Structure-aware Semantic Discrepancy and Consistency for 3D Medical Image Self-supervised Learning

Tan Pan[1,2] Zhaorui Tan[2] Kaiyu Guo[3,2] Dongli Xu[2] Weidi Xu Chen Jiang[2] Xin Guo[2] Yuan Qi[1,2,4] Yuan Cheng[1,2]

[1]Artificial Intelligence Innovation and Incubation Institute, Fudan University [2]Shanghai Academy of Artificial Intelligence for Science
[3]The University of Queensland [4]Zhongshan Hospital, Fudan University

Takeaway

- **A novel insight of intra-structure consistency and inter-structure discrepancy** in the anatomical structure-aware feature learning in 3D medical images.
- We propose S^2DC, a novel training framework that enhances interstructure discrepancy and intra-structure consistency. The framework establishes reliable **patch-to-patch** correspondences to reinforce discrepancy while leveraging **patch-to-structure** semantic connectivity from the similarity distribution to improve consistency.
- Our method demonstrates superior performance over SOTA medical image SSL methods, evaluated across 10 datasets, 4 tasks, and 3 imaging modalities. The code is available at https://github.com/Ashespt/S2DC/tree/main.

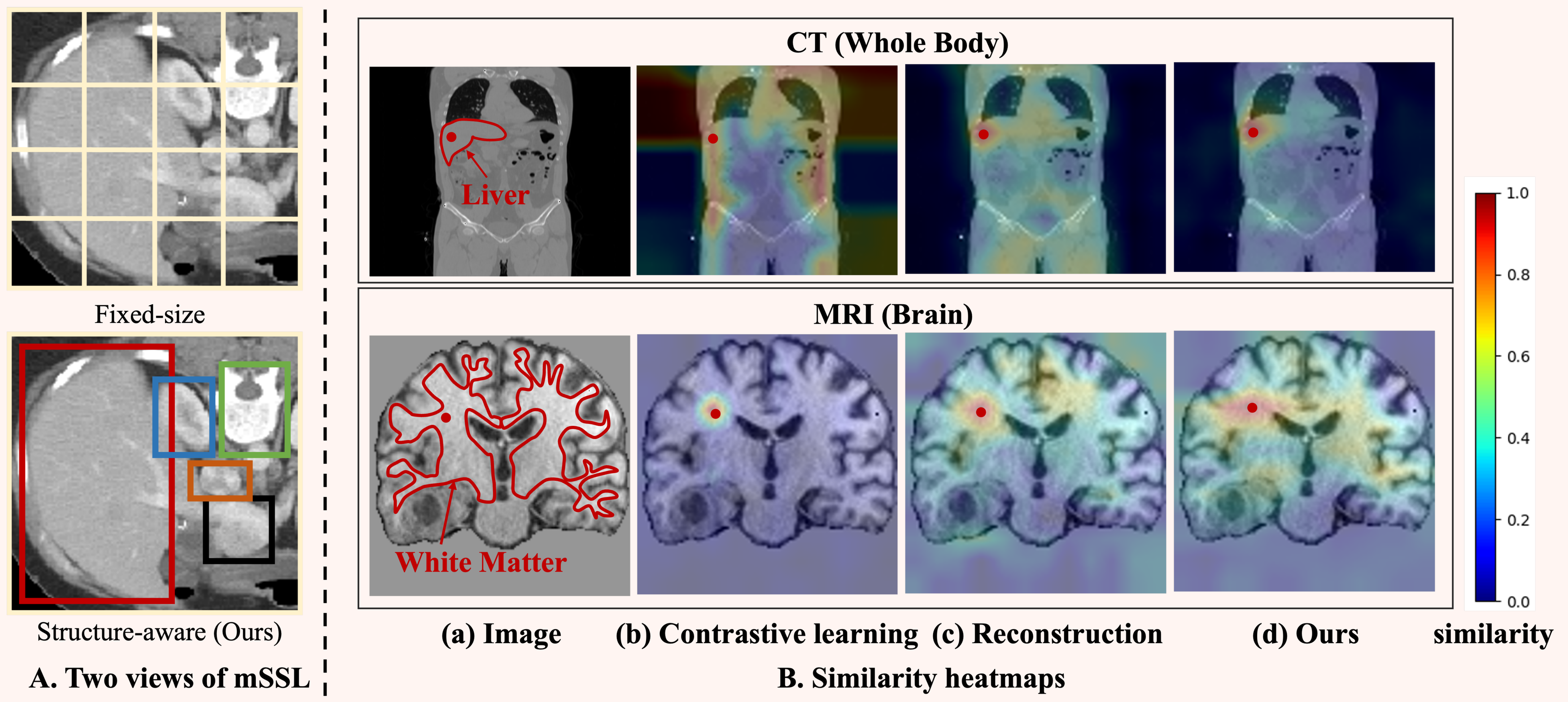

