Keep Your Friends Close, and Your Enemies Farther: Distance-aware Voxel-wise Contrastive Learning for Semi-supervised Multi-organ Segmentation

Haochen Zhao[1], Jianwei Niu[1], Xuefeng Liu[1]*, Xiaozheng Xie[2], Li Kuang[1], Haotian Yang[1], Bin Dai[3], Hui Meng[4], Yong Wang[5]*
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]School of Computer and Communication Engineering, University of Science and Technology Beijing
[3]Hangzhou International Innovation Institute of Beihang University
[4]Hangzhou Institute for Advanced Study, University of the Chinese Academy of Sciences
[5]Chinese Academy of Medical Sciences and Peking Union Medical College
{studyzhao, niujianwei, liu_xuefeng}@buaa.edu.cn

Abstract

Based on pseudo-labels, voxel-wise contrastive learning (VCL) is a prominent approach designed to learn effective feature representations for semi-supervised medical image segmentation. However, in multi-organ segmentation (MoS), the complex anatomical structures of certain organs often lead to many unreliable pseudo-labels. Directly applying VCL can introduce confirmation bias, resulting in poor segmentation performance. A common practice is to first transform these unreliable pseudo-labels into complementary ones, which represent classes that voxels are least likely to belong to, and then push voxels away from the generated complementary labels. However, we find that this approach may fail to allow voxels with unreliable pseudo-labels (unreliable voxels) to fully benefit from the advantages of VCL. In this paper, we propose DVCL, a novel distance-aware VCL method for semi-supervised MoS. DVCL is based on the observation that unreliable voxels, which may not form discriminative feature boundaries, still form clear clusters. Hence, voxels close to each other in the feature space ('neighbors') likely belong to the same semantic class, while distant ones ('outsiders') likely belong to different classes. In DVCL, we first identify neighbors and outsiders for all unreliable voxels, and then pull their neighbors into the same clusters while pushing outsiders away. In this way, unreliable voxels can learn more discriminative features, thereby fully enjoying the advantages of VCL. However, DVCL itself will inevitably introduce the problem of noisy neighbors and outliers. To address these challenges, we further propose a neighbor partitioning strategy and a query outlier strategy to provide more stable feature representations for DVCL. Extensive experiments demonstrate the effectiveness of our method.

*Corresponding author.

1. Introduction

Medical image segmentation [31, 52] is a critical task in computer-aided diagnosis. However, considering the large amount of effort required for labeling data, semi-supervised learning (SSL) is often employed. With the continuous advancement of feature extraction techniques [7, 15, 17, 22], voxel-wise contrastive learning (VCL) [3, 46, 49] has proven highly effective in SSL. Based on the initially generated pseudo-labels, VCL pulls voxels with the same pseudo-labels together in feature space while pushing those with different pseudo-labels away from each other, thereby learning effective representations from unlabeled voxels for the segmentation task.

However, existing VCL methods can encounter difficulty in multi-organ segmentation (MoS). MoS is more challenging than single-organ segmentation due to complex anatomical structures, including large size variations, different shapes, and overlapping organs [39, 53, 54]. Consequently, segmenting these organs (*e.g.*, the left and right adrenal glands) becomes considerably more difficult [4, 39], resulting in many unreliable pseudo-labels, especially at the initial stage. These unreliable pseudo-labels generally lead to confirmation bias [2, 23] in VCL, ultimately lowering the segmentation performance on these organs in MoS.

To address these challenges in VCL, some methods [3, 35, 41] choose not to use voxels with unreliable pseudo-labels (unreliable voxels). However, given the amount of unreliable voxels, this approach wastes the valuable information contained in them. To leverage information contained in a large number of unreliable pseudo-labels, a popular practice [11, 12, 14, 44] is to push the voxel away from

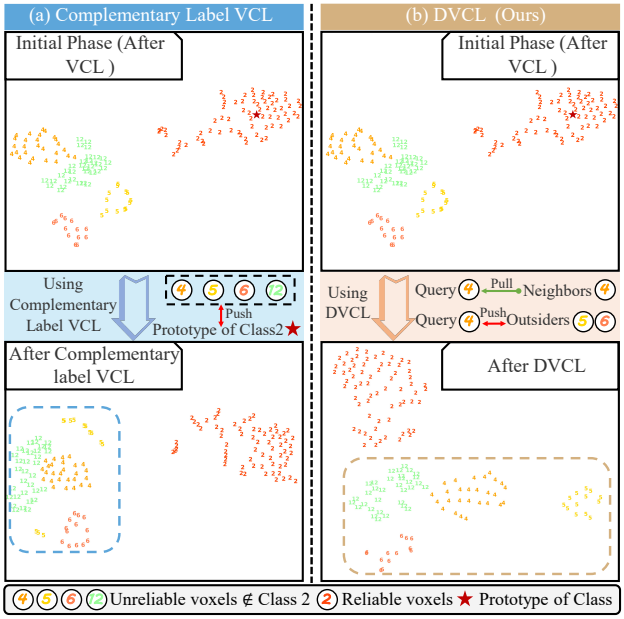

Figure 1. Illustration of the difference between (a) previous method and (b) our method in making full use of unreliable voxels within contrastive learning (CL). t-SNE [38] visualization of feature spaces on FLARE 2022 [29] val image. For better visualization, we show five classes, with the numbers indicate the true classes of the features. As seen, our method encourages the learning of more discriminative decision boundaries among these classes, allowing unreliable voxels to fully enjoy the advantages of CL, as highlighted in the dashed boxes.

the prototypes of its complementary labels. These generated complementary labels represent classes the voxel is least likely to belong to, and therefore can be more reliable than the original pseudo-label. Taking Fig.1(a) as an example, the unreliable voxels (*voxels 4, 5, 6, and 12*), which are least likely to belong to *class 2*, are further pushed away from the prototype of *class 2* based on VCL.

Despite this technique successfully leverages information from unreliable pseudo-labels, it still has limitations. An example is shown in the upper figure of Fig.1(a). After being pushed away from the prototype of *class 2*, *voxels 4* and *5*, as well as *voxels 4* and *6*, which are originally well-distinguished after VCL, become closer to each other. Intuitively, classes that should not be confused are incorrectly brought closer, potentially leading to interference between them. Furthermore, some voxels (*e.g., voxels 5*) may deviate from their original clustering structure. Consequently, this approach focuses solely on pushing unreliable voxels away from the generated complementary labels, which may prevent them from fully enjoying the advantages of VCL, such as inter-class separation and intra-class aggregation.

