Supplemental Material for "Surface-GCN: Learning Interaction Experience for Organ Segmentation in 3D Medical Images"

Fengrui Tian[a,b], Zhiqiang Tian[a], Zhang Chen[a], Dong Zhang[b,c], Shaoyi Du[b]

[a]School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[b]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
[c]School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Version typeset January 29, 2023

Corresponding author: Zhiqiang Tian and Shaoyi Du. Email: zhiqiangtian@xjtu.edu.cn, dushaoyi@gmail.com

Contents

I. More Experiments on PROMISE12

I.A. More Results for Automatic Segmentation

Fig. 1 shows 15 subjects of quantitative segmentation results from PROMISE12 test dataset. Noting that the prostate in apex and base regions of the entire MR volume is not easy to segment, all prostate MR subjects are divided into 3 parts for fair: the apex subregion (30%), mid subregion (40%), and base subregion (30%), respectively. Each subregion shows 5 slices from different subjects. The magenta contours are the ground truths manually labeled by the radiologists, and the yellow contours are the prediction results from the proposed Surface-GCN. The DSC, RVD, HD, and ASD scores are also labeled on each segmentation result. It could be observed that the proposed method achieves accurate results from all three subregions. The prostate segmentation results on the entire PROMISE12 test set are shown in Table 1.

I.B. Interactive Segmentation Results

Fig. 2b shows the IOU score for one prostate subject in the image volume *vs.* the first 20 interactions on PROMISE12 test set. It can be observed that all three prostate subregions reach higher segmentation accuracies with the increase of interactions. The IOU scores increase from 88.9% to 90.1% of the whole gland after 7 clicks and increases from 90.1% to 91.1% after the rest 13 corrections.

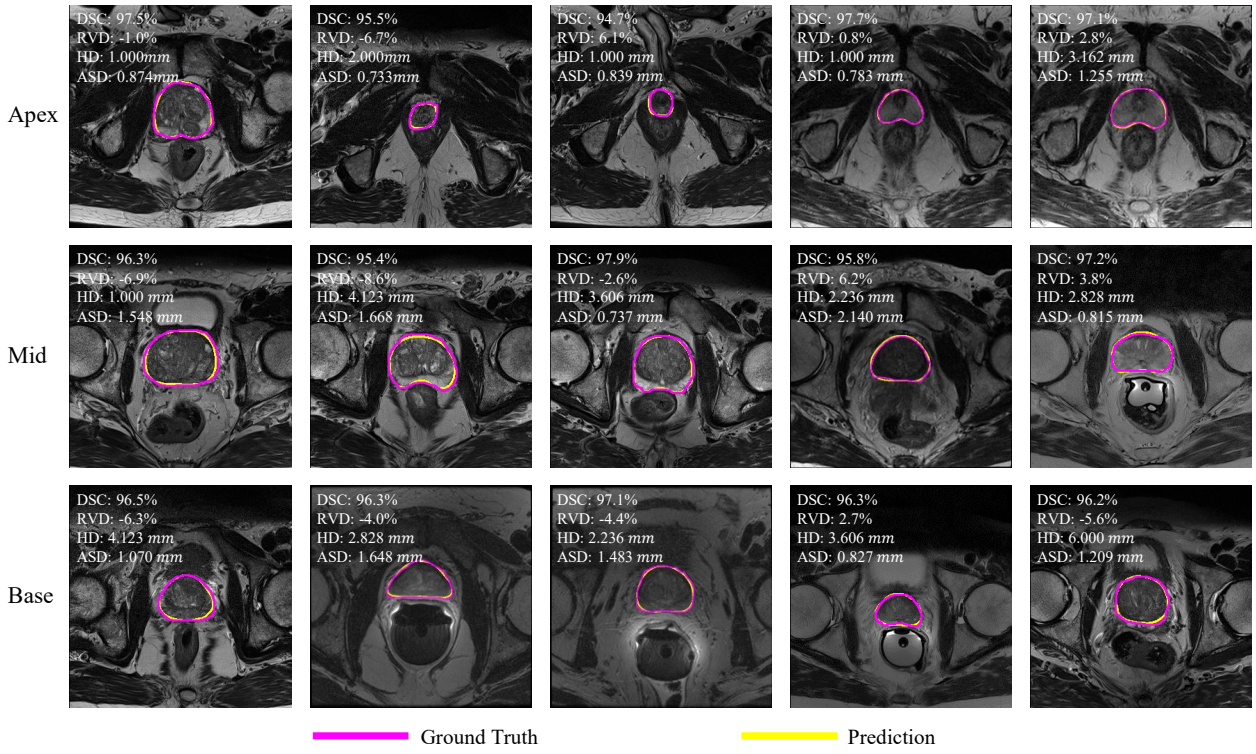
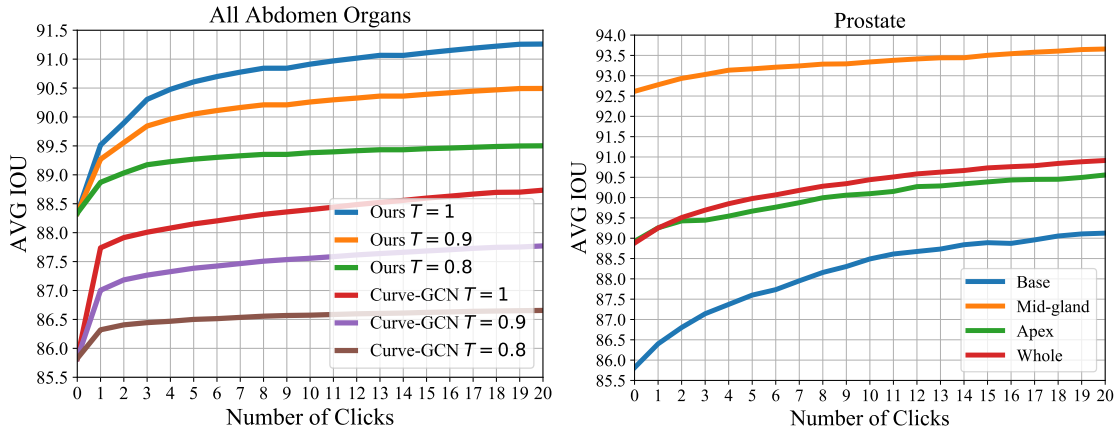


