Supplementary Material for *Learning Semi-Supervised Medical Image Segmentation from Spatial Registration*

Qianying Liu
University of Glasgow
2665227L@student.gla.ac.uk

Paul Henderson†
University of Glasgow
paul.henderson@glasgow.ac.uk

Xiao Gu
University of Oxford
xiao.gu@eng.ox.ac.uk

Hang Dai
University of Glasgow
hang.dai@glasgow.ac.uk

Fani Deligianni†
University of Glasgow
fani.deligianni@glasgow.ac.uk

S1. Baselines.

We first compare with a registration baseline that is not learning-based—we use the transforms to propagate labels from the labeled training cases to the test images, similar to [1, 4, 9], selecting labeled cases with our BRS. We also compare a joint registration and segmentation model, DeepAtlas [18]; this learns registration from scratch simultaneously with segmentation. To stay consistent with our CCT-R, we reimplemented it using a 2D U-Net segmentation model. We evaluate several recent S4 methods with the U-Net [15] backbone: Mean Teacher (MT) [16], Deep Co-Training (DCT) [14], Uncertainty Aware Mean Teacher (UAMT) [19], Interpolation Consistency Training (ICT) [17], Cross Consistency Training (CCT) [13], Cross Pseudo Supervision (CPS) [3], and Cross Teaching Supervision (CTS) [11], which like CCT-R uses Swin-UNet [2] (Transformer) and U-Net backbones. In addition, we include the SOTA S4 method with contrastive learning, MCSC [7]. As a reference we also train the U-Net backbone from the S4 methods on only the labeled subset of cases (LS) without additional tricks. We also include fully-supervised methods—the same U-Net trained under full supervision (FS), and the SOTA fully-supervised methods BATFormer [6] (on ACDC) and nnFormer [20] (on Synapse). We retrain all baseline models using their recommended hyperparameters, and report the results from [11] or our replication, whichever is better.

S2. Implementation Details

For all methods we use random cropping, random flipping and rotations to augment. All methods were trained until convergence, or up to 40,000 iterations. We precomputed a composite pairwise registration (affine for ACDC and affine + B-spline deformable transformation for Synapse) for all training data prior to training, using ITK [10, 12]. The compute time required for each affine registration is approximately 2 minutes per pair, while each deformable pair takes around 3 hours based on 50 CPUs. Consequently, the computational overhead for affine transformations on the ACDC and Synapse datasets is roughly 161 and 10 hours, respectively. For Synapse, the deformable transformations require approximately 918 hours. However, by parallelizing up to 5 registration tasks, we can reduce the effective time to 1/5, maximizing CPU utilization. Additionally, if computational resources are limited, using only affine transformations offers a cost-effective alternative. We used the AdamW optimizer with a weight decay of 5×10^{-4}. The learning rate followed a polynomial schedule, starting at 5×10^{-4} for the U-Net and 1×10^{-4} for the Swin-Unet. Our training batches consisted of 8 images for ACDC and 24 images for Synapse, evenly split between labeled and unlabeled. In the contrastive learning section, each (H_*) was composed of two linear layers, outputting 256 and 128 channels, respectively. In Eq. 6, w_{cps} is defined by a Gaussian warm-up function [11]: $w_{cps}(i) = 0.1 \cdot \exp\left(-5(1 - i/t_{\text{total}})^2\right)$, where i is the index of the current training iteration and t_{total} is the total number of iterations, while w_{cl} is set to a constant value of 10^{-3}. In Eq. 4, temperature $\tau = 0.1$. In REPS module, the bank size $K = (M + K)/5$. We implemented our method in PyTorch. All experiments were run on one RTX 3090 GPU.

S3. Full results on ACDC and Synapse

Here we show extended versions of Table 1 and 2 in the main paper as Table S1 and Table S2. In these extended tables, we provide additional comparisons by separately evaluating the performance of the two branches (CNN and Transformer) of our CCT-R (whereas in the main paper we use the mean of their logits); we also give results for all baselines under three different

Table S1. Segmentation results on ACDC for our method CCT-R and baselines, according to DSC(%) and HD(mm) for organs.