In this paper, we propose a novel distance-aware VCL (DVCL) technique to address the above challenges. First, we observe that after VCL, although distinct discriminative feature boundaries have not been formed between classes of unreliable voxels, those of the same class are still expected to form a cluster in feature space, with some neighborhood relationships remaining evident. For example, *voxels 4* and *12* may be confused, whereas *voxels 4* and *5*, as well as *voxels 4* and *6*, are relatively easy to distinguish (see the top figure of Fig.1(b)). This phenomenon also highlights the core advantage of VCL, which lies in promoting the formation of distinct clustering structures [1, 42]. Based on this observation, the rationale behind DVCL is simple: for an unreliable voxel, its close unreliable voxels in feature space (referred to as neighbors), which likely belong to the same semantic class, while its far-away unreliable voxels (referred to as outsiders), which likely belong to different classes. Specifically, in DVCL, we first identify, for each unreliable voxel, its neighbors and outsiders. During contrastive learning, for all unreliable voxels, we pull their neighbors into the same clusters while pushing outsiders away. An example of using DVCL is shown in Fig.1(b). Take *voxel 4* as an example (Query). Before implementing DVCL, some instances of *voxels 4* are its neighbors, while some instances of *voxels 5* and *6* are its outsiders. Then, neighbors are pulled into the same cluster as the query, while outsiders are pushed away. In this way, we encourage unreliable voxels to learn more discriminative features, allowing them to fully enjoy the advantages of VCL. This further facilitates the generation of higher quality pseudo-labels.

Though our proposed DVCL can lead to good performance, we find two main obstacles that may negatively impact its effectiveness. (1) Noisy Neighbor Dilemma: Voxels from different classes may ultimately become neighbors due to their close proximity to each other in the feature space. (2) Outlier Query Dilemma: When an outlier serves as a query, it may degrade feature representations, as outliers typically lack semantically similar neighbors. To reduce the potential impact of noisy neighbors and outliers, we first adopt the neighbor partitioning strategy (NPS), which classifies neighbors into two categories: relevant neighbors and ordinary neighbors. During DVCL, relevant neighbors are assigned higher weights than ordinary neighbors to reflect their greater contribution to neighbor clustering. Second, we further propose a query detection s trategy (QDS) to exclude potential outliers when selecting queries in DVCL, leading to further performance gains. We apply DVCL on four datasets under different partition protocols and demonstrate its superior performance. Overall, our contributions can be summarized as follows:

- We propose a novel distance-aware voxel-wise contrastive learning (DVCL) technique to encourage unreliable voxels to learn more discriminative features, allowing them to fully enjoy the advantages of VCL.
- We introduce the neighbor partitioning strategy (NPS)

and the outlier detection strategy (ODS) to enhance the stability of feature representations in DVCL.
- Extensive experiments are performed on four public datasets, resulting in new state-of-the-art results on different scenarios.

2. Related Work

Semi-supervised Multi-organ Segmentation. Due to the large variations of appearance and size of different organs, semi-supervised MoS has been a popular yet challenging task [30, 34, 40, 48, 55]. To address the class-imbalance problem, [24] proposed CLD to leverage label distribution to encourage the network to put more effort into small organ. Similarly, [39] proposed DHC, which dynamically utilizes pseudo-labels to guide the model in mitigating data and learning biases. To enhance the quality of pseudo-labels with the help of labeled data, [4] designed MagicNet, which helps unlabeled images learn semantic information about organ positions from labeled images. Additionally, [53] proposed GuidedNet, which refines pseudo-labels by leveraging feature distributions from labeled data. Recently, several methods have sought to enhance voxel classification accuracy by constructing a well-structured embedding space through VCL. [32, 36, 45] utilize VCL to boost the representation capacity under the guidance of pseudo-labels, thus exploring unlabeled data more efficiently. Considering that unreliable pseudo-labels may not be accurate, [25] designed UGCL to select reliable parts from the unreliable pseudo-labels using predicted probabilities for further utilization. In this work, we fully utilize unreliable pseudo-labels and encourage voxels with unreliable pseudo-labels to learn more discriminative features, thereby enhancing segmentation performance for challenging organs in MoS.

Semi-supervised Contrastive Learning. Recently, VCL has achieved significant success in extracting powerful features from unlabeled data in SSL [11, 14, 41, 42]. For example, [42] proposed a VCL algorithm for SSL semantic segmentation that enhances similarity between embeddings of same-class voxels while distinguishing different-class voxels using pseudo-labels. However, unreliable pseudo-labels can introduce confirmation bias in VCL, ultimately degrading task performance. To address this, [41] proposed UGPCL that discard unreliable pseudo-labels based on threshold. Additionally, [18] first introduced a reliable memory bank to optimize memory usage during training and improve performance through memory management. To effectively leverage the valuable supervisory signals within unreliable pseudo-labels, works such as [11, 12, 14, 44] have focused on treating unreliable voxels as negative samples for the most unlikely classes (complementary labels). However, this approach fails to fully enable unreliable voxels to benefit from VCL. In contrast, our method encourages unreliable voxels to learn more discriminative features, allowing them to fully enjoy the advantages of VCL. This capability is a key reason for the superiority of our method.

3. Methodology

3.1. Problem Definition and Framework Overview

Problem Definition. The training dataset, denoted as $\mathcal{S}$, consists of both labeled and unlabeled data, expressed as $\mathcal{S} = \mathcal{S}^l \cup \mathcal{S}^u$. Here, $\mathcal{S}^l = \{(x_i, y_i)\}_{i=1}^N$ constitutes the labeled data, and $\mathcal{S}^u = \{x_i\}_{i=1}^M$ represents the unlabeled data ($M \gg N$ in most cases). In this context, $x_i \in \mathbb{R}^{D\times H\times W}$ refers to the input volume, and $y_i \in \mathbb{R}^{C\times D\times H\times W}$ denotes the ground-truth mask, where C is the number of classes, including the background. H,W, and D represent the height, width, and depth of the input medical volume, respectively. The goal is to train a segmentation network based on $\mathcal{S}^l$ and $\mathcal{S}^u$ that correctly predicts labels for unseen data.

Framework Overview. An overview of the proposed framework is illustrated in Fig.2, which is based on the Cross Pseudo Supervision (CPS) framework [6] and utilizes two parallel segmentation networks of the same architecture without sharing parameters: *model A* and *model B*. Each network consists of four main components: an encoder, a decoder, a segmentation head, and a feature head (*e.g.,* the sequentially-connected MLP layers). At each training step, $\mathcal{B}_l$ labeled data and $\mathcal{B}_u$ unlabeled data are sampled and fed into the *model A* and the *model B*, respectively. For the labeled data, the supervised loss function is applied, guiding each segmentation head to generate a prediction mask that closely aligns with the ground-truth mask:

$$\mathcal{L}_{sup} = \frac{1}{\mathcal{B}_l}\sum_{i=1}^{\mathcal{B}_l}\left[\mathcal{L}_s(p_i^A, y_i) + \mathcal{L}_s(p_i^B, y_i)\right], \quad (1)$$

where $\mathcal{L}_s = \frac{1}{2}[\mathcal{L}_{Dice} + \mathcal{L}_{ce}]$, $\mathcal{L}_{Dice}$ and $\mathcal{L}_{ce}$ represent the Dice and cross-entropy losses, respectively. $p_i^{A\text{ or }B}$ is the predict probabilities, and $p_i^{A\text{ or }B}(c)$ is the value of $p_i^{A\text{ or }B}$ for the c-th dimension. For unlabeled data, each segmentation network generates a pseudo-label, which serves as an additional supervisory signal for the other network, ensuring that both produce consistent outputs:

$$\mathcal{L}_{cps} = \frac{1}{\mathcal{B}_u}\sum_{i=1}^{\mathcal{B}_u}\left[\mathcal{L}_{ce}(p_i^A, \hat{y}_i^B) + \mathcal{L}_{ce}(p_i^B, \hat{y}_i^A)\right], \quad (2)$$

where $\hat{y}_i^{A\text{ or }B} = \arg\max_c p_i^{A\text{ or }B}(c)$ is the pseudo-labels. Specifically, an entropy-based selection module (ESM) adaptively divides pseudo-labels into reliable and unreliable groups, as detailed in Section 3.2. The final objective for adaptation is:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u\mathcal{L}_{cps} + \lambda_c(\mathcal{L}_{hqcl} + \beta\mathcal{L}_{dvcl}), \quad (3)$$

where $\mathcal{L}_{hqcl}$ denotes the high-quality voxel-wise contrastive loss for voxels with reliable pseudo-labels and ground-truth

