Supplementary of "AbdomenCT-1K: Is Abdominal Organ Segmentation A Solved Problem?"

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, Xiaoping Yang

This supplementary includes the detailed dataset information quantitative results in terms of HD95, Jaccard, and ASD, and the violin plots of DSC and NSD.

TABLE 1: Detailed sub-dataset information in AbdomenCT-1K.

Information	Organ	LiTS Plus	KiTS Plus	Spleen Plus	NIH-Pancreas Plus	MSD Pancreas Plus	NJU
Volume mean (min, max)	Liver	173 (58.3,334)	180 (32.5,385)	172 (113, 343)	159 (94.0,27.3)	157 (50.5,450)	105 (70.0,153)
	Kidney	37.0 (10.4,99.8)	53.9 (18.9,195)	38.7 (2.7, 63.7)	34.0 (19.0,54.7)	37.8 (14.9,153)	30.3 (6.0,51.9)
	Spleen	26.3 (5.4,126)	26.6 (0,93.7)	25.2 (6.2,52.3)	23.7 (3.9,104)	24.8 (0,94.8)	23.6 (6.2,62.9)
	Pancreas	9.1 (2.1,17.5)	9.4 (1.2,133)	8.9 (3.7,15.4)	7.3 (4.2,15.0)	9.6 (2.2,86.8)	6.4 (0.4,19.7)
HU mean (min, max)	Liver	99.3 (28.7,164.5)	80.4 (-2.1,148)	94.1 (38.9,166)	95.9 (46.6,140)	121 (52.6,187)	76.8 (37.8,210)
	Kidney	136 (60.5,236)	103 (24.3,206)	112 (16.1, 227)	140 (75.7,213)	161 (33.9,268)	84.1 (23.9,230)
	Spleen	106 (75.5,174)	95.7 (30.0,191)	92.0 (29.7,150)	107 (66.2,151)	121 (43.7,195)	87.6 (36.8,250)
	Pancreas	71.5 (-16,126)	69.1 (-11.6,145)	53.6 (-16.2,135)	86.4 (37.1,127)	80.4 (-19.3,152)	50.3 (3.3,143)
# Slices mean (min, max)	-	448 (74,987)	216 (29,1059)	89.0 (31, 168)	237 (181,466)	95.1 (37,751)	211 (70,437)
Slice Thicknes mean (min, max)	-	1.5 (0.7,5.0)	3.2 (0.5,5.0)	4.4 (1.5, 8.0)	0.99 (0.5,1.0)	2.9 (0.7,7.5)	2.6 (1.3,5.0)
Scanner	-	-	-	-	Philips and Siemens	-	GE

TABLE 2: Supplemental results of multi-organ segmentation results in fully supervised benchmark (corresponding to Table 5 of the main paper).

Training	Testing	Liver			Kidney			Spleen			Pancreas		
		HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)
MSD Pan. Plus (281), NIH Pan. Plus (80) Subtask 1: 361 cases	100 cases	31.4 ±51.7	92.5 ±8.5	4.3 ±6.9	46.8 ±105.0	75.0 ±19.5	8.5 ±17.4	44.8 ±83.3	84.0 ±18.8	9.6 ±22.0	71.4 ±70.0	52.0 ±23.6	21.9 ±24.8
MSD Pan. Plus (281) LiTS Plus (40) KiTS Plus (40) Subtask 2: 361 cases		20.8 ±45.8	94.2 ±5.0	3.0 ±5.7	13.7 ±20.8	86.2 ±15.4	2.5 ±4.5	15.6 ±54.2	89.7 ±15.3	4.3 ±18.5	29.0 ±37.5	66.3 ±18.1	7.3±12.7

TABLE 3: Supplemental results of multi-organ segmentation results in semi-supervised benchmark (corresponding to Table 7 of the main paper).