Labeled	**Methods**	**Mean**		**Myo**		**LV**		**RV**	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
70 (100%)	UNet-FS	91.7	4.0	89.0	5.0	94.6	5.9	91.4	1.2
	BATFormer [6]	92.8	8.0	90.26	6.8	96.3	5.9	91.97	11.3
7 (10%)	Reg. only (Aff)	30.7	16.4	19.7	13.9	42.0	14.4	30.5	20.8
	DeepAtlas [18]	79.4	8.0	79.0	11.7	81.9	3.2	77.3	9.0
	UNet-LS	75.9	10.8	78.2	8.6	85.5	13.0	63.9	10.7
	MT [16]	80.9	11.5	79.1	7.7	86.1	13.4	77.6	13.3
	DCT [14]	80.4	13.8	79.3	10.7	87.0	15.5	75.0	15.3
	UAMT [19]	81.1	11.2	80.1	13.7	87.1	18.1	77.6	14.7
	ICT [17]	82.4	7.2	81.5	7.8	87.6	10.6	78.2	3.2
	CCT [13]	84.0	6.6	82.3	5.4	88.6	9.4	81.0	5.1
	CPS [3]	85.0	6.6	82.9	6.6	88.0	10.8	84.2	2.3
	CTS [11]	86.4	8.6	84.4	6.9	90.1	11.2	84.8	7.8
	MCSC [7]	89.4	2.3	**87.6**	**1.1**	**93.6**	3.5	87.1	2.1
	Ours (CNN, Affine)	<u>89.5</u>	<u>1.8</u>	87.2	2.0	<u>92.9</u>	**1.8**	88.4	<u>1.7</u>
	Ours (Trans, Affine)	89.1	<u>1.8</u>	85.7	<u>1.2</u>	91.7	2.8	<u>89.9</u>	**1.3**
	Ours (mean, Affine)	**90.3**	**1.6**	<u>87.4</u>	1.4	92.7	<u>2.2</u>	**90.9**	**1.3**
3 (5%)	Reg. only (Aff)	32.0	17.8	18.0	15.7	43.9	16.0	34.0	21.7
	DeepAtlas [18]	59.0	8.6	62.8	5.4	67.8	7.7	46.4	12.6
	UNet-LS	51.2	31.2	54.8	24.4	61.8	24.3	37.0	44.4
	MT [16]	56.6	34.5	58.6	23.1	70.9	26.3	40.3	53.9
	DCT [14]	58.2	26.4	61.7	20.3	71.7	27.3	41.3	31.7
	UAMT [19]	61.0	25.8	61.5	19.3	70.7	22.6	50.8	35.4
	ICT [17]	58.1	22.8	62.0	20.4	67.3	24.1	44.8	23.8
	CCT [13]	58.6	27.9	64.7	22.4	70.4	27.1	40.8	34.2
	CPS [3]	60.3	25.5	65.2	18.3	72.0	22.2	43.8	35.8
	CTS [11]	65.6	16.2	62.8	11.5	76.3	15.7	57.7	21.4
	MCSC [7]	73.6	10.5	70.0	8.8	79.2	14.9	71.7	7.8
	Ours (CNN, Affine)	85.2	**1.9**	<u>83.3</u>	<u>1.5</u>	**89.9**	<u>2.9</u>	<u>82.4</u>	<u>2.2</u>
	Ours (Trans, Affine)	<u>85.4</u>	2.6	83.2	1.8	<u>89.3</u>	3.8	**83.5**	**2.1**
	Ours (mean, Affine)	**85.7**	<u>2.0</u>	**83.8**	**1.4**	**89.9**	**2.4**	**83.5**	**2.1**
1 (1.4%)	Reg. only (Aff)	23.4	19.7	13.6	18.7	31.6	19.0	25.1	21.4
	DeepAtlas [18]	40.4	18.5	42.2	11.7	34.7	29.2	44.4	14.6
	UNet-LS	26.4	60.1	26.3	51.2	28.3	52.0	24.6	77.0
	CTS [11]	46.8	36.3	55.1	5.5	64.8	**4.1**	20.5	99.4
	MCSC [7]	58.6	31.2	64.2	13.3	78.1	12.2	33.5	68.1
	Ours (CNN, Affine)	79.6	5.2	77.6	5.3	<u>83.2</u>	5.1	78.0	5.1
	Ours (Trans, Affine)	<u>80.0</u>	<u>4.2</u>	<u>77.7</u>	<u>4.0</u>	83.0	<u>4.2</u>	**79.4**	<u>3.6</u>
	Ours (mean, Affine)	**80.4**	**3.5**	**78.3**	**3.2**	**83.6**	4.3	<u>79.3</u>	**2.9**