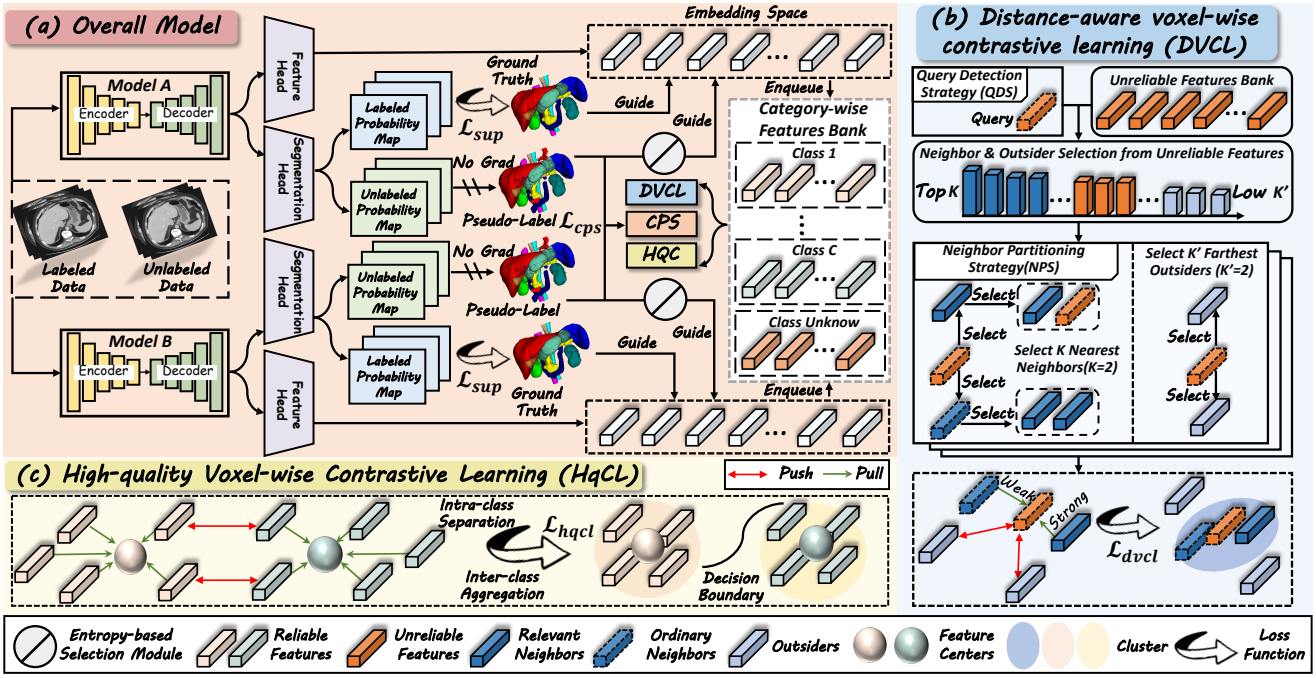

Figure 2. Overall framework. Labeled data is directly fed into both networks for supervised training. For unlabeled data, we separate the initially generated pseudo-labels into reliable and unreliable ones based on ESM. Reliable features (features with reliable pseudo-labels and ground-truth masks) are used to compute the $\mathcal{L}_{hqcl}$ (yellow block). Unreliable features (features with unreliable pseudo-labels) are used to compute the $\mathcal{L}_{dvcl}$, with NPS and QDS incorporated during the computation of $\mathcal{L}_{dvcl}$ (blue block).

masks (Section 3.3), while $\mathcal{L}_{dvcl}$ represents the distance-aware voxel-wise contrastive loss for voxels with unreliable pseudo-labels (Section 3.4). λ_u is set to 0.1 and used the CPS weight ramp-up function [50]. The value of λ_c is set to 0.1. The scalar $\beta = 0.3$ is used to balance the two contrastive loss functions.

3.2. Entropy-based Selection Module

We follow [44] utilize voxel-wise entropy $e_{ij}^{A\text{ or }B}$ as the selection indicator between reliable and unreliable pseudo-labels in unlabeled data as:

$$e_{ij}^{A\text{ or }B} = \sum_{c=1}^{C} -p_{ij}^{A\text{ or }B}(c) \log\left(p_{ij}^{A\text{ or }B}(c)\right), \tag{4}$$

where $p_{ij}^{A\text{ or }B}(c)$ is the value of predict probabilities $p_{ij}^{A\text{ or }B}$ at c-th dimension of the j-th voxel of the i-th data. We define $M_r^{A\text{ or }B}[i,j]$ and $M_u^{A\text{ or }B}[i,j]$ as the masks of low and high entropy for selecting reliable and unreliable pseudo-labels respectively. These masks can be expressed as:

$$\begin{cases} M_r^{A\text{ or }B}[i,j] = \mathbb{I}(e_{ij}^{A\text{ or }B} < \tau_e^{A\text{ or }B}), \\ M_u^{A\text{ or }B}[i,j] = \mathbb{I}(e_{ij}^{A\text{ or }B} > \tau_e^{A\text{ or }B}), \end{cases} \tag{5}$$

where $\tau_e^{A\text{ or }B} = e_{at}^{A\text{ or }B} + \alpha e_{st}^{A\text{ or }B}$ is the selection threshold of the entropy mask. α is factor and is set to 0.5 in our experiments. $e_{at}^{A\text{ or }B}$ represents the average entropy of voxels at training iteration t, while $e_{st}^{A\text{ or }B}$ denotes its variance:

$$\begin{cases} e_{at}^{A\text{ or }B} = \frac{\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{D\times H\times W}}{\mathcal{B}*D\times H\times W} e_{ij}^{A\text{ or }B}, \\ e_{st}^{A\text{ or }B} = \sqrt{\frac{\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{D\times H\times W}}{\mathcal{B}*D\times H\times W} \left(e_{ij}^{A\text{ or }B} - e_{at}^{A\text{ or }B}\right)^2}. \end{cases} \tag{6}$$

This new selection strategy has the following benefits: *i)* ESM allows the model to adapt its confidence across different training stages, thereby enabling a clearer distinction between reliable and unreliable pseudo-labels. *ii)* ESM compares confidence against a threshold, τ_e, without needing to sort all training voxels [28, 43, 44]. This reduces the time complexity from $O(N \log N)$ to $O(N)$, making the strategy more efficient when applied to large-scale datasets.