Task	Liver			Kidney			Spleen			Pancreas			Average		
	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)
Lower Bound	21.8±40.0	92.2±8.5	4.3±9.2	12.8±21.6	86.1±17.5	1.1±1.7	15.5±38.0	89.5±14.6	2.7±6.9	10.0±13.9	70.9±17.2	1.7±2.2	15.0±30.6	84.7±17.0	2.5±6.0
Subtask 1	15.1±32.4	92.9±7.0	2.9±7.2	11.6±17.6	86.1±16.0	1.08±1.5	4.2±8.1	91.2±13.0	0.8±1.2	8.2±10.2	72.4±16.0	1.4±1.2	9.8±19.9	85.7±15.7	1.6±3.8
Subtask 2	23.0±45.3	92.9±6.7	4.3±9.1	10.4±17.0	87.2±15.6	0.9±1.3	11.7±32.9	91.4±12.6	1.6±3.0	8.9±11.9	72.4±15.3	1.7±1.7	13.5±30.3	86.0±15.3	2.1±5.1
Upper Bound	8.3±22.3	95.1±4.0	1.6±4.4	8.5±15.1	91.4±6.5	1.3±2.7	10.1±31.4	93.2±10.7	1.2±2.6	6.4±8.3	75.9±11.7	1.5±1.6	8.3±21.0	88.9±11.6	1.4±3.0

TABLE 4: Supplemental results of multi-organ segmentation results in weakly supervised benchmark (corresponding to Table 8 of the main paper).

Training	Ratio	Testing	HD95 (mm)	Jaccard (%)	ASD (mm)
Spleen Plus (41)	5%	100 cases	44.0 ±51.6	68.4 ±25.6	9.7 ±15.1
	15%		36.0 ±49.5	75.5 ±21.7	7.5 ±12.8
	30%		35.7 ±50.5	76.5 ±20.8	7.7 ±13.9

