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

Qianying Liu[1]
2665227L@student.gla.ac.uk
Xiao Gu[2]
xiao.gu17@imperial.ac.uk
Paul Henderson[1]
paul.henderson@glasgow.ac.uk
Fani Deligianni[1]
fani.deligianni@glasgow.ac.uk

[1] School of Computing Science
University of Glasgow
Glasgow, UK

[2] Department of Computing
Imperial College London
London, UK

Abstract

Semi-supervised learning has demonstrated great potential in medical image segmentation by utilizing knowledge from unlabeled data. However, most existing approaches do not explicitly capture high-level semantic relations between distant regions, which limits their performance. In this paper, we focus on representation learning for semi-supervised learning, by developing a novel Multi-Scale Cross Supervised Contrastive Learning (MCSC) framework, to segment structures in medical images. We jointly train CNN and Transformer models, regularising their features to be semantically consistent across different scales. Our approach contrasts multi-scale features based on ground-truth and cross-predicted labels, in order to extract robust feature representations that reflect intra- and inter-slice relationships across the whole dataset. To tackle class imbalance, we take into account the prevalence of each class to guide contrastive learning and ensure that features adequately capture infrequent classes. Extensive experiments on two multi-structure medical segmentation datasets demonstrate the effectiveness of MCSC. It not only outperforms state-of-the-art semi-supervised methods by more than 3.0% in Dice, but also greatly reduces the performance gap with fully supervised methods. Our code is available at https://github.com/kathyliu579/MCSC.

1 Introduction

Image segmentation serves as a fundamental process in medical image analysis by delineating organ structures and allowing the quantification of their shape and size, thus providing essential information for clinical diagnostics, treatment planning, and patient monitoring [5, 23]. Deep learning approaches have achieved great successes in medical image segmentation in recent years; however, such techniques hinge upon the availability of large-scale and accurately annotated datasets [24]. In the medical domain, such datasets require prohibitive time, cost, and expertise to obtain. To mitigate this issue, *semi-supervised* learning

(SSL) aims to minimize the annotation efforts by training with both labelled and unlabelled data [17, 18, 38].

Several strategies have been proposed for SSL in medical image segmentation. These include iterative pseudo-labeling [22], regularization strategies [8, 17, 18, 28, 32], as well as leveraging domain-specific prior knowledge such as anatomical information [35]. Typically pseudo-labeling iteratively generates approximate segmentation masks for unlabeled data. Integrating these pseudo annotations with ground truth labels for model updates necessitates a meticulously designed approach, which remains an open problem. Differently, several regularization approaches forgo this process, by enforcing prediction consistency over different data transformations [8, 32], different model architectures [17, 18], or different tasks [28]. In particular, recent works [17] have investigated the possibility of making use of two advanced segmentation backbones, e.g., CNN and Transformer, for cross-teaching SSL.

Although these methods are promising, their performance is significantly weaker than fully supervised approaches and thus their practical application in medical image segmentation is limited [21, 25, 36]. To alleviate this issue, *contrastive learning* has been extensively utilized to facilitate robust feature learning. It functions by encouraging feature similarity of positive pairs, as well as dissimilarity of negative pairs. Positive pairs may be defined in a self-supervised manner as different augmentations of the same instance [9] or in a supervised manner based on the actual label [13]. In SSL, pioneering works [3, 31] have made efforts towards directly applying contrastive learning on unlabelled data, by performing global-level image contrast for training. However, this strategy is mostly suited for classification tasks, since it extracts global representations that ignore detailed pixel-level information. To accurately delineate organ boundaries, a local contrastive strategy is required to enable predictions at a pixel level [3, 29, 38]. In particular, for image segmentation that inherently relies on dense-wise prediction, Chaitanya *et al*. highlighted the importance of complementing the global image-level contrast with local pixel-level contrast [3].

