Supplemental: One Sketch for All: One-Shot Personalized Sketch Segmentation

Anran Qi[1], Yulia Gryaditskaya[1,2], Tao Xiang[1], Yi-Zhe Song[1]
SketchX Lab, CVSSP[1], Surrey Institute for People-Centred AI[2]
University of Surrey

VI. Supplemental web-pages

We provide the supplemental web-pages that show the 5 templates, used for each category to compute the results in Table 1 in the main document, and the representative segmentation results for each method.

VII. One-shot segmentation

A. Alternative evaluation

In this section, we provide additional evaluation results to those in Section IV-D in the main document. We provide in Table VII a more restrictive evaluation on subsets of sketches that have the same set of labels as an exemplar. Compared to the evaluation in the main paper, numerical results in Table VII do not account for the cases when the target sketch has less parts and only a part of labels has to be transferred. This is the reason why in the main document we use a less restrictive evaluation strategy. It can be seen that similarly to the results in the main document our approach outperforms the alternative solutions.

The remaining experiments in this document use the evaluation strategy used in the main document.

B. Detailed numerical evaluation after label refinement

In Table VIII we provide the detailed numerical results per category. While on average our method outperforms competing approaches after refinement, our method is outperformed by ISPP method on the 'bulldozer' category and tightly follows FLSS on the 'suitcase' category. The worse performance of our method than the ISPP method on the 'bulldozer' category can be explained by the fact that we solve jointly for the keypoints and stroke-level transformations. In this case, the prediction of keypoints sometimes can degrade, resulting in the method not being able to correctly estimate the global reflection between the two sketches, *e.g.* 'bulldozer' facing right or left. In Section VIII-D we evaluate a separate training strategy, where the keypoints prediction network is trained separately. While separate training does increase the performance on the 'bulldozer' category by 15.3 points, in overall, the joint training strategy results in more stable performance across categories, showing better results on more categories. Please see Section VIII-D for the further comparison of these two strategies.

C. One-shot vs. few-shot

In the main paper we show in Table II that the performance improves if there are several templates available, and our results consistently outperform SGCN. Here in Table IX we show the numerical evaluation per category.

VIII. Ablation studies

A. ISPP: GCN vs PointNet++ encoder

Table X shows that when the PointNet++ encoder is used as was proposed in the original paper, the ISPP method performance on one shot sketch segmentation consistently drops: The point accuracy reduces on average over the five categories by 3.6 points, and the component accuracy – by 5 points.

B. Segmentation module

As we mention in Section IV-C in the main document:

> At inference, to obtain the labeling via Eq. 9, we first estimate our hierarchical deformation, then the label of a point v_i is obtained as follows $\tau(v_i) = \tau_{\theta_3}(v_i, \mathcal{F}^{stroke}_{s_j \in \hat{X}: v_i \in s_j}, \mathcal{F}^{sketch}_{\hat{E}})$.

Here we compare this strategy with the strategy of passing in an encoding of a globally warped target sketch $\mathcal{F}^{sketch}_{\hat{X}}$, instead of an encoding of a stroke-level warped exemplar $\mathcal{F}^{sketch}_{\hat{E}}$. Table XI shows that this strategy slightly loses the one we use in the main paper.

C. Chamfer distance in the stroke-level deformation

Finally, we evaluate the role of the Chamfer distance in Equation 6. Table XII shows the segmentation accuracy if the stroke level-deformation is guided only by the mean square distance between the keypoints of the deformed template $\hat{E}$ and the keypoints of the globally deformed sketch $\hat{X}$: $\mathcal{L}_{MSE}(K_{\hat{E}_{\hat{X}}}, K_{\hat{X}})$. It can be seen that using both losses $\mathcal{L}_{MSE}(K_{\hat{E}_{\hat{X}}}, K_{\hat{X}})$ and $\mathcal{L}_{CD}(\hat{X}, \hat{E}_{\hat{X}})$ gives a slight advantage over using the keypoints loss only.