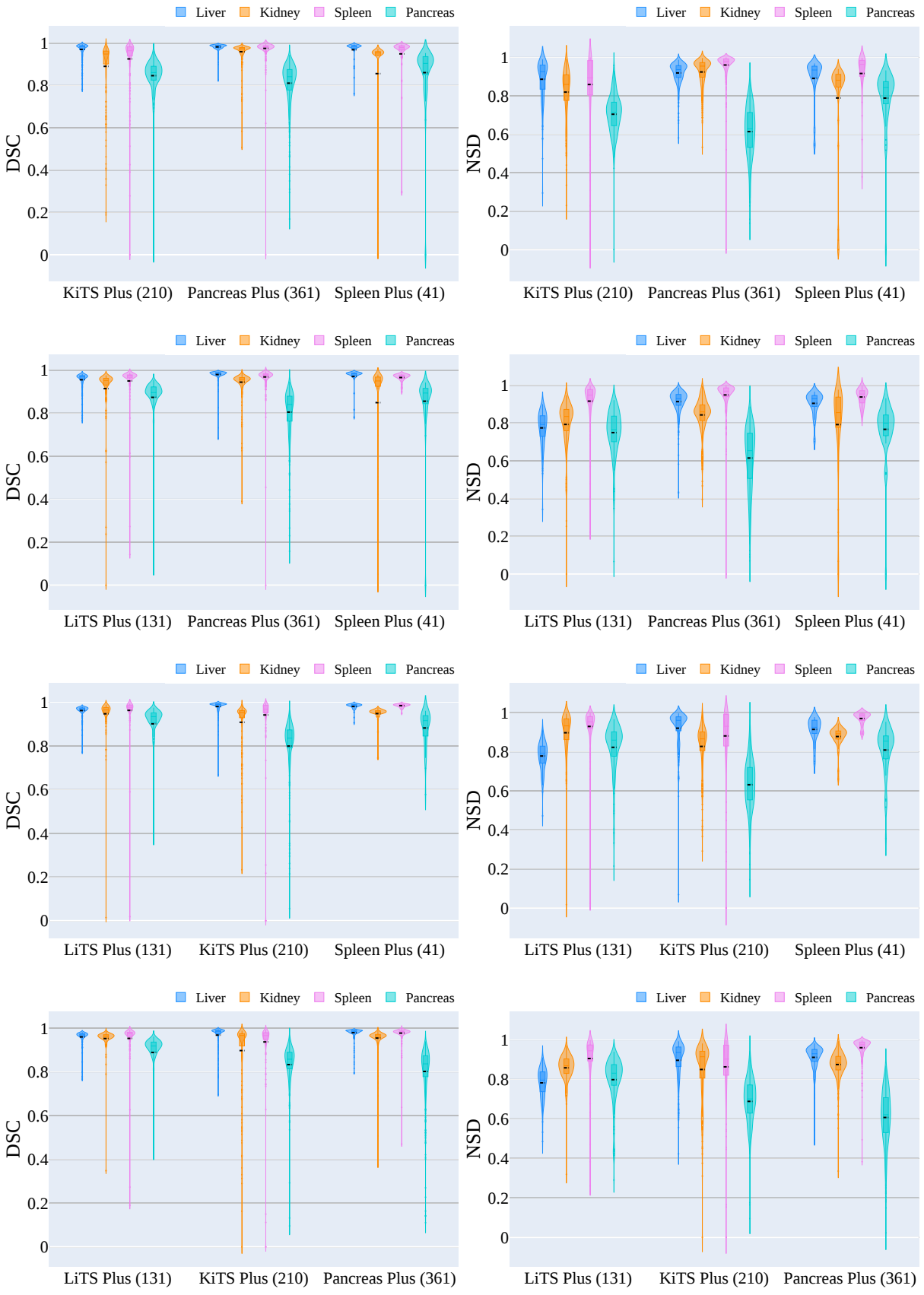

Fig. 1: Violin plots of the performances (DSC and NSD) of different organ segmentation in large-scale fully supervised multiple abdominal organ segmentation tasks.

TABLE 5: Supplemental results of multi-organ segmentation results in weakly supervised benchmark (corresponding to Table 9 of the main paper).

Task	Method	Liver			Kidney			Spleen			Pancreas		
		HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)
5% labels	2D U-Net	56.9±53.0	86.7±10.4	12.3±13.9	40.7±34.2	70.7±23.1	6.8±6.2	47.7±78.0	74.0±24.7	13.2±25.0	30.6±17.3	42.3±17.9	6.6±5.4
	2D U-Net + CRF	51.6±52.2	86.9±9.9	11.2±13.5	40.0±34.8	68.0±23.1	6.4±6.2	47.2±77.5	74.0±25.6	12.7±24.8	29.6±17.0	40.6±18.3	5.9±5.2
15% labels	2D U-Net	51.3±53.6	88.3±9.9	11.0±12.9	32.0±34.8	77.2±21.8	4.9±5.4	39.6±71.2	82.3±18.4	9.8±20.5	21.2±15.8	54.4±18.1	4.2±3.7
	2D U-Net + CRF	46.3±52.4	88.7±9.5	10.2±12.6	31.0±34.8	74.7±21.7	4.5±4.7	38.6±71.1	82.9±18.9	9.4±20.1	20.6±16.4	53.2±19.0	3.8±3.6
30% labels	2D U-Net	51.2±55.6	88.5±9.9	11.2±13.5	34.7±35.7	78.3±20.6	5.7±6.5	38.1±71.4	82.3±17.8	10.5±22.3	18.7±15.7	56.7±18.0	3.6±3.4
	2D U-Net + CRF	47.0±53.7	88.8±9.4	10.3±13.1	33.2±35.8	75.7±20.6	5.2±5.9	37.5±71.4	82.9±18.2	10.1±22.0	18.2±15.5	55.5±18.8	3.2±3.3

TABLE 6: Supplemental results of average organ segmentation results in continual learning benchmark (corresponding to Table 10 of the main paper).

Training		Testing		HD95 (mm)	Jaccard (%)	ASD (mm)
Dataset	Annotation	Dataset	Annotation			
MSD Pancreas Ts (139) KiTS (210) Spleen (41) MSD Pancreas (281)	Liver Kidney Spleen Pancreas	100 cases	Liver, kidney, spleen, and pancreas	35.4±52.1	72.1±25.5	8.1±17.1

TABLE 7: Supplemental results of multi-organ segmentation results in continual learning benchmark (corresponding to Table 11 of the main paper).

Organ	HD95 (mm)	Jaccard (%)	ASD (mm)
Liver	21.2±42.9	90.8±11.2	3.0 ±5.7
Kidney	31.7 ±28.8	69.3 ±22.9	3.8 ±6.9
Spleen	42.2 ±77.6	77.1 ±25.7	10.4 ±24.3
Pancreas	46.2 ±43.5	51.2 ±22.0	15.3 ±20.1

TABLE 8: Supplemental results on the common testing set of the four benchmarks (corresponding to Table 12 of the main paper).