Since self-supervised contrastive learning normally select augmented views of the same sample data point as positive pairs [9], without prior knowledge of the actual class label and its prevalence, it is prone to a substantial number of false negative pairs, particularly when dealing with class-imbalanced medical imaging segmentation datasets [15]. To mitigate the false negative predictions resulting from self-supervised local contrastive learning, existing works have investigated supervised local contrastive learning [9, 12]. Pioneering works [31, 37] applied supervised contrastive learning only on unlabelled data based on conventional iterative pseudo annotation. Some studies [12] also attempted to apply supervised local contrastive loss on labelled data exclusively, whilst performing self-supervised training for unlabelled data. However, the discrepancy in positive/negative definitions leads to divergent optimization objectives, which may yield suboptimal performance.

We propose a novel multi-scale cross contrastive learning framework for semi-supervised medical image segmentation. Both labelled and unlabelled data are integrated seamlessly via cross pseudo supervision and balanced, local contrastive learning across features maps that span multiple spatial scales. Our main contributions are three-fold:

- We introduce **a novel SSL framework** that combines the benefits of cross-teaching with a proposed local contrastive learning. This enhances training stability, and beyond this, ensures semantic consistency in both the output prediction and the feature level.
- We develop the first **local contrastive framework** defined over **multi-scale feature maps**, which accounts for over-locality and over-fitting typical of pixel-level contrast. This benefits from seamlessly unifying pseudo-labels and ground truth via cross-teaching.

- We incorporate a **balanced contrastive loss** which is normalised based on the prevalence of each class to enforce **unbiased representation learning** in SSL medical image segmentation. This tackles the significant imbalance issue for both pseudo label prediction, and the concurrent supervised training based on imbalanced (pseudo) labels.

We evaluate our proposed methodology on two challenging benchmarks of radiological scans: multi-structure MRI segmentation on ACDC [1], and multi-organ CT segmentation on Synapse [14]. Our approach not only significantly outperforms state-of-the-art SSL methods, but also closes the gap between fully supervised approaches with just a small fraction of labelled data. With just 10% labelled data, it achieves remarkable improvement in Hausdorff Distance (HD) from 8.0 to 2.3mm. Our method is also more resilient to the reduction of labelled cases, achieving around 10% improvement in Dice Coefficient (DSC) when labelled data are reduced from 10% to 5% in ACDC and from 20% to 10% in Synapse.

2 Related Work

Consistency Regularization in Semi-Supervised Medical Image Segmentation. Semi-supervised learning has gained popularity in medical image segmentation due to its effectiveness in handling scenarios with limited annotations [2, 5, 17, 20]. Among various approaches, enforcing prediction consistency has emerged as a crucial regularization strategy for extracting and leveraging knowledge from unlabelled data. Such regularization can be based on predictions from different augmentations [2, 20], different architectures [17], or tasks [28]. For instance, inspired by the fact that the predicted mask should undergo the same spatial transformations as the input images, Bortsova *et al*. [2] developed a transformation consistency based semi-supervised framework. Peng *et al*. [20] sought to attain prediction similarity from a batch of co-trained models with identical architectures, while adversarially preserving each model's diversity. Recent works [17] has taken advantage of the advanced U-Net and Transformer, and aimed to achieve the prediction consistency from networks. However, the medical image datasets are typically imbalanced, which poses great challenges in learning unbiased predictions with limited annotations [15]. Tackling such issue in consistency settings for unlabelled data remains an open problem. Furthermore, existing works primarily focus on prediction consistency at the output level [17], neglecting the pursuit of discriminative feature representations for both labelled and unlabelled data.

Contrastive Learning in Medical Image Segmentation. Contrastive learning has contributed to most successful self-supervised visual representation methods [7, 10, 11, 26]. The core idea is to promote the similarity of positive image pairs, whilst distinguishing negative pairs. To tailor for the needs of dense-wise downstream segmentation task, pixel-wise self-supervised contrastive learning has been introduced recently [30, 33]. Recent research has also found that integrating the contrastive loss in both global and local levels, can enhance performance [3]. In the realm of natural images, there has been a growing interest in merging semi-supervised learning with contrastive learning, resulting in a one-stage, end-to-end model that forgoes unsupervised pretraining [34, 38]. This approach has recently been adopted in the medical domain for segmentation tasks [9, 12, 31, 37]. However, as discussed in the Introduction section, existing combinations of contrastive learning and semi-supervised learning do not fully address the inherent challenges posed by size-limited and data-imbalanced medical datasets, thus lacking generality. The question of how to effectively integrate contrastive learning for medical image segmentation remains open.