3.3. High-quality Contrastive Learning

Anchor Voxels Sampling. For labeled data, we select high-quality anchor voxels with accurate predictions based on ground-truth masks, while for unlabeled data, we rely on reliable pseudo-labels (Fig.3 yellow block). The feature sets of the labeled and unlabeled anchor voxels for class c as

$$\begin{cases} \mathcal{R}_c^{Al\text{ or }Bl} = \{\mathbf{r}_{ij}^{A\text{ or }B} \mid y_{ij} = c, \arg\max p_{ij}^{A\text{ or }B}(c) = c\}, \\ \mathcal{R}_c^{Au\text{ or }Bu} = \{\mathbf{r}_{ij}^{A\text{ or }B} \mid \hat{y}_{ij}^{A\text{ or }B} = c, [i,j] \in M_r^{A\text{ or }B}[i,j]\}, \end{cases} \tag{7}$$

where $\mathbf{r}_{ij}^{A\text{ or }B} \in \mathbb{R}^d$ represents the feature of the j-th voxel of i-th labeled image given by feature head. y_{ij} denotes

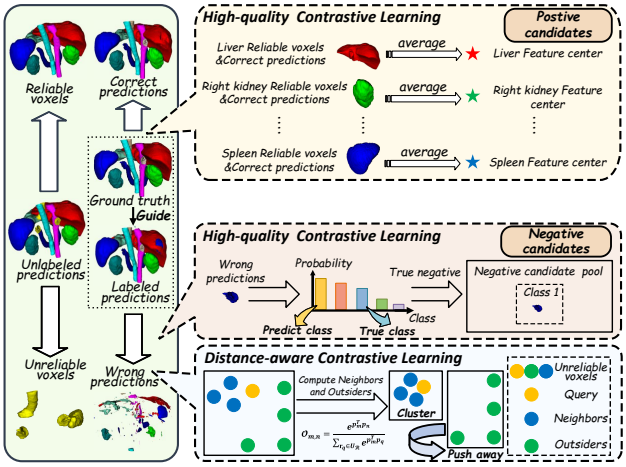

Figure 3. An illustration of positive and negative candidate selection in $\mathcal{L}_{hqcl}$ and neighbor and outsider computation in $\mathcal{L}_{dvcl}$.

the ground-truth mask. $M_r^{A\text{ or }B}[i,j]$ denotes the reliable pseudo-labels masks in Eq.6. The set of all qualified anchor voxels for class c is $\mathcal{R}_c = \mathcal{R}_c^{Al} \bigcup \mathcal{R}_c^{Bl} \bigcup \mathcal{R}_c^{Au} \bigcup \mathcal{R}_c^{Bu}$. To prevent limited anchors leading to an overly localized and unstable sample center, we store $\mathcal{R}_c$ by using a memory bank. When the memory bank is saturated, we remove old features to leave enough space to store the latest features.

Positive Candidate Sampling. The positive candidate for class c is set as the center of all anchor voxels (Fig.3 yellow block):

$$\mathcal{P}_c = \frac{1}{|\mathcal{R}_c|} \sum_{\mathbf{r}_c \in \mathcal{R}_c} \mathbf{r}_c. \tag{8}$$

Negative Candidate Sampling. We consider labeled voxels that are misclassified as class c as true negative candidates for that class, helping to guide these voxels away from misclassified classes in the feature space (Fig.3 pink block). Consequently, the feature sets of negative candidates for class c are defined as:

$$\mathcal{G}_c^{Al\text{ or }Bl} = \{\mathbf{r}_{ij}^{A\text{ or }B} \mid \arg\max p_{ij}^{A\text{ or }B}(c) = c, y_{ij} \neq c\}. \tag{9}$$

Finally, the set of negative candidates of class c is $\mathcal{G}_c = \mathcal{G}_c^{Al} \bigcup \mathcal{G}_c^{Bl}$. With anchors, positive and negative candidates available, we adopt the InfoNCE [33] loss for class c as

$$\mathcal{L}_{hqcl}^c = -\frac{1}{|M|} \sum_{m=1}^{|M|} \log \frac{\Phi_\sigma(\mathbf{r}_{cm}, \mathcal{P}_c)}{\Phi_\sigma(\mathbf{r}_{cm}, \mathcal{P}_c) + \sum_{n=1}^{N} \Phi_\sigma(\mathbf{r}_{cm}, \mathbf{r}_{cmn}^-)}, \tag{10}$$

where M=256 is the number of anchor voxels, and $\mathbf{r}_{cm}$ is the feature of the m-*th* anchor of class c. Each anchor is assigned one positive candidate $\mathcal{P}_c$ and N=50 negative candidates, $\{\mathbf{r}_{cmn}^-\}_{n=1}^N$. $\Phi_\sigma(a,b) = e^{\cos(a,b)/\sigma}$ with a temperature σ set to 1 in this work. Finally, the entire HqCL loss for all classes are obtained with $\mathcal{L}_{hqcl} = \frac{1}{C}\sum_{c=1}^{C} \mathcal{L}_{hqcl}^c$.

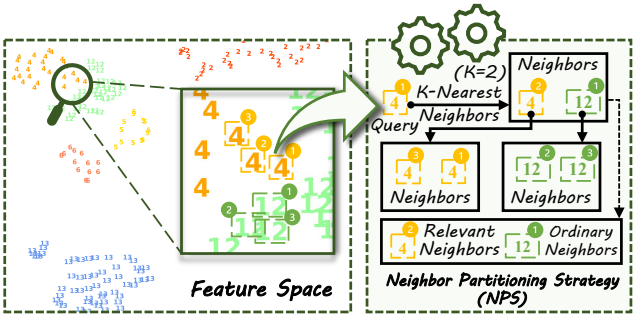

Figure 4. Left: Feature space visualizations—a real example using t-SNE [38] on the FLARE 2022 val image. The numbers represent the true classes of the features. Right: Illustration of the detailed processing procedure of NPS.

3.4. Distance-aware Contrastive Learning

From the perspective that features that are close or distant in the feature space should yield consistent or inconsistent predictions, we propose a DVCL technique to encourage unreliable voxels to learn more discriminative features.

Inspired by [16, 47], we define $\mathcal{O}_{m,n}$ as the probability that the feature $\mathbf{r}_m$ has same prediction to feature $\mathbf{r}_n$:

$$\mathcal{O}_{m,n} = \frac{e^{p_m^T p_n}}{\sum_{\mathbf{r}_q \in U_\mathcal{R}} e^{p_m^T p_q}}, \tag{11}$$

where $U_\mathcal{R} = M_u^A \bigcup M_u^B$ denotes the feature set of all unreliable voxels by following Eq.6. $\mathbf{r}_m$ and $\mathbf{r}_n$ denote the features of two voxels from $U_\mathcal{R}$, while having corresponding predict probabilities p_m and p_n.