Task		Liver			Kidney			Spleen			Pancreas			Average		
		HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)	HD95 (mm)	Jaccard (%)	ASD (mm)
Fully Supervised	Subtask 1	7.0±8.4	92.6±8.9	1.2±1.2	4.6±7.2	90.8±10.1	0.8±0.8	60.9±75.1	79.5±23.5	18.4±28.7	29.8±37.3	66.5±25.6	10.0±22.4	25.6±47.7	82.4±21.2	7.6±19.5
	Subtask 2	4.8±5.7	95.1±4.5	0.7±0.8	1.9±3.5	95.2±6.1	0.4±0.6	9.6±29.7	95.2±8.0	1.3±3.8	26.5±37.1	73.5±21.2	7.0±18.3	10.7±25.7	89.7±15.1	2.4±9.7
Semi-Supervised	Subtask 1	10.3±25.0	94.0±2.5	1.1 ±1.7	3.2 ±6.7	92.6 ±7.0	0.44 ±0.27	7.7 ±22.2	89.1 ±15.8	1.6 ±4.8	19.6 ±17.2	60.2 ±21.8	1.9±2.7	10.2±19.9	84.0±19.6	1.3±2.9
	Subtask 2	9.5±27.4	95.9±1.9	0.77 ±1.6	2.7 ±6.7	94.7 ±6.8	0.56 ±2.0	11.7 ±45.5	94.9 ±6.4	2.6 ±7.4	15.8 ±19.9	71.6 ±15.8	1.7 ±2.4	9.9±28.8	89.3±13.7	1.4±4.1
Weakly Supervised	Subtask 1	185±78.8	73.6±13.4	44.9±24.7	131±81.1	63.0±25.5	25.4±14.8	72.2±73.3	55.2±34.7	17.3±27.2	64.8±54.9	10.6±13.8	21.0±29.7	113±86.4	51.2±33.8	27.2±26.8
	Subtask 2	202±74.9	73.6±12.2	54.4±29.7	147±100	71.6±19.2	30.6±28.0	71.7±72.0	59.8±32.2	14.3±22.2	64.2±57.5	19.5±16.5	16.9±20.5	121±95.8	56.4±30.8	29.1±29.8
	Subtask 3	206±75.5	73.6±11.1	56.9±30.1	153±98.1	72.6±15.5	37.2±30.4	104±96.0	59.3±31.6	24.3±32.6	108±82.4	20.2±17.3	34.0±45.6	143±97.1	56.7±30.0	38.1±37.0
Continual Learning		9.0±11.8	89.1±13.9	1.4±1.7	8.8±13.3	84.6±18.2	1.1±1.6	60.4±69.5	73.0±26.7	19.7±29.7	16.6±20.7	66.0±21.7	3.5±8.6	23.7±42.8	78.2±22.5	6.4±17.2

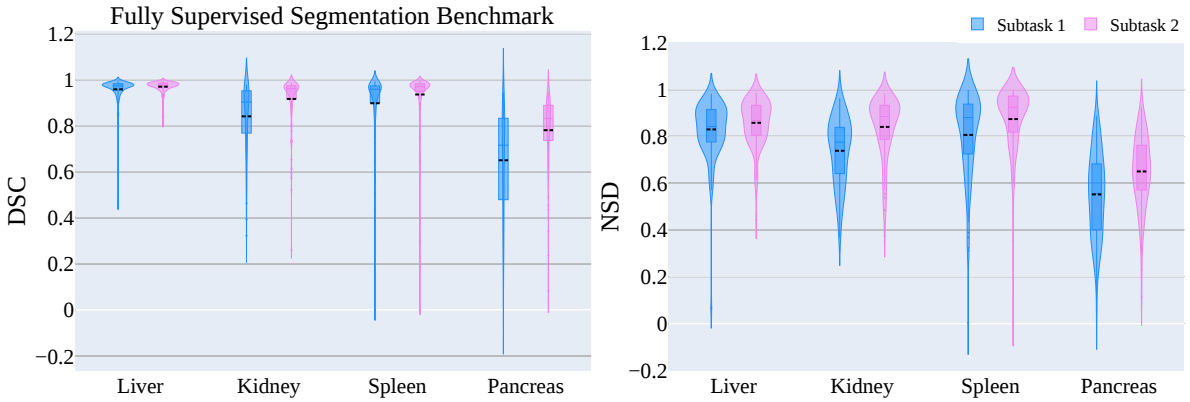

Fig. 2: Violin plots of the performances (DSC and NSD) of different organ segmentation results in fully supervised segmentation benchmark.