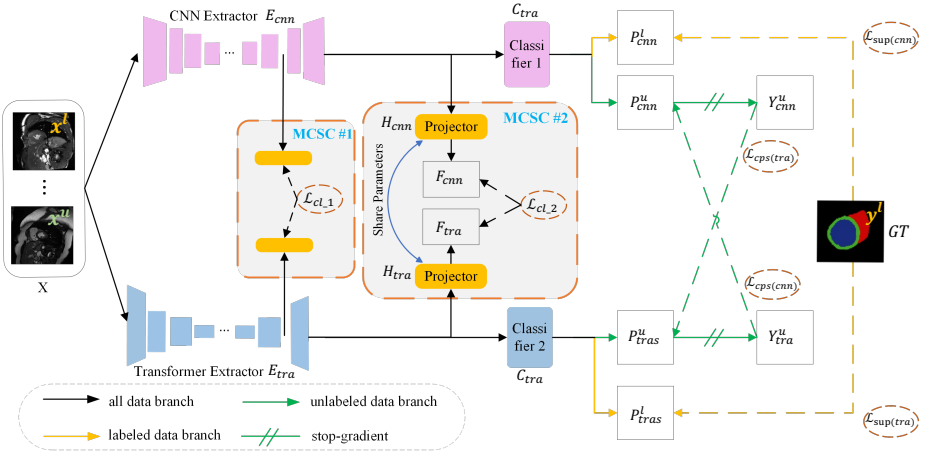

Figure 1: The overall architecture of our MCSC framework for semi-supervised segmentation. Two networks, a CNN (pink) and Transformer (blue), with complementary inductive biases, learn together. When training on unlabelled data, each network generates pseudo labels for the other. These labels are used to define a pseudo supervision loss and a novel local contrastive loss that improves the quality of representations learnt by the models.

3 Method

We adopt a student–student framework based on [17], with cross-teaching between a CNN-based U-Net and a Transformer-based U-Net. This leverages the advantages of convolution-based and Transformer-based segmentation networks for learning local semantic information and long-range dependencies, and enables the two models to achieve consistency on segmentation prediction. However, this framework has some limitations: (i) it only focuses on the prediction consistency on each image slice at output level; (ii) it ignores the dissimilarity and similarity among different and same segmentation categories across the whole dataset. To overcome this, we propose a Multi-Scale Cross Supervised Contrastive Learning (MCSC) framework to pull closer the features of the same category and push away the features of different categories from both networks. It not only ensures the consistency of two models on the feature and output level, but also enhances the distinguishability of features in different categories, thereby improving the segmentation performance. We illustrate the overall architecture of our framework in Figure 1, and provide pseudocode in supplementary section S1. The branch of CNN or Transformer includes a feature extractor $E_*(\cdot)$, a segmentation head $C_*(\cdot)$, and two feature space projectors $H_*(\cdot)$. Both branches only share the parameters of the last layer in the feature space projectors.

Given a training dataset consisting of a small labelled subset $D_l = \{x^l_i, y^l_i\}_{i=1}^K$ and a large unlabelled set $D_u = \{x^u_j\}_{j=1}^M$, where $M \gg K$, the input to our model is a minibatch $X = X^l \cup X^u$ including labelled images and unlabelled images. The minibatch X is first fed into the CNN-based and Transformer-based networks to obtain their feature representations and segmentation logits. In the semi-supervised setting, we employ the following supervision losses for training: (i) on the output level, we calculate the *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 in Figure 1) between the segmentation predictions and the pseudo labels from the CNN-based U-Net or

the Transformer-based U-Net in a cross teaching manner on the output level (Section 3.1); (ii) on the feature level, we employ the proposed multi-scale cross contrastive loss $\mathcal{L}_{cl}$ (black dashed lines in Figure 1) to enhance feature consistency of the same segmentation category and feature distinguishability of the different segmentation categories across the whole dataset (labelled and unlabelled) (Section 3.2).