Best is bold, <u>Second Best</u> is underlined.

settings on both datasets. It can be seen that on the ACDC dataset, the performance of CCT-R's CNN and Transformer branches is quite similar. However, on the more challenging Synapse dataset, the Transformer outperforms the CNN, likely due to its superior ability to capture long-range dependencies, which allows it to better handle the relationships between large and small organs.

Table S2. Segmentation results on Synapse for our method CCT-R and baselines, according to DSC(%) and HD(mm).

Labeled	Methods	DSC↑	HD↓	Aorta	Gallb	Kid_L	Kid_R	Liver	Pancr	Spleen	Stom
18(100%)	UNet-FS	75.6	42.3	88.8	56.1	78.9	72.6	91.9	55.8	85.8	74.7
	nnFormer	86.6	10.6	92.0	70.2	86.6	86.3	96.8	83.4	90.5	86.8
4(20%)	Reg. only (Affine)	27.0	39.6	16.0	7.5	36.4	33.0	56.8	13.1	28.5	25.1
	Reg. only (Aff+Def)	32.5	36.5	29.7	4.8	36.5	29.4	65.5	14.2	48.0	31.7
	DeepAtlas [18]	56.1	85.3	69.2	43.3	50.8	55.2	88.8	30.5	62.7	48.0
	UNet-LS	47.2	122.3	67.6	29.7	47.2	50.7	79.1	25.2	56.8	21.5
	UAMT [19]	51.9	69.3	75.3	33.4	55.3	40.8	82.6	27.5	55.9	44.7
	ICT [17]	57.5	79.3	74.2	36.6	58.3	51.7	86.7	34.7	66.2	51.6
	CCT [13]	51.4	102.9	71.8	31.2	52.0	50.1	83.0	32.5	65.5	25.2
	CPS [3]	57.9	62.6	75.6	41.4	60.1	53.0	88.2	26.2	69.6	48.9
	CTS [11]	64.0	56.4	79.9	38.9	66.3	63.5	86.1	41.9	75.3	60.4
	MCSC [7]	68.5	24.8	76.3	<u>44.4</u>	<u>73.4</u>	<u>72.3</u>	91.8	46.9	79.9	62.9
	Ours (CNN, Affine)	67.3	37.9	79.0	36.5	72.7	70.4	87.9	47.3	77.8	67.0
	Ours (Trans, Affine)	70.5	22.7	**81.0**	34.1	71.1	71.9	93.2	<u>49.9</u>	**87.9**	**75.2**
	Ours (mean, Affine)	70.0	23.2	79.8	34.5	71.0	70.7	92.8	49.6	<u>87.4</u>	<u>74.4</u>
	Ours (CNN, Affine+Deform)	69.5	36.2	80.0	**49.2**	73.0	69.9	89.3	48.5	79.5	66.7
	Ours (Trans, Affine+Deform)	**72.5**	**20.5**	<u>80.9</u>	43.4	**75.6**	**75.1**	<u>93.5</u>	**51.3**	<u>87.4</u>	72.2
	Ours (mean, Affine+Deform)	<u>71.4</u>	<u>21.1</u>	80.4	42.3	73.0	70.0	**93.7**	49.4	**87.9**	74.2
2(10%)	Reg. only (Affine)	25.4	36.8	17.5	3.5	32.7	27.5	53.4	12.6	33.4	22.5
	Reg. only (Aff+Def)	29.1	44.0	27.2	11.3	28.6	26.5	66.4	12.7	29.7	30.3
	DeepAtlas [18]	44.0	67.1	68.0	24.9	37.9	46.0	82.7	18.4	44.2	30.6
	UNet-LS	45.2	55.6	66.4	27.2	46.0	48.0	82.6	18.2	39.9	33.4