D. Two steps training: Isolated training for keypoints

In this section we evaluate the overall performance of our method, if we train in two steps. First, we train a keypoints estimation module with our GCN sketch encoder. Then, we train the deformation and segmentation modules. In this case the GCN encoders are trained separately at each step. Table XIII

	P-metric										C-metric									
	SPGG [1]		SGCN [2]		FLSS [3]		ISPP [4]		Ours		SPGG [1]		SGCN [2]		FLSS [3]		ISPP [4]		Ours	
Category	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
airplane	22.1	8.6	66.0	14.6	56.7	12.3	51.1	13.5	**83.4**	8.7	21.8	7.8	56.7	17.5	38.3	15.0	26.0	15.3	**76.2**	12.7
alarm clock	23.0	5.3	81.8	11.4	60.7	12.9	59.5	13.7	**86.5**	12.3	8.7	2.0	72.5	15.9	42.8	11.8	36.1	17.8	**77.1**	19.7
ambulance	28.5	10.3	76.6	7.2	62.0	11.9	60.7	11.7	**85.8**	6.3	15.1	7.2	64.8	11.0	49.6	8.4	32.5	9.1	**79.0**	10.2
angel	1.8	1.8	57.2	9.8	45.0	19.0	61.3	5.4	**71.7**	14.7	1.9	1.9	46.3	13.2	25.3	17.0	33.0	10.7	**65.1**	13.2
ant	7.5	4.7	44.8	19.5	42.4	17.1	47.6	11.5	**61.8**	22.1	9.9	6.1	34.5	13.4	23.4	11.7	23.0	14.2	**46.4**	21.0
apple	52.1	11.5	83.4	8.2	80.7	6.5	78.9	7.1	**93.4**	4.3	32.9	8.1	78.9	13.5	60.3	11.8	60.9	17.4	**85.6**	8.7
backpack	30.0	8.7	55.4	6.5	35.2	5.0	34.4	5.2	**63.7**	13.0	14.0	7.2	45.2	8.5	5.5	4.1	14.5	7.4	**48.0**	10.9
basket	21.9	8.4	67.7	16.5	67.3	16.5	55.9	18.1	**81.0**	12.8	22.3	11.4	61.3	13.2	44.0	21.3	32.5	21.3	**74.9**	16.3
bulldozer	34.7	15.5	51.8	13.1	53.7	5.5	66.7	2.9	**69.6**	12.3	20.8	11.9	40.8	13.3	33.7	2.7	46.6	4.5	**58.9**	13.0
butterfly	42.0	6.1	79.6	11.3	68.0	8.9	63.7	8.9	**93.2**	3.7	35.4	6.2	71.4	17.3	47.1	26.0	33.7	19.9	**90.4**	5.2
cactus	30.3	11.5	86.6	13.4	49.0	10.9	52.3	9.3	**90.3**	10.5	25.1	11.7	86.4	12.4	27.4	20.2	16.6	11.5	**89.9**	11.2
calculator	25.3	5.2	90.5	5.9	62.3	10.0	49.2	14.3	**94.0**	3.8	18.5	10.2	89.3	7.0	44.4	6.7	21.1	12.4	**91.4**	6.3
campfire	30.8	8.1	91.7	3.2	81.1	4.0	74.0	5.1	**94.5**	1.7	17.6	6.4	89.9	4.6	72.7	10.9	58.3	6.9	**90.6**	4.2
candle	21.0	4.4	90.6	5.5	87.0	2.6	86.4	2.0	**97.0**	1.8	27.0	4.0	80.6	9.2	70.4	7.6	69.4	7.4	**96.2**	2.2
coffee cup	44.6	11.9	79.4	9.2	71.5	7.5	74.2	3.1	**87.8**	4.3	24.9	12.1	81.6	8.0	53.4	6.1	55.0	4.5	**85.5**	4.3
crab	25.7	3.7	60.1	22.6	49.1	12.8	48.7	13.2	**75.4**	14.5	21.0	3.8	57.9	23.7	28.3	9.7	25.4	16.7	**70.2**	17.8
drill	44.5	11.8	71.2	7.9	80.9	1.9	84.9	1.3	**88.6**	8.2	26.7	6.9	55.2	5.2	53.4	2.1	68.6	9.9	**79.7**	8.4
duck	27.3	5.5	66.9	12.7	56.0	10.0	72.7	5.3	**91.1**	4.3	20.6	8.8	58.0	15.4	30.2	16.4	51.4	10.2	**86.6**	7.5
face	11.0	3.8	60.1	18.7	37.8	9.1	39.6	11.0	**70.4**	13.9	16.5	7.8	45.4	15.4	11.6	4.8	16.2	10.0	**61.7**	12.2
flower	18.1	3.4	74.8	14.5	63.0	2.1	58.0	2.8	**88.6**	2.0	25.6	5.1	71.0	11.1	36.2	7.9	28.0	1.7	**90.0**	2.0
house	23.5	9.6	79.9	9.6	59.5	9.3	58.9	8.5	**90.8**	3.8	19.8	7.5	78.8	9.2	36.7	7.2	32.3	9.8	**86.9**	4.4
ice cream	30.3	9.8	83.9	6.0	75.8	4.8	73.0	1.5	**88.8**	7.4	22.3	7.0	81.1	7.5	63.7	6.7	62.5	3.5	**84.3**	8.5
pig	21.7	1.1	68.9	26.4	34.3	17.2	51.2	5.6	**77.3**	26.2	22.4	3.6	59.7	25.4	15.0	12.4	26.2	4.5	**69.1**	26.2
pineapple	29.4	6.2	76.0	11.9	65.1	3.9	58.5	5.3	**79.2**	7.2	29.4	3.1	72.8	10.8	40.4	7.5	36.1	6.4	**73.1**	7.0
suitcase	30.4	14.2	89.3	2.3	82.7	4.8	81.2	4.3	**93.7**	3.0	16.3	6.1	91.0	2.5	72.4	7.4	60.0	7.2	**92.6**	3.4
Average	27.1	7.6	73.4	11.5	61.1	9.1	61.7	7.6	**83.9**	8.9	20.7	7.0	66.8	12.2	41.0	10.5	38.6	10.4	**78.0**	10.3
Airplane [5]	20.0	6.3	53.0	13.9	55.9	19.1	57.8	14.0	**67.2**	19.7	15.2	8.5	40.1	8.2	36.4	8.3	28.6	5.0	**58.9**	23.9
Airplane[6]	16.4	10.2	32.7	19.1	29.7	23.1	35.4	23.8	**37.9**	20.4	9.6	6.0	15.6	11.3	13.7	11.9	6.7	8.1	**26.2**	15.0
Creative birds	13.9	4.1	13.8	5.9	28.2	9.3	28.3	3.3	**29.9**	2.6	15.3	2.2	14.4	1.9	16.5	7.7	12.6	3.8	**19.1**	4.7