3.1 Cross Pseudo Supervision

The CNN and Transformer networks teach each other using the unlabelled data, through a cross pseudo supervision loss $\mathcal{L}_{cps}$ [8, 17]. This regularises their respective predictions to be consistent with each other. Specifically, the predictions made by the CNN become pseudo labels that supervise the Transformer, and vice-versa. The unlabelled images X^u are fed into the feature extractors $E_*(\cdot)$ and classifier heads $C_*(\cdot)$ of the two models respectively, to get class probability maps $P^u_* = \text{softmax}\{C_*(E_*(X^u))\}$, and pseudo one-hot label map $Y^u_* = \text{argmax}\,(P^u_*)$, where $*$ denotes the CNN or Transformer branch. We then define two consistency loss terms: $\mathcal{L}_{cps(cnn)}$ uses the Transformer's pseudo-labels to supervise the CNN, and $\mathcal{L}_{cps(tra)}$ the reverse; these are given by:

$$\mathcal{L}_{cps(cnn)} = \mathcal{L}_{dice}(P^u_{cnn}, Y^u_{tra}), \quad \mathcal{L}_{cps(tra)} = \mathcal{L}_{dice}(P^u_{tra}, Y^u_{cnn}). \tag{1}$$

Here $\mathcal{L}_{dice}$ is the standard Dice loss function, but using pseudo-labels instead of ground-truth segmentation. Note that during training there is no gradient back-propagation between P^u_{cnn} and Y^u_{cnn}, and similar from P^u_{tra} to Y^u_{tra}.

3.2 Multi-Scale Cross Supervised Contrastive Learning (MCSC)

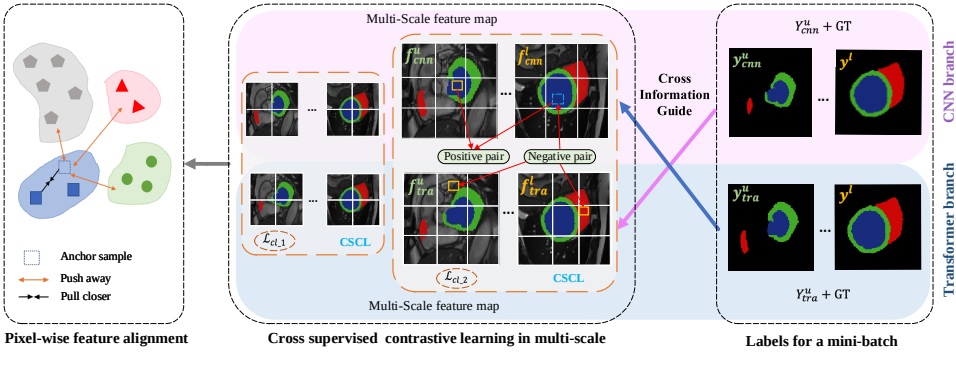

Figure 2: Multi-scale cross supervised contrastive learning. Pseudo labels from cross-teaching (**right**) are combined with ground-truth labels where available, and used to define a local contrastive loss over features of different scales (**middle**, orange dashed boxes). This contrastive pairs of pixels drawn from either the same or different slices; for efficiency it is defined over patches. Features of pixels of the same (pseudo-) class are pulled together (**left**), while those of different classes are pushed apart.

Cross pseudo supervision does not exploit feature regularities across the whole dataset, e.g. similarity between representations of the same organ in different slices. We therefore add a contrastive loss, operating on multi-scale features extracted from the Transformer and the CNN. This has two advantages: (i) It encourages consistency of the two models' internal features (not just outputs) (ii) It captures high-level semantic relationships between distant regions, and between features on both labelled and unlabelled data.

Our MCSC module (Figure 2) is based on local supervised contrastive learning [29], which learns a compact feature space by reducing the distance in the embedding space between positive pairs, and increasing the distance between negative pairs. Firstly, it extracts features from the CNN and Transformer, then projects them into a common embedding space. This is followed by a novel approach of selecting positive and negative pairs using the pseudo labels, and a class-balanced contrastive loss calculated on these.