Figure 1. A. Two views of mSSL. B. Similarity heatmaps on CT and MRI images across different methods. We sample an anchor patch (red dots) and compute feature similarities with all other patches. In the first row, the liver anchor should show low similarity with non-liver patches, while in the second row, the white matter anchor should exhibit high similarity with other white matter patches. Current SOTA contrastive-based (b) and reconstruction-based (c) methods struggle with both patch feature discrepancy in different structures and consistency in the same structure. In contrast, our method (d) advances both discrepancy and consistency.

Experiments and Analysis

Method	Accuracy(%) 10%	50%	100%
Swin-UNETR	77.15	92.28	94.15
SwinMM	87.87	93.50	94.80
VoCo	86.73	92.43	94.60
S^2DC	88.29	93.89	95.34

Table 1. Experiment results on CC-CCII with various ratios of the training data. 10%, 50%, and 100% represent ratios.

Baseline($\mathcal{L}_g$)	$+\mathcal{L}_{p2p}$	$+\mathcal{L}_{p2s}$	BTCV DICE(%)	AUTOPET DICE(%)
•			83.43	45.72
•	•		83.98	45.85
•		•	84.02	46.23
•	•	•	84.14	46.47

Table 2. The ablation results of different constraints.

Seeking postdoctoral positions and collaborations in AI4Healthcare.
pant23@m.fudan.edu.cn