TABLE VII: Numerical evaluation on the SPG dataset [1]: first 25 categories; on the 'airplane' category from TUBerlin [5] and Huang14 [6] datasets; on creative birds [7]. μ denotes the average accuracy over 5 runs with 5 randomly chosen templates, and σ is the standard deviation of the 5 runs results. The evaluation in this table is done only on those sketches that have the same semantic parts as an exemplar sketch.

provides the comparison between SGCN [2], FLSS [3], ISPP [4], ours joint training strategy used in the main document (Ours Joint), and a two steps training (Ours Separate). It can be seen that on average separate training results in a slightly better average segmentation accuracy with P-metric of 84% vs. 83.9%, and C-metric of 77.6% vs. 77.4%. Nevertheless, (Ours Joint) strategy gives higher points accuracy than (Ours Separate) on 14 out of 25 categories on the SPG dataset. Moreover, (Ours Joint) consistently outperforms all other methods, while (Ours Separate) gives lower accuracy than SGCN on the 'backpack' and 'house' categories. We observe that the stroke-level deformation benefits from joint training, although, for some categories, it comes at cost of decreased performance of the keypoints prediction step (*e.g.* the 'bulldozer' category). Joint strategy results in a more robust performance across the categories with the standard deviation of point accuracy equal to 9.6% versus 10.1% for the separate training strategy (Table XIII).

