HHS Public Access

Author manuscript

*IEEE Trans Med Imaging*. Author manuscript; available in PMC 2023 July 06.

Published in final edited form as:

*IEEE Trans Med Imaging*. 2022 September ; 41(9): 2228–2237. doi:10.1109/TMI.2022.3161829.

# SimCVD: Simple Contrastive Voxel-Wise Representation Distillation for Semi-Supervised Medical Image Segmentation

**Chenyu You**,
Department of Electrical Engineering, Yale University, New Haven, CT 06520 USA

**Yuan Zhou [Member, IEEE]**,
Department of Radiology and Biomedical Imaging, Yale University, New Haven, CT 06520 USA

**Ruihan Zhao**,
Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA

**Lawrence Staib [Senior Member, IEEE]**,
Department of Radiology and Biomedical Imaging, the Department of Biomedical Engineering, and the Department of Electrical Engineering, Yale University, New Haven, CT 06520 USA

**James S. Duncan [Life Fellow, IEEE]**
Department of Radiology and Biomedical Imaging, the Department of Biomedical Engineering, and the Department of Electrical Engineering, Yale University, New Haven, CT 06520 USA

## Abstract

Automated segmentation in medical image analysis is a challenging task that requires a large amount of manually labeled data. However, most existing learning-based approaches usually suffer from limited manually annotated medical data, which poses a major practical problem for accurate and robust medical image segmentation. In addition, most existing semi-supervised approaches are usually not robust compared with the supervised counterparts, and also lack explicit modeling of geometric structure and semantic information, both of which limit the segmentation accuracy. In this work, we present SimCVD, a simple contrastive distillation framework that significantly advances state-of-the-art voxel-wise representation learning. We first describe an unsupervised training strategy, which takes two views of an input volume and predicts their signed distance maps of object boundaries in a contrastive objective, with only two independent dropout as mask. This simple approach works surprisingly well, performing on the same level as previous fully supervised methods with much less labeled data. We hypothesize that dropout can be viewed as a minimal form of data augmentation and makes the network robust to representation collapse. Then, we propose to perform structural distillation by distilling pair-wise similarities. We evaluate SimCVD on two popular datasets: the Left Atrial Segmentation Challenge (LA) and the NIH pancreas CT dataset. The results on the LA dataset demonstrate that, in two types of labeled ratios (*i.e.*, 20% and 10%), SimCVD achieves an average Dice score of 90.85% and 89.03% respectively, a 0.91% and 2.22% improvement compared to previous best results. Our method can be trained



(Corresponding author: Chenyu You.).

in an end-to-end fashion, showing the promise of utilizing SimCVD as a general framework for downstream tasks, such as medical image synthesis, enhancement, and registration.

**Index Terms—**

Medical image segmentation; contrastive learning; knowledge distillation; geometric constraints

## I. Introduction

Medical image segmentation is a popular task in both machine learning and medical imaging communities [1]–[5]. Compared to traditional segmentation approaches, deep neural network based segmentation methods have achieved much stronger performance in recent years with huge advances in representation learning [6]–[13]. However, previous state-of-the-art approaches are mostly trained with a large amount of labeled data, which pose significant practical challenges in many medical segmentation tasks where there is a scarcity of labeled data due to the heavy burden of annotating images.

In recent years, a wide variety of semi-supervised methods [14]–[25] have been designed to tackle these issues, which learn from limited labeled data along with a large amount of unlabeled data, achieving significant improvements in accuracy and greatly reducing the labeling cost. The common paradigms include adversarial learning, knowledge distillation, and self-supervised learning. Contrastive learning, a sub-area of self-supervised learning, has recently been noted as a promising direction since it has shown great promise in learning useful representations with limited human supervision [24], [26]–[29]. This is often best understood as pulling together semantically similar (*positive*) samples and pushing apart non-similar (*negative)* samples in a shared latent space. The representations uncovered by these contrastive objectives are capable of boosting the performance of any vision system especially in scenarios where the amount of annotated data available for the downstream tasks is extremely low, which is well suited for medical image analysis.

Despite advances in semi-supervised learning benchmarks, previous methods still face several major challenges: (1) *Suboptimal performance*: although prior works have achieved promising segmentation accuracy in the setting of limited annotations, semi-supervised models are usually not robust due to some information loss, compared with fully-supervised counterparts; (2) *Geometric information loss*: previous segmentation networks are poor at characterizing geometry, *i.e.*, leveraging the intrinsic geometric structure of the images, such as the object boundary. As a consequence, it is often hard to accurately recognize object contours; and (3) *Generalization ability*: considering the limited amount of training data, training deep models is usually deficient due to over-fitting and co-adapting [30], [31].

In this work, we address the question: can we advance state-of-the-art voxel-wise representation learning in a more extreme few-annotation-setting for medical image segmentation? To this end, we present SimCVD, a simple contrastive voxel-wise representation distillation framework, which can be utilized to produce superior voxel-wise representations from unlabeled data for improving network performance. Our proposed SimCVD, built upon the mean-teacher framework [32], can address the above-mentioned

challenges as follows. First, SimCVD predicts the output geometric representations with only two different *dropout* [33] masks (Figure 1). In other words, we pass two views of the geometric representations to the mean-teacher model, obtain two representations as "positive pairs", by applying two independent dropout masks, and learn effective representations by efficiently associating positives and disassociating negatives in the shared latent space. Though this unsupervised learning strategy is simple, we find this approach is strikingly effective compared to other common data augmentation techniques (*e.g.*, inpainting and local shuffle pixel). More importantly, as we will show, it achieves comparable performance to previous fully-supervised approaches. Through a series of thorough analyses, we find that dropout can be viewed as minimal data augmentation for performance improvement, and it can effectively regularize the training of deep neural networks, avoid representation collapse and enhance model generalization.