Figure 1: Qualitative results on PROMISE12 test dataset. Magenta contours are the ground truths and the yellow contours are the automatic segmentation results from the proposed Surface-GCN.



(a) (b)

Figure 2: (a) shows the averages on four abdominal organs at different thresholds T. (b) presents the number of clicks *vs.* IOU improvement on the entire prostate image volume.

Table 1: Quantitative results on PROMISE12 test dataset.

CaseID	DSC↑ (%)				RVD (%)				HD↓ (mm)				ASD↓ (mm)			
	Whole	Base	Mid	Apex	Whole	Base	Mid	Apex	Whole	Base	Mid	Apex	Whole	Base	Mid	Apex
1	92.44	92.42	96.17	88.46	-7.18	-7.16	-6.32	-8.05	3.55	1.80	2.60	6.24	1.94	1.77	1.51	2.54
2	93.22	94.67	95.76	88.28	-8.06	-6.15	-5.90	-12.68	3.83	3.81	2.38	5.68	1.69	1.31	1.37	2.46
3	94.35	92.95	96.43	93.59	-7.23	-8.27	-4.64	-8.77	4.07	4.41	3.50	4.31	1.50	1.73	1.30	1.46
4	94.06	91.28	94.98	95.67	-6.85	-6.48	-8.50	-5.28	5.15	5.44	5.32	4.66	1.84	2.33	1.93	1.23
5	95.36	93.48	97.18	95.68	-3.97	-5.21	-3.69	-2.95	2.08	2.45	2.09	1.71	1.09	1.30	0.94	1.02
6	96.25	95.72	97.67	95.08	-1.11	-4.69	-0.10	1.29	2.54	1.85	1.80	4.08	1.08	1.00	0.88	1.41
7	95.43	95.17	95.65	95.38	-3.77	-8.29	-2.24	-1.28	3.11	3.04	2.75	3.65	1.14	1.01	1.28	1.08
8	94.30	91.85	95.36	95.83	-2.45	1.79	-6.83	-3.19	4.48	3.87	6.14	3.75	2.20	2.66	2.09	1.83
9	93.92	93.86	95.96	91.88	3.94	-2.43	2.06	12.21	4.17	3.72	4.36	4.44	2.74	2.42	2.28	3.53
10	93.54	90.84	96.01	94.14	-2.96	3.79	-3.45	-9.32	6.53	8.37	5.57	5.46	2.70	3.42	1.68	2.80
11	93.77	91.18	96.32	94.11	-3.81	-13.71	-3.65	5.97	4.96	8.31	4.60	1.92	2.97	3.46	2.11	3.20
12	92.07	90.35	97.03	87.74	12.65	14.20	-0.39	25.78	7.33	8.06	4.48	9.82	4.20	4.86	2.09	5.91
13	92.05	90.90	95.07	89.41	0.31	-13.52	-0.28	14.87	4.98	6.02	5.38	3.46	2.71	3.10	2.16	2.98
14	95.50	93.47	97.19	96.06	-3.05	-7.64	-0.23	-0.81	4.06	5.56	3.15	3.32	2.00	2.58	1.52	1.82
15	94.45	94.05	96.91	92.95	3.72	-0.11	-0.29	10.55	5.40	4.87	2.28	8.28	2.21	2.05	1.39	2.99
16	94.13	93.19	96.52	93.10	-3.72	0.54	-0.21	-10.78	4.51	4.20	3.48	5.63	2.35	2.21	1.81	2.92