E. Keypoints sensitivity to rotations and robustness of their prediction

As demonstrated in the supplemental web-pages and in Fig. 9, keypoints prediction is robust to rotations, not affecting the segmentation performance. The mean μ and standard deviation σ of mean L_2-distances between the keypoints from the original sketch and its reflected version (after reflecting back), on the ablation categories is $\mu = 0.058$, $\sigma = 0.008$. All sketches are normalized to fit the [-0.5,0.5] bounding box.

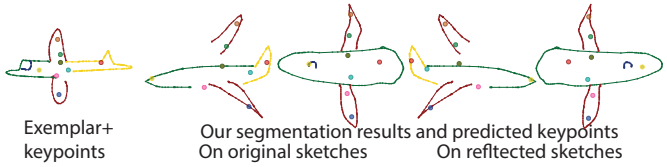

Fig. 9: Keypoints and segmentation results. We visualize 8 keypoints, while use 256 for deformations computations.

REFERENCES

[1] K. Li, K. Pang, J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Universal sketch perceptual grouping," in *Proc. Eur. Conf. Comput. Vis.*, 2018. 2, 3

[2] L. Yang, J. Zhuang, H. Fu, X. Wei, K. Zhou, and Y. Zheng, "Sketchgnn: Semantic sketch segmentation with graph neural networks," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–13, 2021. 2, 3, 4

[3] L. Wang, X. Li, and Y. Fang, "Few-shot learning of part-specific probability space for 3d shape segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020. 2, 3, 4

[4] N. Chen, L. Liu, Z. Cui, R. Chen, D. Ceylan, C. Tu, and W. Wang, "Unsupervised learning of intrinsic structural representation points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020. 2, 3, 4

[5] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, 2012. 2, 3

[6] Z. Huang, H. Fu, and R. W. Lau, "Data-driven segmentation and labeling of freehand sketches," *ACM Trans. Graph.*, vol. 33, no. 6, 2014. 2, 3

[7] S. Ge, V. Goswami, C. L. Zitnick, and D. Parikh, "Creative sketch generation," in *Int. Conf. Learn. Represent.*, 2021. 2, 3

