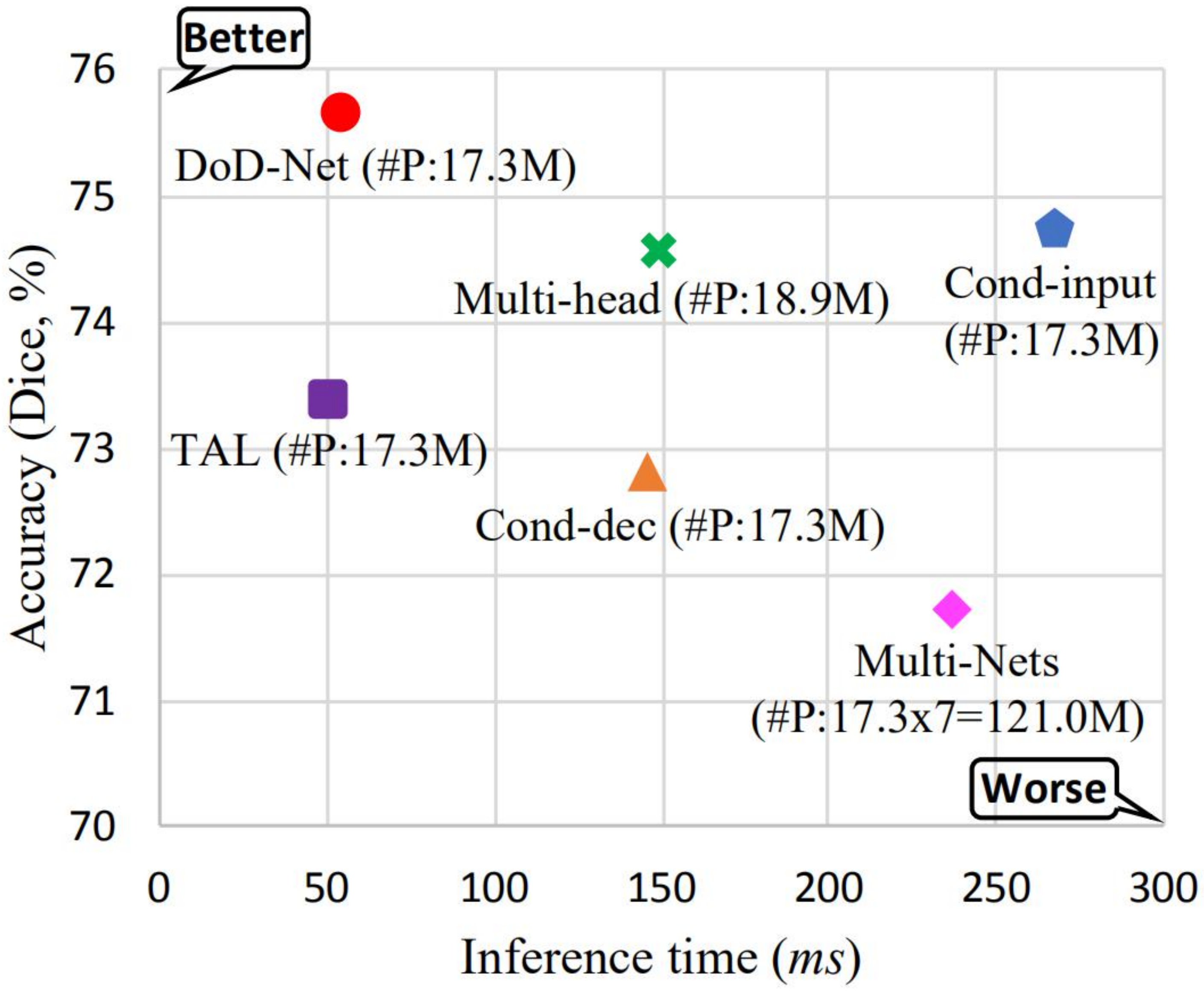
(1) seven individual networks, each being trained on a partially dataset (denoted by Multi-Nets),
(2) two multi-head networks (i.e., MultiHead [3] and TAL [9]),
(3) a single-network method without the task condition (Cond-NO),
(4) two single-network methods with the task condition (i.e., Cond-Input [2] and Cond-Dec [6]).

Experiments

Comparison with SOTA

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3 – Visualization of segmentation results obtained by different methods. (a) input image; (b) ground truth; (c) Multi-Nets; (d) TAL [9]; (e) Multi-Head [3]; (f) Cond-Input [2]; (g) Cond-Dec [6]; (h) DoDNet.