Method: From Patch to Structure

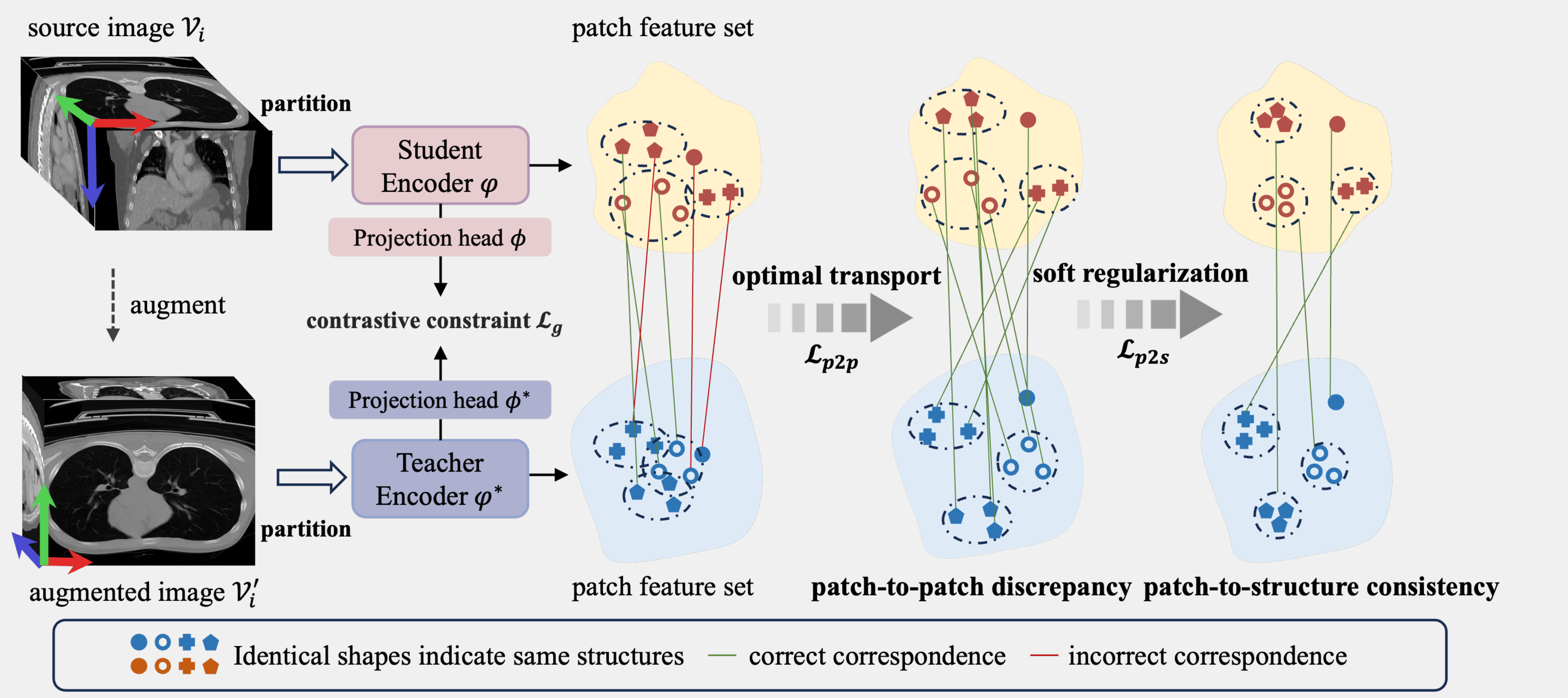

Figure 2. The pipeline of our SSL framework. S^2DC is established on patch features (i.e., token feature in vision transformer) and incorporates two main steps: (1) Patch-to-patch discrepancy. (2) Patch-to-structure consistency.

Stage 1: Patch-to-patch correspondence

Given two patch centers $c_i = (x_i, y_i, z_i)$ and $c_j = (x_j, y_j, z_j)$ from a volume $\mathcal{V}_i$ and its augmented $\mathcal{V}_i'$, we have the GT correspondence:

$$\mathcal{M}_{gt}(i,j) \triangleq \begin{cases} 1 \; if \; \langle H(c_i), c_j \rangle \wedge \left\langle H^{-1}(c_j), c_i \right\rangle, \\ 0 \; else. \end{cases} \tag{1}$$

Then, we can calculate the similarity map $\mathcal{M}_t$ between tokens and get the loss between $\mathcal{M}_t$ and $\mathcal{M}_{gt}$. By applying the dual-softmax operator:

$$\hat{\mathcal{M}}_t(i,j) = softmax(\mathcal{M}_t(i,\cdot)) \cdot softmax((\mathcal{M}_t(\cdot,j)). \tag{2}$$

The patch-to-patch loss $\mathcal{L}_{p2p}$ is:

$$\mathcal{L}_{p2p} = -\frac{1}{|\mathcal{M}_{gt}|} \sum_{i=0}^{N} \sum_{j=0}^{N} \mathcal{M}_{gt}(i,j) \times log(\hat{\mathcal{M}}_t(i,j)), \tag{3}$$