17	95.39	95.82	97.02	92.82	-0.50	-2.85	-1.10	2.58	3.47	2.93	2.33	5.45	1.71	1.41	1.48	2.29
18	95.84	96.19	97.30	93.59	-4.35	-2.85	-1.66	-9.20	4.91	4.61	4.49	5.73	2.11	1.46	1.63	3.35
19	95.55	96.75	96.44	93.16	-2.15	-2.07	-4.33	0.51	3.37	2.59	4.44	2.81	1.88	1.32	2.00	2.28
20	94.75	93.66	95.96	94.35	0.08	-2.41	5.46	-3.88	4.17	3.76	5.60	2.86	2.00	1.52	2.15	2.31
21	95.05	94.53	97.11	93.98	-1.03	3.29	0.24	-6.31	6.57	6.24	5.25	7.89	1.94	1.94	1.48	2.28
22	94.06	93.98	95.59	92.56	-4.56	-6.24	-3.50	-3.95	6.15	5.66	6.68	6.10	2.59	2.10	2.23	3.45
23	93.66	93.25	95.50	92.18	1.94	-6.96	-0.28	13.06	2.50	2.40	2.10	3.00	1.27	1.13	1.09	1.58
24	94.39	94.99	94.59	93.60	-4.48	-1.70	-9.66	-3.12	2.43	3.02	2.02	2.17	1.23	0.86	1.53	1.37
25	95.63	95.29	96.97	94.14	1.71	1.37	-0.54	5.06	2.42	1.82	2.79	2.54	0.87	0.89	0.71	1.06
26	95.04	94.57	96.33	93.87	2.11	-2.09	-0.27	9.30	2.72	1.72	2.39	4.12	1.29	0.99	1.10	1.81
27	94.84	92.21	96.77	95.44	3.92	4.12	1.93	5.71	2.32	3.02	2.27	1.66	1.21	1.39	0.90	1.34
28	95.03	94.23	97.06	94.27	1.54	-3.51	1.53	6.60	2.30	3.42	1.67	1.65	0.97	1.02	0.66	1.16
29	94.67	92.84	95.81	94.93	3.17	5.29	-2.33	8.36	1.78	1.86	1.56	2.00	1.12	1.23	1.08	1.06
30	95.85	95.65	97.21	94.18	2.26	4.86	1.21	1.06	1.71	1.80	1.10	2.43	0.84	0.71	0.61	1.27
Avg.	94.49	93.65	96.33	93.35	-1.13	-2.50	-1.93	1.11	3.92	4.02	3.49	4.23	1.85	1.84	1.50	2.19
Std.	1.09	1.72	0.79	2.23	3.27	5.87	3.34	8.91	1.52	1.97	1.57	2.08	0.76	0.95	0.51	1.07

II. More Experiments on Abdomen Datasets

II.A. More Automatic Segmentation Results for Abdomen Datasets

Six subjects of quantitative segmentation results are shown in Columns 1-3 in Fig. 3. Column 4 shows the segmentation overlaid on the entire CT. Blue, green, orange, and cyan contours denote the segmentation results of the left kidney, gallbladder, spleen, and esophagus respectively. The magenta contours denote the corresponding ground truths. The DSC, 95%HD, and ASD scores are labeled on each result.

