Multi-Scale Cross Contrastive Learning for Semi-Supervised Medical Image Segmentation

Qianying Liu[1], Xiao Gu[2], Paul Hendersons[1], Fani Deligianni[1]

1. School of Computing Science, University of Glasgow 2. Department of Computing, Imperial College London

Code is on GitHub

Background and Purpose

- Most existing semi-supervised learning has demonstrated great potential in medical image segmentation by utilizing knowledge from unlabelled data.
- However, they do not explicitly capture high-level semantic relations between distant regions, which limits performance.
- We develop a novel Multi-Scale Cross Supervised Contrastive Learning framework, to jointly train CNN and Transformer models, regularising their features to be semantically consistent across different scales based on ground-truth and cross-predicted labels.
- MCSC outperforms state-of-the-art methods by more than 3.0% in Dice on two benchmarks.
- Code is available on GitHub (QR Code).

Ablations

SCL	DB	CroLab	Balanced	MulSca	Unet DSC↑	Unet HD↓	Transformer DSC↑	Transformer HD↓
					86.40	8.6	85.22	5.1
✓	✓				87.50	7.4	86.02	4.5
✓	✓	✓			88.23	3.4	86.13	3.2
✓		✓	✓		88.80	4.6	86.53	**2.4**
✓		✓	✓	✓	**89.38**	**2.3**	**87.28**	3.5

Tab. A1 Ablation study for the primary components of our model. SCL denotes supervised local contrastive loss. DB denotes discarding background pixels as anchor. CroLab stands for cross label information of two models to select contrastive sample. Balanced means averaging the instances of each class in denominator of SCL. MulSca means contrasting multi-scale feature maps.

Branches 256	56	28	Mean DSC↑	Mean HD↓
✓			88.80	4.6
	✓		88.88	4.2
		✓	88.39	4.5
✓		✓	**89.38**	**2.3**
✓	✓		88.92	2.9
✓	✓	✓	88.35	4.3

Tab. A2 Ablation on the choice of feature maps for the multi-scale (ACDC, 7 labelled cases).

MCSC framework

- Two networks, a CNN (pink) and Transformer (blue), with complementary inductive biases, learn together.
- On the output level, supervision loss $\mathcal{L}_{sup}$ (yellow dashed lines in Figure 1) between the segmentation predictions and the limited labelled data, as well as the cross pseudo supervision loss $\mathcal{L}_{cps}$ (green dashed lines) between the segmentation predictions and the pseudo labels from the U-Net or the Transformer in a cross teaching manner.
- On the feature level, we employ the proposed multi-scale cross contrastive loss $\mathcal{L}_{cl}$ (black dashed lines to enhance feature consistency of the same category and feature distinguishability of the different categories across the whole data (labelled and unlabelled) .

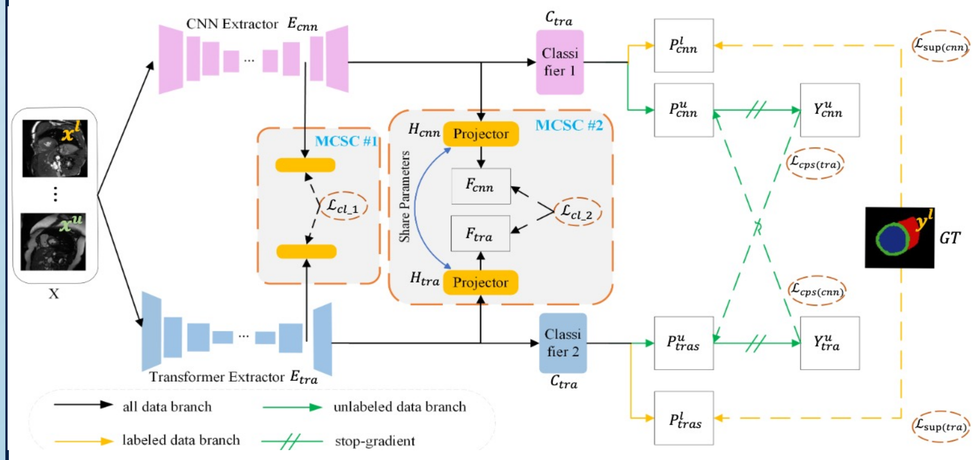

Fig. M1 The overall architecture of our MCSC framework.