Feature Embedding. After $X = \{x_i\}_{i=1...N}$ is passed into $E_{cnn}(\cdot)$ and $E_{tra}(\cdot)$ respectively, the resulting features are projected by passing them through projectors $H_{\text{cnn}}(\cdot)$ and $H_{\text{tra}}(\cdot)$ into a unified feature space, where we will sample pairs to contrast. Overall, we get a feature batch F consisting of $2N$ feature maps $f_i = H(E(x_i)) \in \mathbb{R}^{h \times w \times c}$, where $f_{1...N}$ come from the CNN and $f_{N+1...2N}$ from the Transformer (middle of Figure 2).

Cross Supervised Sampling. For cross supervised sampling, we follow these strategies: (i) We exchange class information from two models to guide the sampling, using the prediction of Transformer to be the supervisory information for CNN and vice-versa (Figure 2, right). This is consistent with the cross-prediction loss $\mathcal{L}_{cps}$, and implicitly it also makes the features predicted by the two models on the same slice consistent. (ii) We contrast features on both unlabelled and labelled data. Since the pseudo labels are of varying quality, labelled data is included in the contrastive loss to reduce the noise. (iii) We contrast pixels both within and between slices. Previous work focuses on inter-slice samples and ignores useful anatomical information within slices. For example, compared to different slices, the features of different class of organ boundaries in the image should be more similar. By focusing on them, we can refine the details of the hardest boundary segmentation. Therefore, our strategy differs significantly from existing approaches to sampling pairs in supervised contrastive learning with semi-supervised segmentation, where positive or negative pairs are selected based on pseudo labels on unlabelled data [4, 31, 37].

The computational complexity and memory for the supervised contrastive loss is very high; however, comparing many samples is crucial for improving the performance of contrastive learning [11]. To address this problem, inspired by [12], we compute the local contrastive loss over patches. We divide all the feature maps in F into patches with size of $h' \times h'$. Let us assume there are M patches of each f. We randomly select (without replacement) a patch from each feature map in F, and finally we get M batches of $2N$ patches. The loss is evaluated on $2N$ patches from each batch in turn, until the entire f has been traversed.

Balanced Supervised Local Contrastive Loss. After sampling positive/negative pairs of pixels, a contrastive loss is introduced to pull positive pairs closer and push negative pairs apart within the $2N$ patches. Given the extreme imbalance between background and foreground (different organs), a randomly sampled batch tends to consist of a significantly larger number of positive and negative pairs for the background, compared to the foreground organs. This inherent imbalance inevitably biases conventional supervised contrastive learning towards the background, consequently neglecting the differentiation of foreground categories. Simply eliminating the background during contrastive learning [12] is not an optimal solution, as (i) the remaining number of foreground pixels is extremely small, and (ii) this fails to capture the relationship between the background and the foreground.

Inspired by [40], we average both the inter-class (positive) and intra-class (negative) feature contrast within the pixels of each class, and then forward it to calculate the supervised contrastive loss. In this way, each class makes an approximately balanced contribution. This

balanced contrastive loss is implemented as follows:

$$\mathcal{L}_{bcl} = -\frac{1}{|A|}\sum_{a_i \in A}\frac{1}{|A_y|-1}\sum_{p \in A_y \setminus \{i\}}\log\frac{\exp(a_i \cdot a_p/\tau)}{\sum_{j \in Y_A}\frac{1}{|A_j|}\sum_{a_k \in A_j}\exp(a_i \cdot a_k/\tau)}, \tag{2}$$

where A is the pixel-level feature sets of the $2N$ patches, a_i represents the i^{th} feature, A_y is a subset that contains all samples of class y, $A_y \setminus \{i\}$ represents all the pixels in A_y excluding a_i, Y_A represents the set of all the unique classes in current A, and τ is a temperature constant. By balancing the contribution of each class during contrastive learning, we avoid the learned representations being biased towards the dominant background. Note that $\mathcal{L}_{bcl}$ is calculated over each $2N$ patches, and then averaged over M batches of $2N$ patches for back-propagation.