Stage 2: Patch-to-structure semantic connectivity

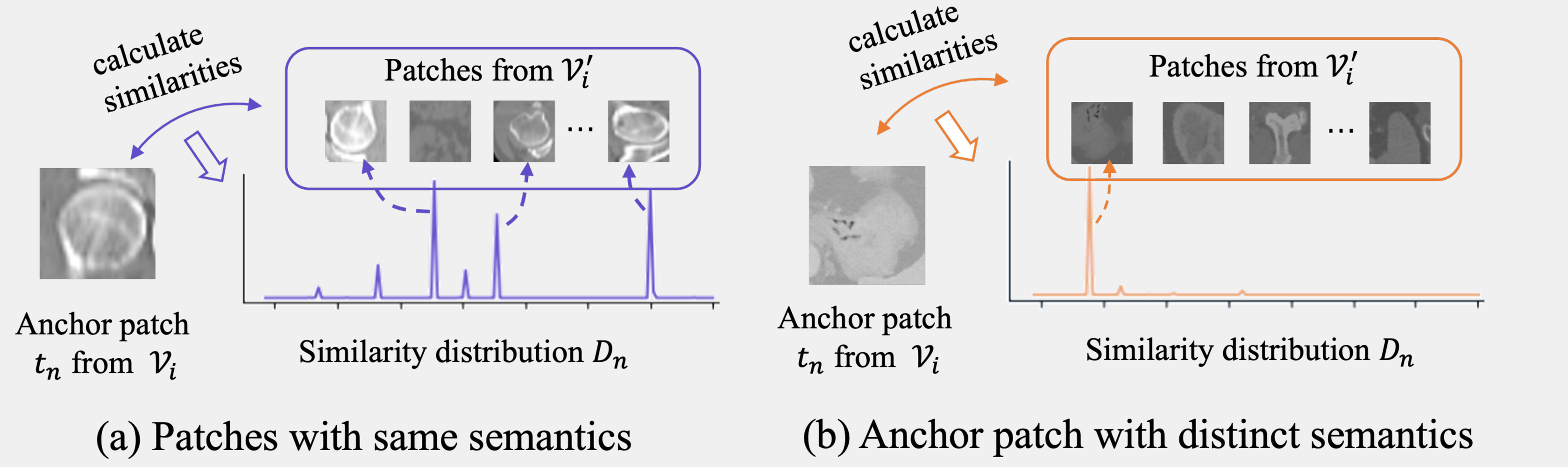

Figure 3. Illustration of the similarity distribution D_n. (a) Patches with the same semantics (e.g., bone). Given an anchor patch, patches from the same semantics form several peaks in D_n. (b) Anchor patch with distinct semantics (e.g., pancreas), D_n shows only one large peak (its augmented patch).

Method: From Patch to Structure

Stage 2 We define neighborhood similarity distribution as $\mathcal{D}_n$, which represents the similarity vector of an anchor patch feature t_n from $\mathcal{V}_i$ with all the patch features from $\mathcal{V}_i'$.

$$sr^n_{\mathcal{V}_i} = \frac{\max(\mathcal{D}_n) - \frac{1}{N}\sum_{m=1}^{N} \mathcal{D}_n}{\sigma_{\mathcal{D}_n}}, \tag{4}$$

where, $\sigma_{\mathcal{D}_n}$ is the variance of vector $\mathcal{D}_n$ and the notation $\max(\mathcal{D}_n)$ is the maximal value of $\mathcal{D}_n$.

$$l_{n,m} = \frac{(softmax(sr_{\mathcal{V}_i})_n + softmax(sr_{\mathcal{V}_i'})_m)\mathcal{L}_{p2p}(n,m)}{2}. \tag{5}$$

Finally, the patch-to-structure loss $\mathcal{L}_{p2s}$ is

$$\mathcal{L}_{p2s} = \frac{\sum_{m=0}^{N}\sum_{n=0}^{N} l_{n,m}}{N \times N}. \tag{6}$$

Experiments and Analysis

Overall comparisons on 10 downstream datasets. The results demonstrate that S^2DC performs effectively across 10 datasets and 4 tasks. Compared to training from scratch, which achieves an average score of 77.93%, S^2DC pre-training delivers a **3.5% improvement**, reaching 81.43%. Additionally, S^2DC outperforms the second-best SSL method (VoCo, average score 80.65%) for all tasks, with an average gain of 0.78%.