	UAMT [19]	49.5	62.6	71.3	21.1	62.6	51.4	79.3	22.8	58.2	29.0
	ICT [17]	49.0	59.9	68.9	19.9	52.5	52.2	83.7	25.4	53.2	36.0
	CCT [13]	46.9	58.2	66.0	26.6	53.4	41.0	82.9	21.2	48.7	35.6
	CPS [3]	48.8	65.6	70.9	21.3	58.0	45.1	80.7	23.5	58.0	32.7
	CTS [11]	55.2	45.4	71.5	25.6	62.6	67.5	78.2	26.3	75.9	34.3
	MCSC [7]	61.1	32.6	73.9	26.4	69.9	72.7	90.0	33.2	79.4	43.0
	Ours (CNN, Affine)	60.4	37.1	77.0	27.8	70.8	69.0	88.4	35.4	67.0	47.7
	Ours (Trans, Affine)	64.2	<u>22.1</u>	<u>77.4</u>	22.1	75.0	74.2	92.2	<u>39.6</u>	78.2	54.8
	Ours (mean, Affine)	65.1	22.5	75.7	28.4	74.5	**75.0**	91.8	38.0	<u>82.3</u>	55.1
	Ours (CNN, Affine+Deform)	62.6	44.3	76.5	<u>37.7</u>	73.0	68.0	87.0	32.3	76.5	49.9
	Ours (Trans, Affine+Deform)	**68.3**	23.1	74.8	**49.1**	**75.2**	<u>74.7</u>	**92.8**	**39.7**	**84.1**	**56.2**
	Ours (mean, Affine+Deform)	<u>66.5</u>	**19.7**	**77.6**	34.4	<u>75.1</u>	74.2	<u>92.6</u>	39.5	82.1	<u>56.1</u>
1(5%)	Reg. only (Affine)	26.4	45.0	16.3	6.6	35.8	32.8	53.5	14.4	28.7	22.7
	Reg. only (Aff+Def)	27.4	52.2	26.4	11.3	30.5	27.1	61.6	12.8	26.3	23.6
	DeepAtlas [18]	16.1	72.3	18.4	14.9	1.2	10.1	57.1	0.6	14.4	12.2
	UNet-LS	13.7	116.5	11.6	**17.8**	0.8	1.8	56.9	0.1	8.7	11.6
	UAMT [19]	10.7	90.2	8.0	9.3	0.3	8.1	31.7	1.1	13.1	14.3
	ICT [17]	15.9	82.3	13.8	11.9	0.3	2.7	70.5	0.8	16.4	10.6
	CCT [13]	11.7	107.5	10.0	13.0	0.1	1.9	47.5	3.7	8.0	9.3
	CPS [3]	15.0	123.5	19.6	9.6	5.6	6.9	59.4	2.3	9.4	7.2
	CTS [11]	26.3	96.5	44.6	4.0	11.2	5.5	60.3	9.6	54.1	21.2
	MCSC [7]	34.0	53.8	50.9	13.0	17.6	54.6	64.3	5.5	43.1	23.5
	Ours (CNN, Affine)	39.5	66.5	61.7	<u>17.0</u>	9.2	65.2	71.1	**12.3**	54.3	25.3
	Ours (Trans, Affine)	43.2	67.5	58.5	12.5	20.2	66.6	**78.9**	10.3	<u>72.9</u>	26.5
	Ours (mean, Affine)	43.4	<u>40.8</u>	62.5	13.3	17.9	**71.0**	<u>77.0</u>	<u>11.4</u>	65.4	**28.7**
	Ours (CNN, Affine+Deform)	44.2	54.2	<u>63.8</u>	10.8	48.7	61.6	74.6	5.4	61.8	26.6
	Ours (Trans, Affine+Deform)	<u>45.3</u>	46.9	62.9	9.9	<u>56.5</u>	65.6	70.9	0.1	72.8	24.2
	Ours (mean, Affine+Deform)	**47.6**	**38.4**	**65.5**	9.3	**61.6**	<u>70.2</u>	72.7	0.1	**73.9**	<u>27.8</u>

Best is bold, <u>Second Best</u> is underlined.

t

Table S3. Comparisons with SoTA contrastive learning methods combined with CTS, on ACDC and Synapse.