Experiments

MOTS pre-training for downstream tasks

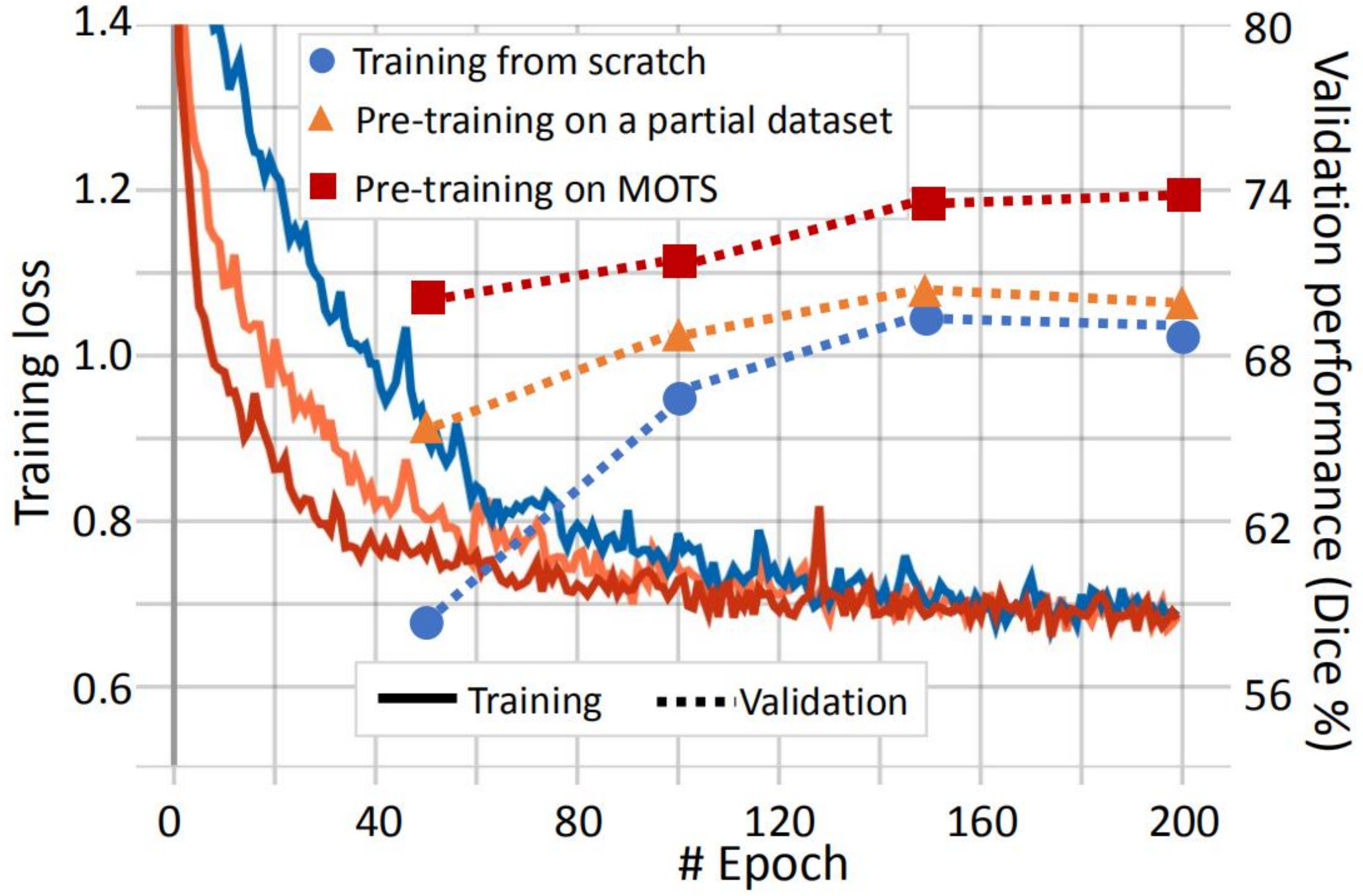


Table 7 – Comparison of state-of-the-art methods on the BCV test set. SD: Mean surface distance (lower is better); TFS: Training network from scratch; MOTS: Pre-training on MOTS. The values of three metrics were averaged over 13 categories.

Methods	Avg. Dice	Avg. SD	Avg. HD
Auto Context [27]	78.24	1.94	26.10
DLTK [24]	81.54	1.86	62.87
PaNN [42]	84.97	1.45	18.47
nnUnet [16]	**88.10**	1.39	17.26
TFS	85.30	1.46	19.67
MOTS	86.44	**1.17**	**15.62**