Multi-Scale Contrastive Loss. Existing works on local contrastive learning pass the features of the last layer before the classifier into the projector. However, the feature maps from earlier layers focus on coarser geometric information like the shape of organs, and later feature maps on details; both are important for segmentation, which depends both on relationships among multiple organs and gross anatomic structure (global) and textures of the specific tissues (local). We therefore pass features with n different scales from n layers of extractors and separate projectors, and then calculate each scale balanced contrastive loss $\mathcal{L}_{bcl}$ as $\mathcal{L}_{cl\_i}$. The overall loss $\mathcal{L}_{cl}$ is given by summing over each scale loss: $\mathcal{L}_{cl} = (\mathcal{L}_{cl\_1} + \ldots + \mathcal{L}_{cl\_n})$.

3.3 Optimization

The two networks are trained to minimize a weighted sum of the losses described in the previous sections: $\mathcal{L}_{cnn} = \mathcal{L}_{sup(cnn)} + w_{cps}\mathcal{L}_{cps(cnn)} + w_{cl}\mathcal{L}_{cl}$ and $\mathcal{L}_{tra} = \mathcal{L}_{sup(tra)} + w_{cps}\mathcal{L}_{cps(tra)} + w_{cl}\mathcal{L}_{cl}$, where w_* are weighting factors used to balance the impact of individual loss terms. w_{cps} is defined by a Gaussian warm-up function [17]: $w_{cps}(t_i) = 0.1 \cdot e^{(-5(1-t_i/t_{total})^2)}$, where t_i is i^{th} iteration of training and t_{total} is the total number of iterations, while w_{cl} is set to a constant value of 10^{-3} based on performance of the validation. Note that the Transformer is used only during training, and does not contribute to the final inference – the CNN is less computationally expensive, but has distilled the Transformer's knowledge.

4 Experiments

We evaluate our method on two benchmark datasets, ACDC [1] and Synapse [14]. ACDC contains 200 short-axis cardiac MR images from 100 cases (i.e. patients) with masks of the left ventricle (LV), myocardium (Myo), and right ventricle (RV) to be segmented; we follow the data split and the selection of labelled cases in [17]. Synapse contains abdominal CT scans from 30 cases with eight organs including aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas and stomach; the splits follow [6]. To quantitatively assess performance, we report two popular metrics: Dice coefficient (DSC) and 95% Hausdorff Distance (HD). Further implementation details for our method and the baselines, as well as more results, are given in the supplementary section S2 and S3.

4.1 Comparison with Other Semi-Supervised Methods

We compare our proposed method to several recent SSL methods that use U-Net as backbone, including Mean Teacher (MT) [25], Deep Co-Training (DCT) [21], Uncertainty Aware Mean Teacher (UAMT) [36], Interpolation Consistency Training (ICT) [27], Cross Consistency Training (CCT) [19], Cross Pseudo Supervision (CPS) [8], and the state-of-the-art

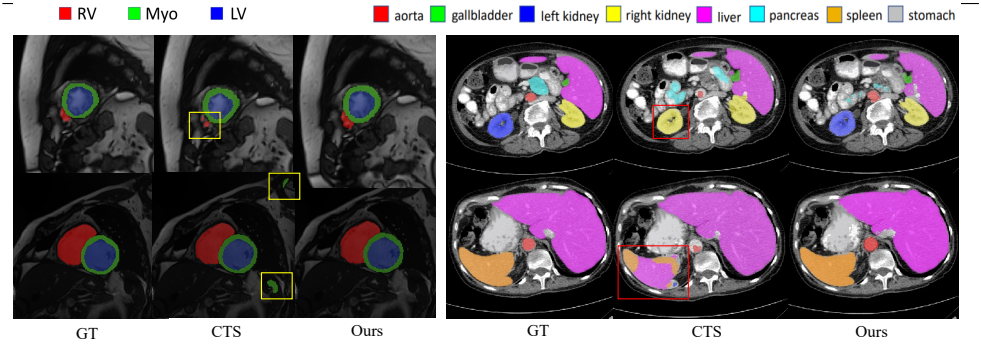

Figure 3: Qualitative results from our method and the best baseline CTS [17] trained on 4 and 7 labelled cases on ACDC (**left**) and Synapse (**right**), respectively.