II.B. More Experiments about Interactive Segmentation

Fig. 2a shows the IOU averages on four abdominal organs at different thresholds T. Table 2 shows the interactive segmentation details of the proposed method. "+One click" denotes the segmentation results after one correction. "+Max IOU" means the maximum accuracy that can be achieved by using interactive clicks. It is believed that the first click correction is more important than the rest clicks[1], especially when the automatic prediction accuracy is not enough. For the gallbladder and the esophagus, our method reaches a greater segmentation improvement than Curve-GCN[2] prediction after one click. It demonstrates that the proposed method has better performance on capturing small organ features.

Table 2: Interactive segmentation IOU scores (%) on abdominal multi-organ CT images. "+One click" denotes the segmentation results after one correction. "+Max IOU" means the maximum accuracy that can be achieved by using interactive clicks.

Method	Spl.	L. Kid.	Gall.	Esoph.	Avg.
Curve-GCN[2]	88.82	88.76	82.82	78.02	85.82
+One click	+1.39	+0.94	+1.33	+1.61	+1.27
+Max IOU	91.95	91.44	85.83	81.28	88.80
Ours	91.88	90.75	85.36	80.57	88.34
+One click	+0.90	+0.88	**+1.51**	**+1.90**	+1.18
+Max IOU	**93.87**	**92.89**	**89.22**	**85.69**	**91.26**

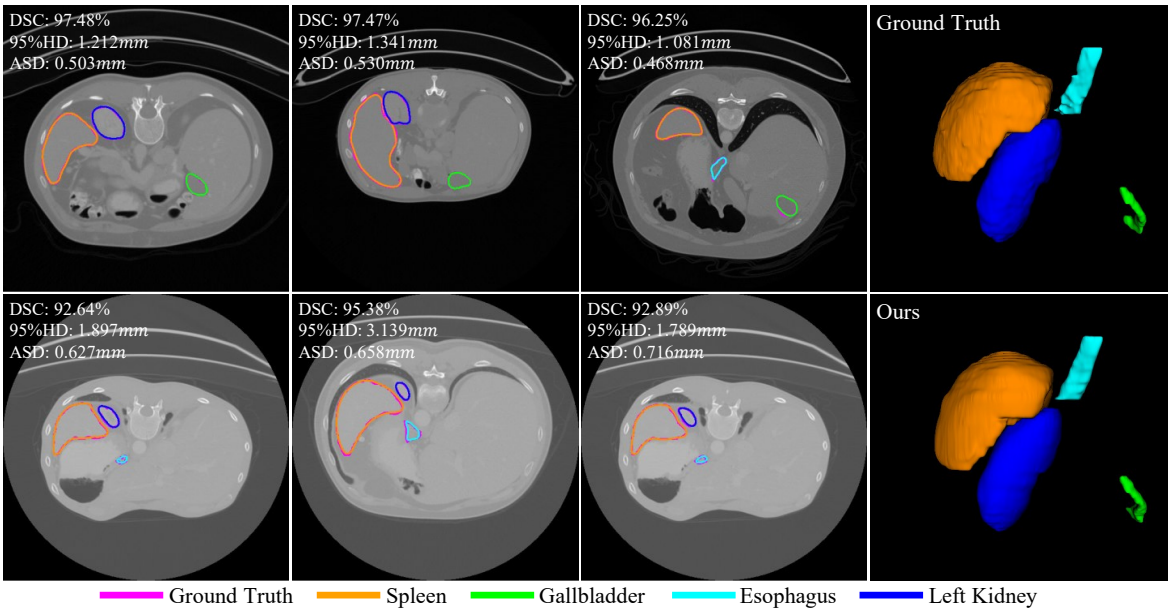


Figure 3: Columns 1-3: Qualitative results for abdominal segmentation. Orange, green, cyan, and blue contours are segmentations of the spleen, gallbladder, esophagus, and left kidney, separately. Magenta contours are the corresponding ground truths. The mean scores of DSC, 95%HD, and ASD in each slice are also labeled. Column 4: Segmentations overlaid on CT.