Task	Dataset	Modality
Segmentation	BTCV [19]	CT
	MSD-Liver [29]	CT
	MSD-Lung [29]	CT
	MSD-Spleen [29]	CT
	BraTs 21 [29]	MRI
	AUTOPET [10]	PET
Classification	CC-CCII [46]	CT
	ADNI-cls [22]	PET
Reconstruction	UDPET [4]	PET
I2I translation	BraTs 23 [29]	MRI

Figure 4. The downstream tasks and modalities.

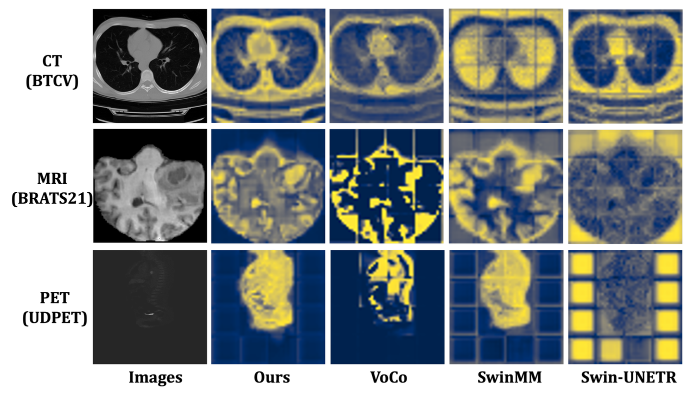

Figure 5. The visualization of the first principal components after applying PCA to token features.

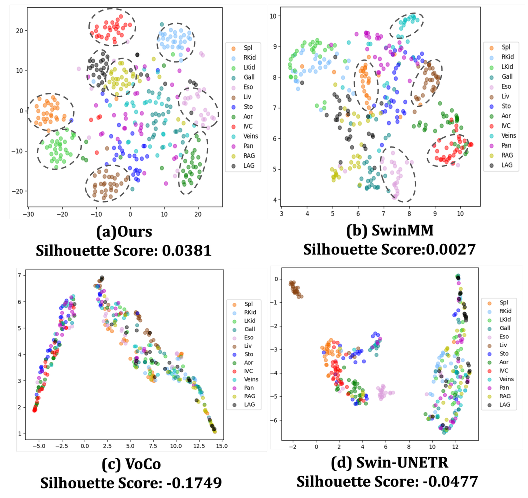

Figure 6. The t-SNE feature visualization of different losses on 13 organs on the BTCV dataset.

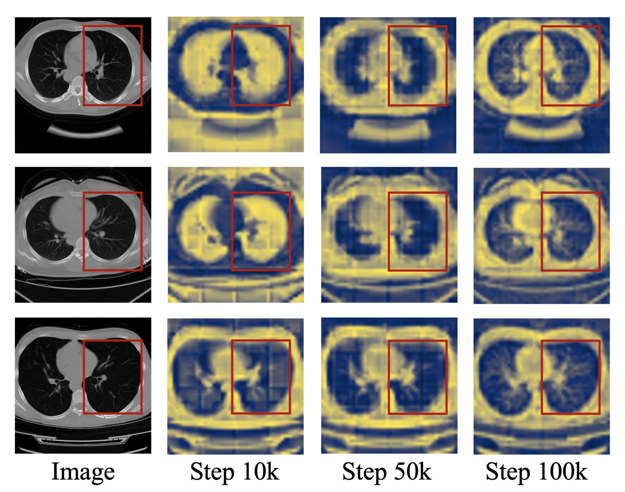

Figure 7. The evolution of patch-to-structure correspondences.