	P-metric										C-metric									
	SPGG [1]		SGCN [2]		FLSS [3]		ISPP [4]		Ours		SPGG [1]		SGCN [2]		FLSS [3]		ISPP [4]		Ours	
Category	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
airplane	23.8	10.5	66.8	13.7	71.5	12.8	57.4	20.7	**86.5**	4.6	23.4	9.9	59.1	15.4	62.4	15.7	50.3	20.7	**81.0**	7.7
alarm clock	28.1	8.4	79.9	9.9	78.9	11.4	76.5	10.7	**86.6**	8.9	17.0	7.4	68.9	15.6	65.7	16.8	63.8	14.9	**76.4**	14.9
ambulance	28.7	9.3	78.3	3.1	70.5	13.5	68.3	9.5	**87.3**	3.3	16.1	7.9	67.6	7.3	62.8	11.5	48.5	8.3	**81.2**	5.4
angel	2.3	0.8	54.3	12.2	54.3	13.9	69.9	6.6	**71.0**	11.4	2.1	1.2	46.5	13.8	44.3	16.6	59.2	7.7	**67.1**	7.2
ant	10.1	1.3	44.1	17.4	41.5	14.3	48.6	14.5	**60.5**	17.9	9.9	6.0	35.4	9.3	30.8	10.7	40.4	12.9	**51.0**	18.0
apple	59.1	8.8	82.6	11.4	90.6	5.0	91.3	4.9	**94.3**	5.4	40.8	4.7	78.9	16.2	73.9	12.4	80.1	13.7	**87.4**	11.5
backpack	34.5	5.8	59.4	4.1	48.2	6.0	44.1	10.4	**65.3**	9.4	27.0	4.0	47.7	2.5	35.3	5.9	34.0	7.0	**52.5**	10.6
basket	28.6	2.6	68.9	15.8	77.5	14.9	66.9	21.3	**79.2**	11.1	27.3	8.2	61.9	13.0	68.4	17.9	58.3	23.4	**70.4**	11.3
bulldozer	42.2	7.6	53.6	15.6	59.8	10.7	**78.6**	7.2	69.2	10.9	27.2	10.7	44.7	17.7	53.4	11.5	**72.7**	6.6	59.0	12.8
butterfly	44.6	3.8	78.0	9.9	82.0	13.8	74.0	8.4	**91.8**	3.5	43.3	5.3	68.9	14.0	73.9	19.8	62.6	8.7	**87.3**	5.1
cactus	37.9	8.2	84.6	4.6	36.8	15.5	49.1	21.4	**89.0**	7.4	20.9	10.6	80.5	8.3	30.7	15.6	36.9	21.4	**84.1**	10.2
calculator	24.8	2.2	89.2	4.6	83.1	4.3	69.7	20.8	**92.8**	2.9	27.7	3.5	88.0	4.5	69.0	6.9	57.4	17.5	**90.5**	5.5
campfire	33.5	9.6	92.1	3.2	91.3	6.7	79.9	12.6	**94.0**	1.6	22.2	6.7	88.6	4.6	88.4	9.2	75.3	11.5	**88.6**	3.9
candle	19.1	5.4	89.9	5.7	95.4	1.8	94.1	2.5	**96.4**	1.4	28.3	4.7	81.6	10.2	87.2	5.9	82.7	8.6	**94.2**	2.0
coffee cup	50.8	11.0	78.6	11.2	80.8	6.7	79.3	10.9	**82.4**	7.8	34.6	9.2	77.9	14.9	76.2	9.3	66.0	13.7	**82.2**	5.0
crab	28.4	5.5	56.2	13.7	49.0	14.9	51.2	15.9	**75.6**	13.1	25.5	5.6	52.2	11.7	39.2	16.0	40.2	15.3	**70.3**	15.1
drill	46.6	14.3	71.5	8.0	82.1	1.0	85.0	4.2	**89.1**	7.8	31.9	10.1	57.6	5.2	55.8	3.1	77.0	10.5	**80.7**	9.0
duck	28.4	6.8	61.1	10.4	54.2	5.8	82.8	7.6	**89.7**	3.8	25.2	4.7	54.0	9.4	41.9	8.2	71.4	9.0	**83.5**	7.3
face	12.3	1.8	70.0	14.3	57.1	10.7	58.8	21.5	**83.7**	6.8	20.0	5.5	55.8	16.1	39.1	7.2	46.5	18.7	**74.7**	9.6
flower	14.1	1.8	75.7	14.4	73.6	2.7	78.7	4.0	**89.2**	1.9	31.5	1.4	73.9	10.4	58.2	4.2	72.7	9.6	**89.9**	2.1
house	20.0	7.5	82.3	9.3	76.3	7.7	72.1	5.8	**89.5**	2.5	20.7	7.5	82.1	7.9	68.0	8.4	66.8	6.1	**86.2**	2.2
ice cream	28.5	9.2	82.6	5.8	78.7	11.5	81.9	4.9	**86.7**	8.5	27.8	9.7	79.3	6.2	76.0	10.3	81.1	4.2	**81.0**	8.5
pig	21.5	2.5	67.0	20.8	44.5	17.9	60.2	15.2	**77.4**	12.6	25.3	0.6	56.1	20.8	34.5	15.1	45.7	12.7	**66.0**	16.9
pineapple	25.2	7.9	76.9	13.6	81.0	4.4	64.5	17.7	**81.0**	8.2	30.1	4.8	74.8	12.5	72.6	3.8	60.7	13.0	**76.9**	6.8
suitcase	28.3	9.8	89.3	1.5	**94.7**	2.4	88.1	5.7	93.7	1.7	22.4	9.4	90.8	1.6	**93.7**	2.9	81.4	7.8	93.0	2.0
Average	28.9	6.5	73.3	10.2	70.1	9.2	70.8	11.4	**84.1**	7.0	25.1	6.4	66.9	10.8	60.1	10.6	61.3	12.1	**78.2**	8.4
Airplane [5]	20.8	5.8	54.8	13.6	62.0	13.6	64.2	17.4	**65.3**	13.8	13.8	5.9	42.2	9.8	52.6	15.8	53.3	19.8	**53.6**	8.7
Airplane [6]	16.8	5.2	44.7	5.0	50.7	8.1	44.4	10.9	**53.1**	6.8	14.9	4.5	30.3	7.6	28.8	7.2	19.2	10.8	**33.4**	11.3
Creative birds	14.5	4.4	12.5	4.6	29.5	7.1	29.6	3.9	**30.4**	1.7	13.6	3.9	12.5	2.5	20.8	5.3	20.6	5.8	**20.1**	1.0