In the current batch, we randomly select a subset of features from $U_\mathcal{R}$ as queries. We then define two sets for each query $\mathbf{r}_m$: close neighbor set $\mathcal{C}_m^K$ (features potentially from same classes) and distant outsider set $\mathcal{D}_m^{K'}$ (features potentially from different classes). $\mathcal{C}_m^K$ is formed by selecting the K-Nearest Neighbors (KNN) [10] of $\mathbf{r}_m$ using cosine similarity as the distance metric. Conversely, $\mathcal{D}_m^{K'}$ is constructed using the K'- Farthest Neighbors of $\mathbf{r}_m$:

$$\begin{cases} \mathcal{C}_m^K = \{\mathbf{r}_n | top\text{-}K(\cos(\mathbf{r}_m, \mathbf{r}_n), \forall \mathbf{r}_n \in U_\mathcal{R})\}, \\ \mathcal{D}_m^{K'} = \{\mathbf{r}_k | low\text{-}K'(\cos(\mathbf{r}_m, \mathbf{r}_k), \forall \mathbf{r}_k \in U_\mathcal{R})\}. \end{cases} \tag{12}$$

Following [13], we use queues as candidate sets $\mathcal{C}_m^K$ and $\mathcal{D}_m^{K'}$, respectively, where each element represents a feature. During network training with sample batches, we store features in $\mathcal{C}_m^K$ and $\mathcal{D}_m^{K'}$, updating the candidate sets using a first-in-first-out strategy.

Neighbor Partitioning Strategy. However, features from different classes may ultimately become neighbors due to their close proximity to each other in the feature space, especially in regions where organs overlap. Taking Fig.4 as an example, when *voxel* 4^1 serves as the query, it may

mistakenly treat *voxel 12*[1] ('noisy neighbor') as a neighbor, thereby providing incorrect supervision. To avoid the Noisy Neighbor Dilemma, we apply neighbor partitioning strategy (NPS) to divide the neighbors into two groups: relevant neighbors and ordinary neighbors (See Fig.4 right). The feature $\mathbf{r}_n$ is regarded as the relevant neighbor of the query $\mathbf{r}_m$ if it meets the following condition:

$$n \in \mathcal{C}_m^K \wedge m \in \mathcal{C}_n^K. \tag{13}$$

Other neighbors which do not meet the above condition are ordinary neighbors. Considering that relevant neighbors have a higher potential to belong to the same cluster as the query. Thus, we assign a high affinity value to the relevant neighbors. Specifically for query $\mathbf{r}_m$, the affinity value of its n-th neighbor is defined as

$$\mathcal{A}_{m,n} = \begin{cases} 1 & \text{if } n \in \mathcal{C}_m^K \wedge m \in \mathcal{C}_n^K, \\ r & \text{otherwise,} \end{cases} \tag{14}$$

where r is a hyperparameter set to 0.5. It connects the query to relevant neighbors with strong connectivity while considering a weaker connection to ordinary neighbors.

Query Detection Strategy. Outliers are typically not retrieved as neighbors due to their large distance from the majority of voxels in the feature space. However, when an outlier exists as a query, it is often unsure whether its neighbors belong to the same semantic class. This is why the Outlier Query Dilemma may deteriorate the feature representation. To detect the outlier, we define the probability of each feature $\mathbf{r}_m$ being an outlier by measuring the number of samples that consider it as their neighbor. where

$$\mathcal{E}(m) = \{n | n \in \mathcal{C}_m^K \wedge m \in \mathcal{C}_n^U\}. \tag{15}$$

U is significantly larger than K. The more samples in $\mathcal{E}(m)$, the lower the probability that $\mathbf{r}_m$ is an outlier. We consider $\mathbf{r}_m$ is an outlier if $|\mathcal{E}(m)| < 1$. when the feature $\mathbf{r}_m$ is an outlier, which means $\mathcal{E}(m)$ is the empty set, it will be excluded from the objective computation as a query.

Distance-aware Voxel-wise Contrastive Learning. Returning to our motivation, for each query (exclude outliers) $\mathbf{r}_m$, the features in $\mathcal{D}_m^{K'}$ should have different predictions than those in $\mathcal{C}_m^K$. This process involves moving features in $\mathcal{C}_m^K$ towards $\mathbf{r}_m$ while pushing away features in $\mathcal{D}_m^{K'}$. Specifically, treating all neighbors equally is not reasonable. To achieve this, we first define two likelihood functions:

$$\begin{cases} \mathcal{N}\left(\mathcal{C}_m^K | \theta_{\mathcal{S}}\right) = \prod_{\mathbf{r}_n \in \mathcal{C}_m^K} \frac{e^{\mathcal{A}_{m,n} p_m^T p_n}}{\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}}, \\ \mathcal{N}\left(\mathcal{D}_m^{K'} | \theta_{\mathcal{S}}\right) = \prod_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} \frac{e^{p_m^T p_k}}{\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}}, \end{cases} \tag{16}$$

where $\theta_{\mathcal{S}}$ denotes the parameters of the segmentation head in our network. We then propose to simultaneously maximize the log-likelihood of the close neighbor set and minimize the log-likelihood of the distant outsider set, denoted as

$$\psi(\mathcal{C}_m^K, \mathcal{D}_m^{K'}) = -\log \frac{\mathcal{N}\left(\mathcal{C}_m^K \mid \mathbf{r}_m, \theta_{\mathcal{S}}\right)}{\mathcal{N}\left(\mathcal{D}_m^{K'} \mid \mathbf{r}_m, \theta_{\mathcal{S}}\right)}. \tag{17}$$

One problem optimizing Eq.17 is that it requires the participation of all features in $U_{\mathcal{R}}$ to compute $e^{p_m^T p_q}$ in Eq.16, leading to potential inefficiencies in time and resources in practice. Here we resort to get an upper-bound of Eq.17:

$$\begin{aligned} \psi(\mathcal{C}_m^K, \mathcal{D}_m^{K'}) =& -\log \frac{\mathcal{N}\left(\mathcal{C}_m^K \mid \theta_{\mathcal{S}}\right)}{\mathcal{N}\left(\mathcal{D}_m^{K'} \mid \theta_{\mathcal{S}}\right)} \\ =& -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{A}_{m,n} p_m^T p_n + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} p_m^T p_k + (K - K') \log\Big(\sum_{\mathbf{r}_q \in U_{\mathcal{R}}} e^{p_m^T p_q}\Big) \\ \leq& -\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{A}_{m,n} p_m^T p_n + \sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} p_m^T p_k + (K - K')\Big(\log |U_{\mathcal{R}}| + \sum_{\mathbf{r}_q \in U_{\mathcal{R}}} \frac{p_m^T p_q}{|U_{\mathcal{R}}|}\Big) \\ <& \underbrace{-\sum_{\mathbf{r}_n \in \mathcal{C}_m^K} \mathcal{A}_{m,n} p_m^T p_n}_{\text{Nerghbors Cluster}} \underbrace{+\sum_{\mathbf{r}_k \in \mathcal{D}_m^{K'}} p_m^T p_k}_{\text{Outsiders Separate}} + (K - K')(\log |U_{\mathcal{R}}| + 1) \\ =& \overline{\psi}(\mathcal{C}_m^K, \mathcal{D}_m^{K'}). \end{aligned} \tag{18}$$

Details for our proof can be found in Sec.G of Supplementary Material (SM). Finally, our DVCL loss is defined as follows:

$$\mathcal{L}_{dvcl} = \frac{1}{|U_{\mathcal{R}}|} \sum_{\mathbf{r}_m \in U_{\mathcal{R}}} \overline{\psi}(\mathcal{C}_m^K, \mathcal{D}_m^{K'}). \tag{19}$$

4. Experiment

4.1. Dataset and Implementation Details

We evaluate our method using four widely recognized MoS datasets, including FLARE 2022 [29], AMOS [19], MMWHS [56], and BTCV [20]. See Sec.A of SM for more dataset details. For both datasets, the data are preprocessed before the network is trained. For more implementation details, *e.g.,* data preprocessing, network parameters, metrics, learning rate policy, batch sizes, *etc.,* please refer to the Sec.B of SM. We implement the proposed framework with PyTorch, using 1 NVIDIA A100 GPUs. Model weights are determined based on performance on the validation set, while comparisons of different methods are made using segmentation metrics on the test set.