(SOTA) method Cross Teaching Supervision (CTS) [17]. Results for the weaker methods MT, DCT and ICT are given in the supplementary material. We also compare against a U-Net trained with full supervision (UNet-FS), and one trained only on the labelled subset of data (UNet-LS). Finally we compare with the SOTA fully-supervised Transformer based methods BATFormer [16] on ACDC, and nnFormer [39] on Synapse. We retrained all the semi-supervised baselines using their original settings (optimizer and batch size), and report whichever is better of our retrained model or the result quoted in [17].

Labelled	Methods	Mean		Myo		LV		RV	
		DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	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 [15]	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 [15]	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.

Table 1: Segmentation results on DSC(%) and HD(mm) of our method and baselines on ACDC, across different numbers of labelled cases.

Results on ACDC. Table 1 shows evaluation results of MCSC and the best-performing baseline under three different levels of supervision (7, 3 and 1 labelled cases). Our MCSC method trained on 10% of cases improves both DSC and HD metrics compared to previous best SSL methods by a significant margin (more than 3% on DSC and 5mm on HD). More importantly, it achieves 2.3mm HD, significantly better than even the fully supervised U-Net and BATFormer, which achieve 4.0 and 8.0 respectively. It also demonstrates competitive DSC of 89.4 %, compared with 91.7 % and 92.8 % of U-Net and BATFormer. In addition, MCSC performance is highly resilient to the reduction of labelled data from 10% to 5%, outperforming the previous SOTA SSL methods by around 10% on DSC. The improvement is even more profound for the minority and hardest class, RV, with performance gains of 14 % on DSC and 13.6mm on HD. Figure 3 shows qualitative results from UNet-LS, CPS,

CTS and our method. MCSC produces a more accurate segmentation, with fewer under-segmented regions on minority class- RV (top) and fewer false-positive (bottom). Overall, results prove that MCSC improving the semantic segmentation capability on unbalanced and limited-annotated medical image dataset by a large margin.

Results on Synapse. Table 2 shows the segmentation results of the best-performing baselines on Synapse with 4 and 2 labelled cases. Compared to ACDC, Synapse is a more challenging segmentation benchmark as it includes a larger number of labelled regions with far more imbalanced volumes. Nevertheless, our method outperforms the baselines by a large margin. This demonstrates the robustness of our proposed framework, and the benefit of regularising multi-scale features from two models to be semantically consistent across the whole dataset. This is further highlighted in the qualitative results provided in Figure 3.

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.

Table 2: Comparison with different models on Synapse. The performance is reported by class-mean DSC (%) and HD (mm), as well as the DSC value for each organ.

4.2 Ablation study

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

Table 3: 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	Branches 56	Branches 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

Table 4: Ablation analysis on the choice of feature maps for the multi-scale contrastive loss (ACDC, 7 labelled cases). Full table is in the supplementary material.

In Table 3 we explore the influence of proposed modules on the performance on ACDC with 7 labelled cases. Starting from CTS [17] (top row), and adding supervised local contrastive learning (SCL) with a prior approach for balancing the loss (DB [12]), we observe a significant improvement of 1.1% on DSC; this emphasizes the importance of enforcing consistency between features of the two models. By exchanging class information from CNN and Transformer to select contrasted samples (instead of using each model's own predictions as pseudo-labels), we see an improvement in DSC and HD from 87.50 to 88.23 and 7.4 to

3.4 respectively. Our approach to balancing different classes (Balanced), instead of just discarding background pixels (DB), improves DSC by 0.7%, since minority classes are better separated. Finally, utilizing multi-scale instead of just final-layer features further improves performance by 0.58% and 2.3% DSC and HD respectively. In Table 4, we compare results using different feature maps as input to the contrastive loss; we see best performance is achieved by using both 256×256 and 28×28 feature maps. Thus, combining coarser geometric information in global features and detailed local features does indeed benefit medical image segmentation.