TABLE VIII: Numerical evaluation on the SPG dataset [1]: first 25 categories; on the 'airplane' category from TUBerlin [5] and Huang14 [6] datasets; on creative birds [7]. μ denotes the average accuracy over 5 runs with 5 randomly chosen templates, and σ is the standard deviation of the 5 runs results. The results after refining each point label with a label dominant for each stroke.

		1 template		3 templates		5 templates	
	Category	(P)	(C)	(P)	(C)	(P)	(C)
Ours	ambulance	**86.0**	**77.4**	**92.8**	**90.5**	**93.2**	**91.1**
	apple	**83.4**	**69.6**	**89.1**	**81.3**	**89.8**	**80.3**
	duck	**88.2**	**78.3**	**91.3**	**84.8**	**92.9**	**89.0**
	face	**86.0**	**77.6**	**91.9**	**85.0**	**93.8**	**89.9**
	pig	**85.2**	**76.8**	**91.4**	**85.6**	**92.3**	**88.8**
SGCN [2]	ambulance	77.5	63.4	90.1	86.3	92.1	87.7
	apple	82.1	65.3	84.0	72.0	86.0	77.0
	duck	74.0	59.0	86.2	80.8	77.2	68.8
	face	78.4	64.6	86.8	78.9	89.9	82.1
	pig	77.4	64.7	85.6	79.2	87.2	80.4

TABLE IX: One shot vs. few shot. See Sec.VII-C for the details.

	P-metric				C-metric			
	$\mathcal{F}^{sketch}_{\hat{X}}$		$\mathcal{F}^{sketch}_{\hat{E}}$		$\mathcal{F}^{sketch}_{\hat{X}}$		$\mathcal{F}^{sketch}_{\hat{E}}$	
Category	μ	σ	μ	σ	μ	σ	μ	σ
ambulance	86.7	4.3	**87.1**	3.7	80.2	7.5	**81.4**	6.2
apple	94.1	5.6	**94.3**	5.4	86.6	12.5	**87.1**	11.9
duck	89.1	4.5	**89.6**	4.1	83.6	7.6	**84.0**	7.5
face	82.2	6.4	**83.3**	6.9	70.1	11.5	**72.2**	11.7
pig	76.4	16.4	**76.6**	14.6	64.4	20.0	**64.5**	18.1

TABLE XI: Numerical evaluation of alternative strategies in the segmentation module.

	P-metric				C-metric			
	ISPP PointNet++		ISPP GCN		ISPP PointNet++		ISPP GCN	
Category	μ	σ	μ	σ	μ	σ	μ	σ
ambulance	58.3	8.2	60.1	10.3	28.5	10.8	33.7	8.7
apple	73.9	9.9	78.2	7.6	50.5	18.3	**56.7**	14.9
duck	67.2	10.6	71.2	6.0	45.9	14.8	48.4	10.5
face	39.2	8.2	41.8	10.8	11.0	6.2	16.6	11.5
pig	40.7	9.1	45.8	9.7	15.1	7.5	20.6	8.6
Average	55.9	9.2	59.4	8.9	30.2	11.5	35.2	10.8

TABLE X: Segmentation accuracy comparison for the ISPP [4] method, when the originally proposed PointNet++ encoder is used instead of our GCN encoder.