4.2. Comparison to SOTA Methods

For experiments, we compare our method to thirteen state-of-the-art semi-supervised segmentation methods: (1) DAN [51], (2) MT [37], (3) UA-MT [50], (4) SASSnet [21], (5) DTC [26], (6) CPS [6], (7) CLD [24], (8) DHC [39], (9) MagicNet [4], (10) UGPCL [41], (11) U^{2}PL [44], (12) BaCon [14], (13) CCL [11], and the fully supervised 3D U-Net [9]. Note that among all the above evaluated methods, several methods use a contrastive learning objective, including UGPCL, U^{2}PL, BaCon, and CCL. For all semi-supervised methods, we utilize 3D U-Net as the backbone.

Method	FLARE 2022		AMOS		MMWHS		BTCV	
	ratio=0.1	ratio=0.5	ratio=0.1	ratio=0.5	ratio=0.1	ratio=0.5	ratio=0.1	ratio=0.5
DAN [51]	75.1±0.6	74.9±0.6	57.8±0.8	61.4±1.1	74.3±0.8	83.9±0.8	37.7±1.4	57.7±1.2
MT [37]	77.5±0.4	77.2±0.4	45.1±1.9	66.2±0.7	81.2±0.5	86.1±0.2	36.6±0.6	60.2±0.9
UA-MT [50]	79.1±0.3	79.0±0.8	61.7±1.1	65.5±0.8	82.5±0.8	86.4±0.1	44.9±1.1	58.0±0.8
SASSnet [21]	76.8±0.3	76.4±0.6	58.4±1.4	63.8±1.1	81.3±0.4	83.0±0.4	46.2±1.0	61.9±0.1
DTC [26]	78.5±0.8	78.8±0.7	60.8±1.2	66.9±1.7	82.7±0.5	85.8±0.2	47.8±1.4	60.8±0.9
CPS [6]	79.3±0.4	79.2±0.2	63.5±0.3	66.6±1.2	83.6±0.2	86.5±0.1	47.7±0.7	61.0±0.4
CLD [24]	79.7±0.1	79.5±0.2	65.8±1.2	69.1±1.1	84.2±0.8	86.9±0.3	49.0±0.9	61.4±0.8
DHC [39]	80.5±0.4	80.2±1.0	65.2±1.4	68.6±0.5	83.7±0.8	86.9±0.1	49.0±0.3	61.2±1.5
MagicNet [4]	80.8±0.7	80.2±0.3	65.3±1.3	69.0±0.5	79.4±0.6	83.9±0.6	50.2±0.5	61.6±0.3
UGPCL [41]	79.6±0.7	80.6±0.8	61.5±1.0	70.7±0.1	84.0±0.3	86.4±0.4	46.6±1.2	59.0±0.8
U^{2}PL [44]	82.1±0.3	82.0±0.3	64.7±1.3	70.0±0.2	84.3±0.1	86.7±0.1	48.4±0.6	61.3±0.2
BaCon [14]	82.0±0.3	81.6±0.5	64.4±1.5	70.1±0.3	84.2±0.1	86.3±0.1	47.9±1.5	61.1±0.1
CCL [11]	81.8±0.1	81.3±0.3	64.0±0.1	68.8±1.6	83.8±0.3	85.6±0.6	48.0±1.4	60.4±0.8
Our-Method	**84.2±0.1**	**83.7±0.3**	**71.4±0.5**	**72.5±0.7**	**86.3±0.2**	**88.2±0.1**	**52.9±0.2**	**65.3±0.7**

Table 1. Comparison with state-of-the-art methods on four datasets under different partition protocols. **Best results are boldfaced.** To reduce the randomness of network training, experiments are calculated in triplicate for all methods.

We demonstrate that our method achieves superior performance across all datasets and label ratios (10%, 50%). As shown in Table.1, our methods consistently outperform all the compared SSL methods by a considerable margin across all datasets and label ratios. Compared to the second-best results, our method under {10%, 50%} label ratios achieves {2.06%↑, 1.70%↑}, {5.55%↑, 1.74%↑}, {2.08%↑, 1.29%↑}, {2.76%↑, 3.39%↑} in mean Dice across FLARE 2022, AMOS, MMWHS, and BTCV, respectively. Detailed results for each organ in four datasets can be found in Sec.D of SM. We provide qualitative illustrations of FLARE 2022, AMOS, MMWHS, and BTCV in Fig.1 (Sec.E of SM), Fig.2 (Sec.E of SM), Fig.3 (Sec.E of SM), Fig.4 (Sec.E of SM), respectively. These results indicate that our method is well segmented relative to other methods.

Furthermore, Fig.5 illustrates the Dice curves of four difficult organs (Sto, Pan, RAG, and LAG) for different methods on the validation set during network training, where a significant number of unreliable pseudo-labels are prevalent. Firstly, our method demonstrates the ability to learn organ knowledge earlier than other methods. For example, as shown by the LAG curve, our approach successfully segments the organ structures at least 2000 iterations ahead of the other methods. Secondly, UGPCL (discard unreliable pseudo-labels) exhibits lower performance compared to the CPS +VCL. This suggests that removing voxels with unreliable pseudo-labels from these difficult organs results in substantial information loss, ultimately leading to a performance decline. In contrast, by fully utilizing a large number of unreliable voxels, our algorithm enhances performance and outperforms U^{2}PL (using complementary labels).

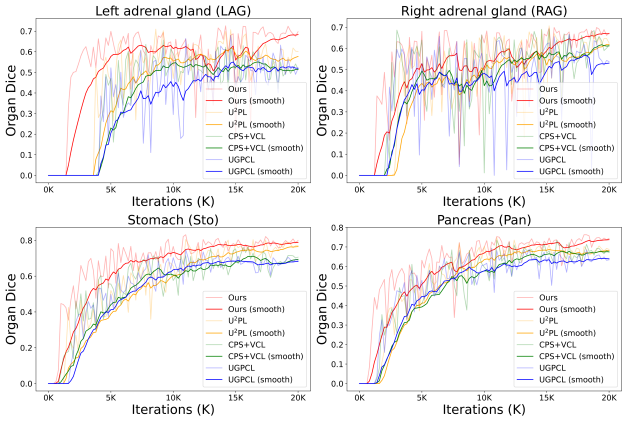

Figure 5. Dice curves generated by different methods on validation set of FLARE 2022 during network training. As seen, our method can significantly improve the performance on difficult organs.