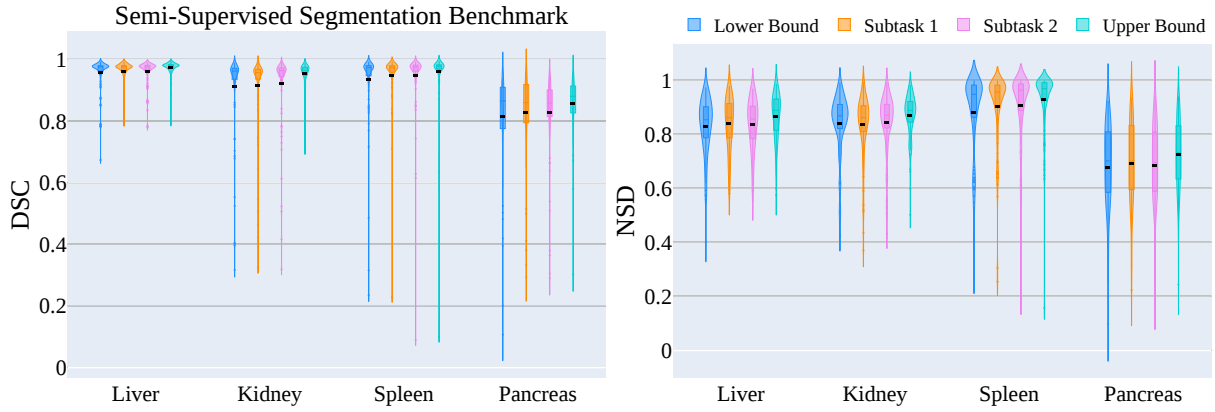

Fig. 3: Violin plots of the performances (DSC and NSD) of different organ segmentation results in semi-supervised segmentation benchmark.

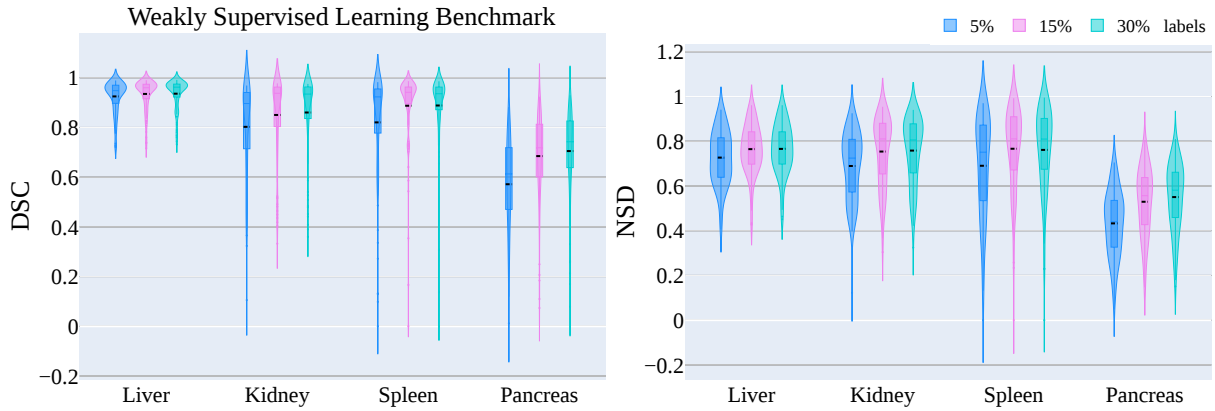

Fig. 4: Violin plots of the performances (DSC and NSD) of different organ segmentation results in weakly supervised segmentation benchmark.

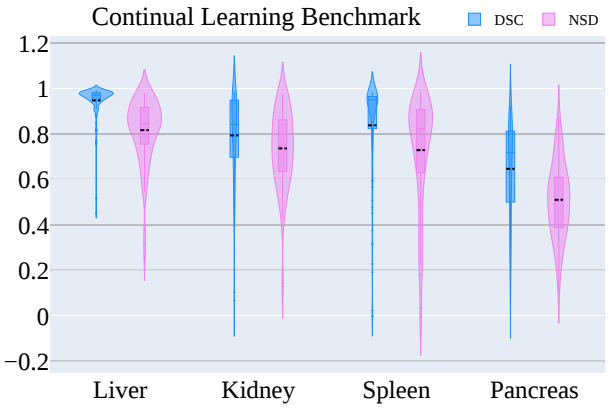

Fig. 5: Violin plots of the performance (DSC and NSD) of different organ segmentation results in continual learning benchmark.