5 Computational Complexity

Theoretical complexity of patch-level contrastive learning. Existing works subsample a smaller set of pixel coordinates as positive pairs to fit in GPU memory [4]. However, using more samples to compare is crucial for improving the performance of contrastive learning [11]. Without subsampling, the overall computational complexity for the supervised local loss is $O(h^4)$, where h is the size of an image, 256 in our case, which would necessitate $O(10^9)$ multiplications. Our proposed approach uses patches with size of $h' \times h'$ to do contrastive learning. This reduces the computational complexity from $O(h^4)$ to $O((h/h')^2 \cdot h'^4)$ and alleviates out-of-memory issues. If we set $h' = 19$, complexity will be $O(10^7)$.

Practical calculation time for different methods. We compare the computational cost of different methods on ACDC using a single Nvidia RTX 3090 GPU. 'ForwardT' refers to the number of times each image needs to be processed through the network during one training iteration. 'BatchT' refers to the training time (in seconds) for a single minibatch (two labelled and two unlabelled images) processed during one iteration, including forward pass, loss calculation, and backward pass. 'InferenceT' refers to the inference time for a single image (in seconds). For our method, we give the inference time of the CNN (pink) and the Transformer (blue); recall however in practice, we use only the CNN during testing.

		MT	UAMT	CCT	CPS	CTS	Ours
Train	ForwardT/image	2	6	1	2	2	2
	BatchT/ batch	0.10	0.16	0.21	0.17	0.22	0.83
Test	InferenceT/ case	0.56	0.56	0.75	0.56	0.56	0.58/0.87
	Gflops/ image	3.00	3.00	8.77	3.00	3.00	3.00/6.03

Table 5: Comparison of the computational cost of different models.

6 Conclusion

We have presented a novel SSL framework for medical image segmentation based on cross-teaching between a Transformer and a CNN. This incorporates a supervised local contrastive loss, named MCSC, that encourages intra-class feature similarity and inter-class discriminativity across the whole dataset. Furthermore, it addresses class imbalance with a loss that eliminates the negative effects of excessive background pixels. Finally, it contrasts multi-scale feature maps, to combine global and local feature understanding. Our experiments on two commonly used medical datasets demonstrate that the proposed framework can fully take advantage of labelled and unlabelled data, and demonstrates remarkably resiliant performance even when the labelled data are significantly reduced.

7 Acknowledgements

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

References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[2] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 810–818. Springer, 2019.

[3] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020.

[4] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Medical Image Analysis*, 87:102792, 2023.

[5] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in Cardiovascular Medicine*, 7:25, 2020.

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] 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.

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[12] Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 481–490. Springer, 2021.

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[14] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[15] Zeju Li, Konstantinos Kamnitsas, and Ben Glocker. Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE transactions on medical imaging*, 40(3):1065–1077, 2020.

[16] Xian Lin, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. Batformer: Towards boundary-aware lightweight transformer for efficient medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[17] 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.

[18] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[19] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[20] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107:107269, 2020.

[21] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018.

[22] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 2171–2179, 2022.

[23] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[24] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[25] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[26] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[27] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. *Neural Networks*, 145:90–106, 2022.

[28] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79:102447, 2022.

[29] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021.

[30] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[31] Huisi Wu, Zhaoze Wang, Youyi Song, Lin Yang, and Jing Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11666–11675, 2022.

[32] Yingda Xia, Fengze Liu, Dong Yang, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3646–3655, 2020.

[33] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[34] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14421–14430, 2022.

[35] J. Yang, Y. Tao, Q. Xu, Y. Zhang, X. Ma, S. Yuan, and Q. Chen. Self-supervised sequence recovery for semi-supervised retinal layer segmentation. In *IEEE Journal of Biomedical and Health Informatics*, pages 3872–3883. IEEE, 2022.

[36] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium 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 605–613. Springer, 2019.

[37] Xiangyu Zhao, Zengxin Qi, Sheng Wang, Qian Wang, Xuehai Wu, Ying Mao, and Lichi Zhang. Rcps: Rectified contrastive pseudo supervision for semi-supervised medical image segmentation. *arXiv preprint arXiv:2301.05500*, 2023.

[38] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021.

[39] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.

[40] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. 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.