Contrastive learning method		**ACDC 3 (5 %)**		**/ 1 (1.4 %)**		**Synapse 4 (20 %)**		**/ 2 (10%)**	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
Patch-level	GLCL [5] (MICCAI'21)	71.7	3.8	47.4	35.8	67.7	42.6	59.7	34.6
	MCSC [7] (BMVC'23)	73.6	10.5	58.6	31.2	68.5	24.8	61.1	32.6
Slice-level	ReCo [8] (ICLR'22)	70.2	6.1	48.3	33.5	68.3	25.9	60.4	20.7
	Ours	**85.4**	**2.6**	**80.0**	**4.2**	**71.4**	**21.1**	**66.5**	**19.7**
None (Vanilla CTS [11])		65.6	16.2	46.8	36.3	64.0	56.4	57.2	45.7

Best is bold.

S4. Comparison with Alternative Supervised Contrastive Learning Losses

In Table S3, we compare our proposed approach with the state-of-the-art contrastive S4 method MCSC [7], and with incorporating other recent patch-level and slice-level contrastive learning techniques (GLCL [5] and ReCo [8]) into CTS. While all the contrastive losses improve on vanilla CTS, our CCT-R achieves higher segmentation accuracy on nearly all datasets and labelling rates.

References

[1] Noah C Benson, Omar H Butt, David H Brainard, and Geoffrey K Aguirre. Correction of distortion in flattened representations of the cortical surface allows prediction of v1-v3 functional organization from anatomy. *PLoS computational biology*, 10(3):e1003538, 2014. 1

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 1

[3] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 1, 2, 3

[4] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002. 1

[5] Xinrong Hu et al. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *MICCAI*, pages 481–490. Springer, 2021. 4

[6] Xian Lin et al. Batformer: Towards boundary-aware lightweight transformer for efficient medical image segmentation. *IEEE JBHI*, 2023. 1, 2

[7] Qianying Liu, Xiao Gu, Paul Henderson, and Fani Deligianni. Multi-scale cross contrastive learning for semi-supervised medical image segmentation. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 1, 2, 3, 4

[8] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew Davison. Bootstrapping semantic segmentation with regional contrast. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[9] Maria Lorenzo-Valdés, Gerardo I Sanchez-Ortiz, Raad Mohiaddin, and Daniel Rueckert. Atlas-based segmentation and tracking of 3d cardiac mr images using non-rigid registration. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part I 5*, pages 642–650. Springer, 2002. 1

[10] Bradley C Lowekamp, David T Chen, Luis Ibáñez, and Daniel Blezek. The design of simpleitk. *Frontiers in neuroinformatics*, 7:45, 2013. 1

[11] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International Conference on Medical Imaging with Deep Learning*, pages 820–833. PMLR, 2022. 1, 2, 3, 4

[12] Matthew McCormick, Xiaoxiao Liu, Julien Jomier, Charles Marion, and Luis Ibanez. Itk: enabling reproducible research and open science. *Frontiers in neuroinformatics*, 8:13, 2014. 1

[13] Yassine Ouali et al. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, pages 12674–12684, 2020. 1, 2, 3

[14] Siyuan Qiao et al. Deep co-training for semi-supervised image recognition. In *ECCV*, pages 135–152, 2018. 1, 2

[15] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1

[16] Antti Tarvainen et al. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NIPS*, 30, 2017. 1, 2

[17] Vikas Verma et al. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022. 1, 2, 3

[18] Zhenlin Xu and Marc Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 420–429. Springer, 2019. 1, 2, 3

[19] Lequan Yu et al. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, pages 605–613. Springer, 2019. 1, 2, 3

[20] Hong-Yu Zhou et al. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 1