	P-metric				C-metric			
	No $\mathcal{L}_{CD}$		Ours full		No $\mathcal{L}_{CD}$		Ours full	
Category	μ	σ	μ	σ	μ	σ	μ	σ
ambulance	**88.4**	2.4	87.1	3.7	**82.5**	4.3	81.4	6.2
apple	90.8	7.0	**94.3**	5.4	79.9	14.2	**87.1**	11.9
duck	83.1	11.2	**89.6**	4.1	74.7	15.5	**84.0**	7.5
face	81.5	6.9	**83.3**	6.9	69.3	8.2	**72.2**	11.7
pig	75.9	15.8	**76.6**	14.6	62.4	20.0	**64.5**	18.1
Average	83.9	8.7	**86.2**	6.9	73.8	12.4	**77.8**	11.1

TABLE XII: The role of Chamfer distance for stroke-level deformation estimation.

	P-metric										C-metric									
	SGCN [2]		**FLSS [3]**		**ISPP [4]**		**Ours (Joint)**		**Ours (Separate)**		**SGCN [2]**		**FLSS [3]**		**ISPP [4]**		**Ours (Joint)**		**Ours (Separate)**	
Category	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
airplane	66.6	14.0	56.3	11.2	50.8	14.2	**86.0**	4.9	85.3	7.7	57.3	17.8	34.6	15.5	23.0	12.7	**80.6**	8.5	79.6	11.3
alarm clock	79.7	9.9	59.7	10.4	59.4	11.5	**86.4**	9.1	85.4	14.4	68.4	15.7	36.5	17.4	32.9	17.8	76.0	15.7	**77.1**	22.3
ambulance	78.1	3.4	61.5	12.0	60.1	10.3	87.1	3.7	**89.1**	4.6	66.9	7.2	46.2	10.4	33.7	8.7	81.4	6.2	**85.6**	7.3
angel	54.2	12.0	47.6	9.4	57.8	6.0	**70.7**	11.5	69.8	13.0	45.7	13.2	22.1	10.3	31.4	7.2	**64.9**	7.8	62.5	6.8
ant	44.2	17.5	41.7	14.5	47.3	12.6	60.8	18.0	**70.1**	10.1	35.2	9.5	22.5	12.6	27.2	15.0	50.6	18.6	**55.5**	11.5
apple	83.4	10.7	82.0	8.6	78.2	7.6	**94.3**	5.4	94.2	5.3	78.3	16.9	59.5	13.1	56.7	14.9	**87.1**	11.9	86.8	11.7
backpack	59.2	3.9	35.9	3.8	33.7	6.0	**64.6**	9.2	55.8	13.8	46.6	2.1	6.4	3.8	8.0	4.5	**50.2**	11.0	41	13.8
basket	68.7	15.7	65.9	14.2	55.2	15.1	**79.1**	10.3	72.4	14.1	61.1	12.7	41.6	20.4	28.8	18.2	**70.1**	11.3	61.5	15.5
bulldozer	53.4	15.7	56.0	9.2	67.9	5.1	69.1	11.0	**84.4**	5.0	43.1	18.4	38.8	11.2	49.1	8.9	58.5	13.4	**77.4**	6.4
butterfly	78.2	9.3	70.2	7.5	65.0	8.1	**91.7**	3.7	86.9	7.3	67.4	13.2	54.1	9.1	38.9	14.1	**86.2**	5.8	79.3	10.5
cactus	84.6	4.6	41.9	9.1	47.7	11.7	89.2	6.8	**91.0**	5.2	80.4	8.2	18.9	12.2	14.1	5.6	83.3	9.4	**85.7**	9.5
calculator	89.2	4.6	67.3	4.1	52.7	12.1	**92.6**	2.8	92.3	2.9	87.7	4.8	44.2	6.2	24.5	7.7	**90.1**	5.0	88	6.5
campfire	91.2	3.2	80.7	4.2	73.5	5.0	93.9	1.6	**95.0**	1.4	88.4	4.6	71.4	11.4	57.1	7.2	89.0	4.2	**92.8**	2.8