III. Discussion about Surface Convolution Unit

The key of the proposed Surface-GCN is the Surface Convolution Unit (SCU), because it fits the organ surface to capture boundary information. To find the best shape of the surface convolution kernel, six different types of SCU vertex connections are presented in Fig. 4. Similar to the Fig. 5 in the main body of this paper, take the red vertex for example. Vertices from left to right are in 5 different slices. Blue vertices are in the same slice of the red vertex, and orange vertices are in the neighboring slices. Blue and orange connections denote the intra- and inter-dependency connections, separately. Fig. 4a denotes the connection type that is used in this paper, and Fig. 4b, Fig. 4c, Fig. 4d, Fig. 4e, and Fig. 4f are other 5 different situations of SCU connections. In Fig. 4b, the number of inter-dependency connections is reduced. To evaluate the importance of the vertices dropped in Fig. 4b, in Fig. 4c, the connections of these parts of vertices are weakened by setting the corresponding weight in the adjacent matrix as 0.1, which is represented by orange dotted lines. In Fig. 4d, we tried to extend the range of inter-dependency connections by establishing connections with more vertices. However, in our experiment, when the weights of these connections in the adjacent matrix set 1, the proposed model becomes hard to converge. One possible reason is that

Table 3: Comparisons of different types of SCU vertex connections (PROMISE12 test dataset).

Type	DSC↑ (%)	RVD (%)	HD↓ (mm)	ASD↓ (mm)
Fig. 4a	**94.49**	-1.13	**3.92**	**1.85**
Fig. 4b	94.41	-0.38	4.19	1.89
Fig. 4c	94.20	-0.36	4.11	1.92
Fig. 4d	94.13	-0.89	4.01	1.96
Fig. 4e (W/o inter)	94.06	1.35	4.14	1.98
Fig. 4f (W/o intra)	94.29	**-0.13**	4.05	1.90

these vertices are so far from the red vertex and could not transfer useful information by the connections. So the weights are changed to 0.1, which is denoted by orange dotted lines in Fig. 4d. Fig. 4e and Fig. 4f are the situations without inter-dependency connections and intra-dependency connections, respectively. The above situations are tested in the proposed model and the results are listed in Table 3. From the results, it can be concluded as follows.

1. SCU shows robustness on different types of vertex connections.
2. Establishing inter- and intra-dependency connections reaches an accurate segmentation result.
3. Fig. 4d shows that connections cannot be established between two far vertices.
4. The comparison of Fig. 4a, Fig. 4b, and Fig. 4c indicates that the connections established by 2 closed vertices are statistically significant for higher segmentation accuracy.
5. The comparison of Fig. 4e and Fig. 4f concludes that connections from the neighboring slices are more important than from the identical slices.