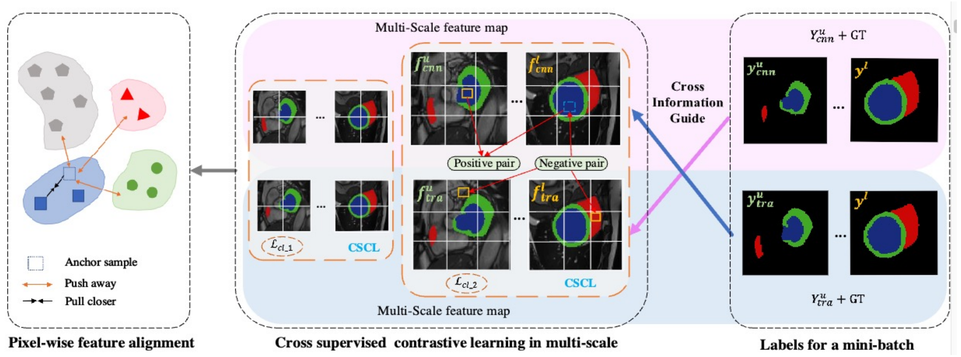

Fig. M2 CST Multi-scale cross supervised contrastive learning. Pseudo labels from cross-teaching (right) are combined with ground-truth, and used to define a local contrastive loss over features of different scales (middle, orange dashed boxes)

Results

- Two benchmark datasets: ACDC contains 200 short-axis cardiac MRI with masks of the left ventricle (LV), myocardium (Myo), and right ventricle (RV). Synapse contains abdominal CT scans from 30 cases with eight organs including aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas and stomach.

Labelled	Methods	Mean DSC↑	Mean HD↓	Myo DSC↑	Myo HD↓	LV DSC↑	LV HD↓	RV DSC↑	RV HD↓
70 cases (100%)	UNet-FS	91.7	4.0	89.0	5.0	94.6	5.9	91.4	1.2
	BATFormer [16]	92.8	8.0	90.26	6.8	96.30	5.9	91.97	11.3
7 cases (10%)	UNet-LS	75.9	10.8	78.2	8.6	85.5	13.0	63.9	10.7
	CCT [19]	84.0	6.6	82.3	5.4	88.6	9.4	81.0	5.1
	CPS [8]	85.0	6.6	82.9	6.6	88.0	10.8	84.2	2.3
	CTS [17]	86.4	8.6	84.4	6.9	90.1	11.2	84.8	7.8
	MCSC (Ours)	**89.4**	**2.3**	**87.6**	**1.1**	**93.6**	**3.5**	**87.1**	**2.1**
3 cases (5%)	UNet-LS	51.2	31.2	54.8	24.4	61.8	24.3	37.0	44.4
	CCT [19]	58.6	27.9	64.7	22.4	70.4	27.1	40.8	34.2
	CPS [8]	60.3	25.5	65.2	18.3	72.0	22.2	43.8	35.8
	CTS [17]	65.6	16.2	62.8	11.5	76.3	15.7	57.7	21.4
	MCSC (Ours)	**73.6**	**10.5**	**70.0**	**8.8**	**79.2**	**14.9**	**71.7**	**7.8**
1 case	UNet-LS	26.4	60.1	26.3	51.2	28.3	52.0	24.6	77.0
	CTS [17]	46.8	36.3	55.1	**5.5**	64.8	**4.1**	20.5	99.4
	MCSC (Ours)	**58.6**	**31.2**	**64.2**	13.3	**78.1**	12.2	**33.5**	**68.1**

Best is reported as bold, Second Best is underlined.

Tab. R1 Segmentation results on the ACDC dataset.

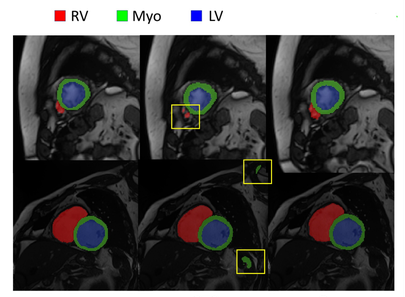

Fig. R1 Visualizations on the ACDC.