4.3. Ablation Study

In this subsection, we conduct various ablations to better understand our design choices. For all the ablation experiments the models are trained on FLARE 2022 dataset with 10% labeled ratio.

Key Component Analysis. In Table.2, we first validate the importance of our proposed components by attaching them one at a time. The 1^{st}row represents the CPS baseline, on which our method is based. Next, in the 2^{nd}row, we enhance the model's performance by learning effective voxel feature representations through HqCL. Moreover, the 3^{rd}row gives the score when the DVCL is further incorporated. As seen, encouraging unreliable voxels to learn more

Baseline	HqCL	DVCL	NPS	QDS	Mean Dice for each organ													Mean Dice	Mean Jaccard
					Liv	Spl	Sto	L.kid	R.kid	Aor	Pan	IVC	Duo	Gal	Eso	RAG	LAG		
✔					96.92	91.86	77.02	92.70	92.71	92.25	69.39	81.91	65.94	75.12	72.78	63.56	58.96	79.32±0.46	68.14±0.61
✔	✔				97.15	91.12	77.26	94.33	**93.87**	92.15	74.48	**84.81**	69.79	84.60	75.23	67.23	63.89	82.00±0.16	71.50±0.16
✔	✔	✔			96.84	**95.04**	82.65	94.69	93.43	92.14	76.83	83.60	70.50	85.72	74.17	70.65	68.08	83.41±0.31	72.50±0.54
✔	✔	✔	✔		97.14	94.78	84.34	94.83	93.57	92.58	**77.18**	84.69	71.13	**86.24**	75.59	70.02	71.13	84.09±0.05	73.72±0.30
✔	✔	✔	✔	✔	**97.19**	93.37	**84.39**	**95.22**	93.58	**92.71**	77.03	84.72	**71.98**	85.81	**75.62**	**70.97**	**71.38**	**84.15±0.07**	**73.62±0.29**

Table 2. Ablation study on the effectiveness of each component under 10% partition protocol on FLARE 2022. **Best results are boldfaced.**

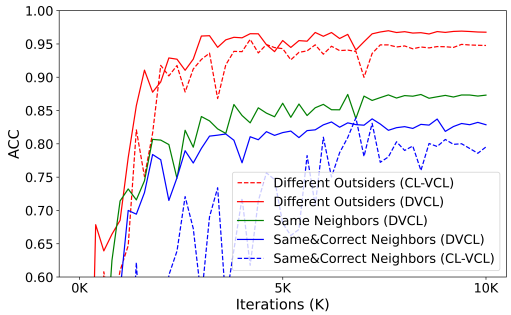

Figure 6. The five curves are: the ratio of features and their K'outsiders with different labels after various methods (solid/dashed red), the ratio of features with K same-labeled neighbors after DVCL (solid green), and the ratio of features with K correct same-labeled neighbors after different methods (solid/dashed blue).

discriminative features leads to significant improvements in Dice for difficult organs (*e.g.,* Sto:82.65%; Pan:76.83%; RAG:70.65%; LAG:68.08%;). Finally, as shown in the 4^{th} and 5^{th}rows, NPS and QDS provide more stable feature representations for DVCL and achieve the highest mean Dice (84.15%) and mean Jaccard (73.62%), outperforming the baseline by 4.83% and 5.48%, respectively.

Effectiveness of DVCL. We present the evolution of various statistical metrics during the training process on the MMWHS dataset in Fig.6. The solid and dashed red curves represent the ratio of unreliable voxels and their K' outsiders that do not belong to the same class after applying DVCL and the complementary label VCL (CL-VCL), respectively. The solid green curve represents the ratio of unreliable voxels with K neighbors, all of which belong to the same class, illustrating that unreliable voxels cluster during training. The solid and dashed blue curves represent the ratio of voxels with K neighbors that not only belong to the same class but are also correctly classified after applying DVCL and the CL-VCL, respectively. This demonstrates that, compared to the complementary labeling approach, our method enables unreliable voxels to learn more discriminative features, thereby achieving better intra-class aggregation and inter-class separation.

Effectiveness of DVCL Loss Term and NPS. We use class relationship likelihoods instead of directly pulling or pushing voxel features, as in traditional contrastive learning (*e.g.,* InfoNCE [33], SimCLR [5], NNCLR [13]). This approach surpasses conventional positive-negative comparisons, capturing distributional differences more effectively and clarifying semantic relationships. As shown in rows 1-4 of Table 3, our strategy achieves the best performance. Additionally, to reduce the potential impact of noisy neighbors, NPS demonstrates a greater performance improvement in rows 5-7 of Table.3, compared to methods that apply different weights based on cosine similarity (SNCLR [8]).

Extra Study. More investigations about ablation study for hyper-parameters in Sec.F of SM.

Methods	Mean Dice	Mean Jaccard
InfoNCE [33]	83.05±0.28	73.21±0.31
SimCLR [5]	82.77±0.15	72.45±0.33
NNCLR [13]	83.17±0.34	72.09±0.41
Log-likelihood [47]	83.41±0.31	72.50±0.54
SNCLR [8]	83.67±0.71	72.74±0.67
Log-likelihood+SNCLR [8]	83.69±0.33	72.71±0.26
Log-likelihood+NPS	**84.09±0.05**	**73.72±0.30**

Table 3. Ablation study on the DVCL loss terms and NPS.

5. Conclusion

In this work, we introduce a semi-supervised method for MoS, termed DVCL. Contrary to the previous works that use complementary labels, DVCL encourages unreliable voxels to learn more discriminative features, allowing them to fully enjoy the advantages of VCL. Additionally, we propose NPS and QDS to enhance the stability of feature representations in DVCL. Experimental findings illustrate that our method outperforms existing SSL frameworks, especially in difficult organs.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No.62372027, Grant No.62372028), Beijing Natural Science Foundation - Daxing Innovation United Foundation (No.L256027) and the National Natural Science Foundation of China (Grant No.62303127, Grant No.62273009, Grant No.62402032).