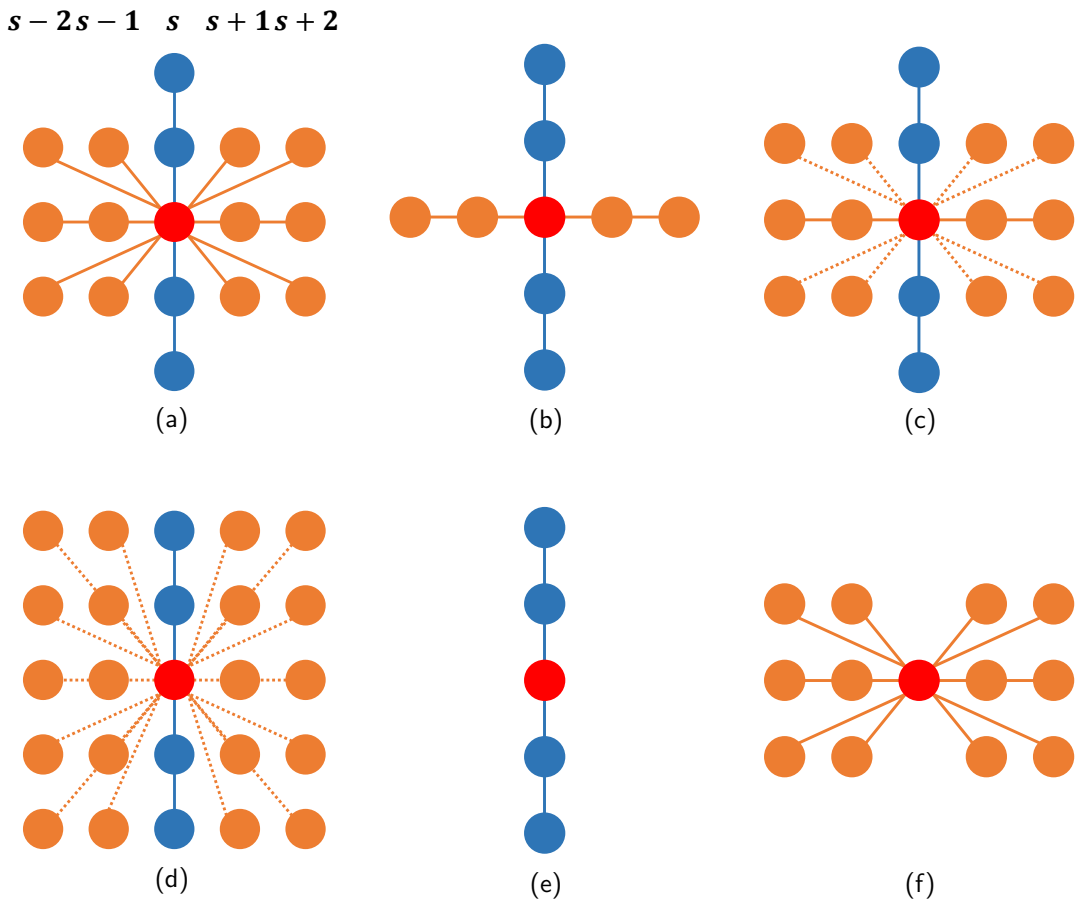


Figure 4: Six different types of SCU vertex connections. In (a)-(f), vertices from left to right columns are in 5 transverse slices. Similar to the Fig. 5 in the main body of this paper, take the red vertex for example. The blue vertices and the red vertex are adjacent in the same slice. The orange vertices are adjacent in different slices. The blue and orange connections are the intra- and inter-dependency connections, respectively. Moreover, the orange solid line denotes the strong connection that weights 1.0 in adjacent matrix, and the orange dotted lines denote the weak connection that only weighs 0.1 in adjacent matrix. (a) presents the GCN connection of Surface-GCN. (b), (c), and (d) are another 3 different types of vertex connections for comparison. (e) and (f) are the situations without inter-dependency connections and without intra-dependency connections, respectively. The evaluation results are listed in Table 3.

IV. Discussion about the Evaluation Procedure of Interactive Segmentation

The simulation procedure "select the point with the largest distance from the ground truth" may be too strict to imitate the radiologists' real operation when existing several suspicious points. We hence test the performance with the following strategy: train and test the model by randomly selecting one point from the top k points with the largest distance. This strategy may be more appropriate to the real situation and Table 4 shows the IOU score of one imitated click using different k values:

Table 4: Comparisons of different evaluation procedures of interactive segmentation. DSC (%) scores are reported in the table.

Method	Spl.	L. Kid.	Gall.	Esoph.
Auto seg	91.88	90.75	85.36	80.57
k=1	92.78	91.63	86.87	82.47
k=2	92.61	91.58	86.48	82.34
k=3	92.59	91.50	86.43	82.30

According to Table 4, we could see that larger k values slightly damage the performance while using one click. However, with the growth of the imitated clicks, we further found that the final segmentation reaches comparable performance with different k values. So in this way, radiologists' corrections on organ boundaries may achieve similar results by using our model.

References

[1] Z. Lin, Z. Zhang, L.-Z. Chen, M.-M. Cheng, and S.-P. Lu, Interactive image segmentation with first click attention, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[2] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, Fast interactive object annotation with Curve-GCN, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5252–5261, 2019.