Labelled	Methods	DSC↑	HD↓	Aorta	Gallb	Kid_L	Kid_R	Liver	Pancr	Spleen	Stom
18 cases(100 %)	UNet-FS	75.6	42.3	88.8	56.1	78.9	72.6	91.9	55.8	85.8	74.7
	nnFormer [39]	86.6	10.6	92.0	70.2	86.6	86.3	96.8	83.4	90.5	86.8
4 cases(20 %)	UNet-LS	47.2	122.3	67.6	29.7	47.2	50.7	79.1	25.2	56.8	21.5
	CCT [19]	51.4	102.9	71.8	31.2	52.0	50.1	83.0	32.5	65.5	25.2
	CPS [8]	57.9	62.6	75.6	41.4	60.1	53.0	88.2	26.2	69.6	48.9
	CTS [17]	64.0	56.4	**79.9**	38.9	66.3	63.5	86.1	41.9	75.3	60.4
	MCSC (Ours)	**68.5**	**24.8**	76.3	**44.4**	**73.4**	**72.3**	**91.8**	**46.9**	**79.9**	**62.9**
2 cases(10 %)	UNet-LS	45.2	55.6	66.4	**27.2**	46.0	48.0	82.6	18.2	39.9	33.4
	CCT [19]	46.9	58.2	66.0	26.6	53.4	41.0	82.9	21.2	48.7	35.6
	CPS [8]	48.8	65.6	70.9	21.3	58.0	45.1	80.7	23.5	58.0	32.7
	CTS [17]	52.0	63.7	73.2	12.7	67.2	64.7	82.9	31.7	40.9	42.4
	MCSC (Ours)	**61.1**	**32.6**	**73.9**	26.4	**69.9**	**72.7**	**90.0**	**33.2**	**79.4**	**43.0**

Best is reported as bold, Second Best is underlined.

Tab. R2 Segmentation results on the Synapse dataset.

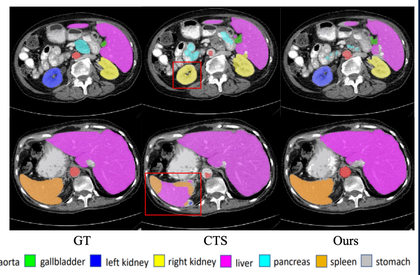

Fig. R2 Visualizations on the Synapse.

Supervision loss functions

- cross pseudo supervision loss (unlabelled data)

$$\mathcal{L}_{cps(cnn)} = \mathcal{L}_{dice}(P^u_{cnn}, Y^u_{tra}), \quad \mathcal{L}_{cps(tra)} = \mathcal{L}_{dice}(P^u_{tra}, Y^u_{cnn}).$$

- Multi-Scale Contrastive loss(whole data): $\mathcal{L}_{cl} = (\mathcal{L}_{cl\_1} + \ldots + \mathcal{L}_{cl\_n})$, each scale $\mathcal{L}_{bcl}$ as $\mathcal{L}_{cl\_i}$

Balanced contrastive loss:

$$\mathcal{L}_{bcl} = -\frac{1}{|A|}\sum_{a_i \in A}\frac{1}{|A_y|-1}\sum_{p \in A_y \setminus \{i\}}\log\frac{\exp(a_i \cdot a_p/\tau)}{\sum_{j \in Y_A}\frac{1}{|A_j|}\sum_{a_k \in A_j}\exp(a_i \cdot a_k/\tau)},$$

- Total loss function:

$$\mathcal{L}_{cnn} = \mathcal{L}_{sup(cnn)} + w_{cps}\mathcal{L}_{cps(cnn)} + w_{cl}\mathcal{L}_{cl} \qquad \mathcal{L}_{tra} = \mathcal{L}_{sup(tra)} + w_{cps}\mathcal{L}_{cps(tra)} + w_{cl}\mathcal{L}_{cl}$$

References

1. Xiangde Luo etal. 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.
2. Wenguan Wang, etal. Exploring cross-image pixel contrast for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7303–7313, 2021.
3. ianggang Zhu, etal. Balanced contrastive learning for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6908–6917, 2022.

Acknowledgements

We acknowledge funding by China Scholarship Council, EPSRC (EP/W01212X/1) and Royal Society (RGS/R2/212199).

University of Glasgow | School of Computing Science

Contact: 2665227l@student.gla.ac.uk