candle	89.8	5.7	86.7	4.5	85.2	1.7	96.3	1.7	**96.7**	1.5	81.0	10.0	71.9	10.3	67.8	7.4	93.9	2.2	**94.8**	2.1
coffee cup	73.6	10.6	73.7	5.9	66.2	7.1	**82.6**	6.9	82.3	18.9	76.2	13.6	54.6	11.8	38.3	15.5	**81.3**	5.7	78.8	22.9
crab	56.2	13.8	49.5	10.3	48.6	13.2	**75.4**	12.9	72.1	13.4	51.9	11.8	27.0	8.4	21.6	10.5	**69.9**	14.8	66.4	14.9
drill	71.3	8.2	80.6	1.9	84.1	1.5	88.7	8.0	**96.7**	1.0	56.4	5.3	55.1	1.9	68.1	8.6	79.8	10.3	**95.6**	1.3
duck	61.2	10.5	53.6	4.5	71.2	6.0	89.6	4.1	**91.4**	3.0	53.9	9.3	26.5	8.4	48.4	10.5	84.0	7.5	**86.6**	5.8
face	69.9	14.4	38.3	6.8	41.8	10.8	**83.3**	6.9	83.0	5.0	55.2	16.0	12.4	8.2	16.6	11.5	72.2	11.7	**72.7**	6.6
flower	75.6	14.2	62.6	3.5	58.1	3.4	88.3	2.0	**89.0**	4.2	72.7	11.2	36.4	6.5	27.3	3.7	87.1	3.2	**88.4**	3.4
house	82.2	9.3	57.8	10.7	58.4	9.3	**89.4**	2.4	75.2	17.4	81.7	7.9	34.7	13.4	32.8	13.3	**85.3**	2.4	68	17.5
ice cream	82.5	5.7	75.2	4.3	72.9	1.0	**86.5**	8.3	84.5	10.1	79.2	5.9	62.2	5.6	60.0	2.8	**80.6**	8.3	79.1	10.2
pig	66.8	20.8	37.1	12.2	45.8	9.7	76.6	14.6	**78.9**	10.2	55.6	20.9	14.8	10.2	20.6	8.6	**64.5**	18.1	64.3	16.8
pineapple	76.9	13.5	66.6	5.1	56.5	8.4	80.8	8.3	**85.4**	5.8	74.4	12.3	43.5	11.4	35.0	7.3	75.8	6.8	**78.3**	6.3
suitcase	89.2	1.6	82.4	5.7	81.7	2.6	**93.8**	1.6	93.7	1.0	90.7	1.5	72.9	7.2	61.0	5.3	**93.2**	1.7	93.1	1.1
Average	73.0	10.1	61.2	7.7	60.8	8.0	83.9	7.0	**84.0**	7.9	66.2	10.8	40.4	10.3	36.9	9.9	**77.4**	8.9	**77.6**	9.8
Min	44.2		35.9		33.7		**60.8**		55.8		35.2		6.4		8.0		**50.2**		41	
Max	91.2		86.7		85.2		96.3		**96.7**		90.7		72.9		68.1		93.9		**95.6**	
Std.	12.9		15.5		13.6		9.6		10.1		15.6		19.0		17.2		12.2		13.5	

TABLE XIII: The comparison of training strategies for our proposed method. Ours (Joint) refers to the joint training strategy used in the main document. Ours (Separate) refers to a two a two-steps training strategy, where we first train the keypoints prediction network, as described in Section VIII-D. We also compute the minimum average accuracy across categories (Min), the maximum average accuracy across categories (Max), and the standard deviations across categories (Std.). These numbers allow to evaluate how consistent are the segmentation results of each method across different categories.