References

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8219–8228, 2021. 2

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. 1

[3] 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. 1

[4] Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, and Yan Wang. Magicnet: Semi-supervised multi-organ segmentation via magic-cube partition and recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23869–23878, 2023. 1, 3, 6, 7, 4, 5

[5] 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

[6] 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. 3, 6, 7, 4, 5

[7] Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. Offset: Segmentation-based focus shift revision for composed image retrieval. *arXiv preprint arXiv:2507.05631*, 2025. 1

[8] GE Chongjian, Jiangliu Wang, Zhan Tong, Shoufa Chen, Yibing Song, and Ping Luo. Soft neighbors are positive supporters in contrastive visual representation learning. In *The Eleventh International Conference on Learning Representations*. 8

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 6, 4, 7

[10] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27, 1967. 5

[11] Qinyi Deng, Yong Guo, Zhibang Yang, Haolin Pan, and Jian Chen. Boosting semi-supervised learning with contrastive complementary labeling. *Neural Networks*, 170:417–426, 2024. 1, 3, 6, 7, 4, 5

[12] Zhongjing Du, Xu Jiang, Peng Wang, Qizheng Zhou, Xi Wu, Jiliu Zhou, and Yan Wang. Lion: Label disambiguation for semi-supervised facial expression recognition with progressive negative learning. In *IJCAI*, pages 699–707, 2023. 1, 3

[13] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 5, 8

[14] Qianhan Feng, Lujing Xie, Shijie Fang, and Tong Lin. Bacon: Boosting imbalanced semi-supervised learning via balanced feature-level contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11970–11978, 2024. 1, 3, 6, 7, 4, 5

[15] Zhiheng Fu, Zixu Li, Zhiwei Chen, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. Pair: Complementarity-guided disentanglement for composed image retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1

[16] Jacob Goldberger, GeoffreyE. Hinton, SamT. Roweis, and Ruslan Salakhutdinov. Neighbourhood components analysis. *Neural Information Processing Systems,Neural Information Processing Systems*, 2004. 5

[17] Qinlei Huang, Zhiwei Chen, Zixu Li, Chunxiao Wang, Xuemeng Song, Yupeng Hu, and Liqiang Nie. Median: Adaptive intermediate-grained aggregation network for composed image retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1

[18] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18145–18155, 2023. 3

[19] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 6, 1, 5

[20] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas 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*, page 12, 2015. 6, 1

[21] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 552–561. Springer, 2020. 6, 7, 3, 4, 5

[22] Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5101–5109, 2025. 1

[23] Zixu Li, Zhiheng Fu, Yupeng Hu, Zhiwei Chen, Haokun Wen, and Liqiang Nie. Finecir: Explicit parsing of fine-grained modification semantics for composed image retrieval. *arXiv preprint arXiv:2503.21309*, 2025. 1

[24] Yiqun Lin, Huifeng Yao, Zezhong Li, Guoyan Zheng, and Xiaomeng Li. Calibrating label distribution for class-imbalanced barely-supervised knee segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–118. Springer, 2022. 3, 6, 7, 4, 5

[25] Hengyang Liu, Pengcheng Ren, Yang Yuan, Chengyun Song, and Fen Luo. Uncertainty global contrastive learning framework for semi-supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2024. 3

[26] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8801–8809, 2021. 6, 7, 3, 4, 5

[27] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. 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 318–329. Springer, 2021. 1

[28] Jie Ma, Chuan Wang, Yang Liu, Liang Lin, and Guanbin Li. Enhanced soft label for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1185–1195, 2023. 4

[29] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, Fan Zhang, Wentao Liu, YuanKe Pan, Shoujin Huang, Jiacheng Wang, Mingze Sun, Weixin Xu, Dengqiang Jia, Jae Won Choi, Natália Alves, Bram de Wilde, Gregor Koehler, Yajun Wu, Manuel Wiesenfarth, Qiongjie Zhu, Guoqiang Dong, Jian He, the FLARE Challenge Consortium, and Bo Wang. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023. 2, 6, 1

[30] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023. 3

[31] Hui Meng, Haochen Zhao, Ziniu Yu, Qingfeng Li, and Jianwei Niu. Uncertainty-aware mean teacher framework with inception and squeeze-and-excitation block for miccai flare22 challenge. In *MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation*, pages 245–259. Springer, 2022. 1

[32] Juzheng Miao, Si-Ping Zhou, Guang-Quan Zhou, Kai-Ni Wang, Meng Yang, ShouJun Zhou, and Yang Chen. Sc-ssl: Self-correcting collaborative and contrastive co-training model for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2023. 3

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5, 8

[34] Wenbo Qi, Jiafei Wu, and SC Chan. Gradient-aware for class-imbalanced semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pages 473–490. Springer, 2024. 3

[35] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, and Shaozi Li. Mutual learning with reliable pseudo label for semi-supervised medical image segmentation. *Medical Image Analysis*, 94:103111, 2024. 1

[36] Cheng Tang, Xinyi Zeng, Luping Zhou, Qizheng Zhou, Peng Wang, Xi Wu, Hongping Ren, Jiliu Zhou, and Yan Wang. Semi-supervised medical image segmentation via hard positives oriented contrastive learning. *Pattern Recognition*, 146: 110020, 2024. 3

[37] 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. 6, 7, 3, 4, 5

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2, 5

[39] Haonan Wang and Xiaomeng Li. Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 582–591. Springer, 2023. 1, 3, 6, 7, 4, 5

[40] Haonan Wang and Xiaomeng Li. Towards generic semi-supervised framework for volumetric medical image segmentation. *arXiv preprint arXiv:2310.11320*, 2023. 3

[41] Tao Wang, Jianglin Lu, Zhihui Lai, Jiajun Wen, and Heng Kong. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 1444–1450, 2022. 1, 3, 6, 7, 4, 5

[42] 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. 2, 3

[43] Xiaoyang Wang, Bingfeng Zhang, Limin Yu, and Jimin Xiao. Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3114–3123, 2023. 4

[44] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022. 1, 3, 4, 6, 7, 5

[45] Lu Wen, Zhenghao Feng, Yun Hou, Peng Wang, Xi Wu, Jiliu Zhou, and Yan Wang. Dcl-net: Dual contrastive learning network for semi-supervised multi-organ segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1876–1880. IEEE, 2024. 3

[46] Huisi Wu, Baiming Zhang, Cheng Chen, and Jing Qin. Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning. *IEEE Transactions on Medical Imaging*, 43(2):649–661, 2024. 1

[47] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 5, 8

[48] Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical image analysis*, 65:101766, 2020. 3

[49] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David Clifton, S Kevin Zhou, Lawrence Staib, and James Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. *Advances in neural information processing systems*, 36, 2024. 1

[50] 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. 4, 6, 7, 3, 5

[51] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 408–416. Springer, 2017. 6, 7, 3, 4, 5

[52] Haochen Zhao, Jianwei Niu, Hui Meng, Yong Wang, Qingfeng Li, and Ziniu Yu. Focal u-net: A focal self-attention based u-net for breast lesion segmentation in ultrasound images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1506–1511. IEEE, 2022. 1

[53] Haochen Zhao, Hui Meng, Deqian Yang, Xiexiao zheng, Xiaoze Wu, Qingfeng Li, and Jianwei Niu. Guidednet: Semi-supervised multi-organ segmentation via labeled data guide unlabeled data. In *ACM Multimedia 2024*, 2024. 1, 3

[54] Qianfei Zhao, Lanfeng Zhong, Jianghong Xiao, Jingbo Zhang, Yinan Chen, Wenjun Liao, Shaoting Zhang, and Guotai Wang. Efficient multi-organ segmentation from 3d abdominal ct images with lightweight network and knowledge distillation. *IEEE Transactions on Medical Imaging*, 42(9):2513–2523, 2023. 1

[55] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, and Alan Yuille. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140. IEEE, 2019. 3

[56] Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis*, 31:77–87, 2016. 6, 1