Second, we attribute the cause of the *geometric information loss* to the need for geometric shape constraints. We address this challenge by performing multi-task learning that jointly predicts a segmentation map along with a signed distance map (SDM) [12], [21], [34]–[36]. The SDM calculates the signed distance function of the object, *i.e.*, the distance of a voxel from the boundary of the object, with the sign determined by whether the voxel is within the object. Thus, it can be viewed as a global shape constraint on the labeled data. Considering that the SDM can provide a more flexible geometric measure of the object boundary, we move beyond the supervised learning scheme and exploit the regularity in geometric shapes among different object classes through distilling "boundary-aware" knowledge via a contrastive objective among the unlabeled data. This enables the model to learn boundary-aware features more effectively by encouraging the networks to produce segmentation maps with similar distance map distributions on the entire dataset.

Third, it is challenging to train the segmentation model on small training sets since deep neural networks trained on a limited amount of data are prone to over-fitting. To this end, we propose to use knowledge distillation (KD), which has been shown to be effective in segmentation and classification tasks [37]–[39]. The key idea of KD is that a teacher model is first trained, and then used to guide the training of the student model for improving generalization ability. In the medical domain, most existing KD methods [40], [41] simply consider the segmentation problem as a pixel/voxel-level classification problem. In contrast, considering that medical image semantic segmentation is a structured prediction problem, we present a novel structured knowledge *pair-wise distillation*, which further use the structural knowledge from the mean-teacher model, while avoiding co-adapting and over-fitting.

Our contributions are summarized as follows. First, we propose a novel contrastive distillation model termed SimCVD featured by (i) boundary-aware representations that incorporate rich information of the object shape, (ii) a distillation objective which contrasts different distance map distributions jointly in the shared latent space, and (iii) a pair-wise distillation objective to further distill pair-wise structural knowledge. Second, we demonstrate that, in the setting of very limited annotation, simply using dropout can deliver more robust end-to-end segmentation performance compared to heavily relying on a large amount of labeled data. Third, we conduct experiments on two popular benchmark datasets

to evaluate SimCVD. The results demonstrate that SimCVD significantly outperforms other state-of-the-art semi-supervised approaches, while achieving competitive performance compared to fully-supervised counterparts.

## II. Related Work

### A. Semi-Supervised Medical Image Segmentation

In recent years, substantial efforts [14]–[17], [20], [42]–[47] have been devoted to incorporating unlabeled data to improve network performance due to limited annotations. *Yu et al.* [20] investigated an uncertainty map based on the mean-teacher framework [32] to guide the student network to capture better features. Li *et al.* [21] proposed to use signed distance fields for boundary prediction to improve the performance. Also, Luo *et al.* [23] proposed a dual-task-consistency (DTC) model for semi-supervised medical image segmentation by jointly predicting the pixel-wise segmentation maps and the global-level level set representations on the unlabeled data. Our method aims at a more practical and challenging scenario: we train our model in a more extreme few-annotation setting that relies only on a small number of annotations, while achieving superior segmentation accuracy.

### B. Contrastive Learning

Self-supervised learning (SSL) [46], [48]–[50] has provided robust benefits to vision tasks by learning effective visual representations from unlabeled data in an unsupervised setting. It is based on a commonly-held belief that superior performance gains can be achieved through improved representation learning. Recently, contrastive learning, a type of self-supervised learning, has received a lot of interests [24], [26], [29], [48], [51]–[56]. The key idea of contrastive learning is to learn powerful representations that optimize similarity constraints to discriminate similar pairs (*positive*) and dissimilar pairs (*negative*) within a dataset. The primary stream of subsequent work focuses on the choice of dissimilar pairs, which is critical to the quality of learned representations. The loss function used to quantify the contrast is chosen from several options, such as InfoNCE [57], Triplet [58], and so on. Recent studies [51], [53] introduced memory bank or momentum contrast to use more negative samples for contrast computation. In the context of medical imaging, Chaitanya *et al.* [24] extended a contrastive learning framework to extract global and local cues in a stage-wise way, which requires human intervention and extensive training time. In contrast to Chaitanya *et al.* [24], our unified work focuses on explicit modeling of the intrinsic geometric structure of the semantic objects in an end-to-end manner, and hence is able to recognize object boundaries more effectively and efficiently.

### C. Knowledge Distillation

The idea of knowledge distillation is to minimize the KL-divergence between the output distributions of the teacher model and the student model, and thus avoid over-fitting. KD has been applied to a variety of tasks [59]–[62], including image classification [37], [63]–[65] and semantic segmentation [39], [66]. Recent works [63], [64] found that the student model can outperform the teacher model when they share the same network architecture. Zhang *et al.* [67] proposed to collaboratively train multiple student models with co-distillation, which improves performance of those individual models. At the same

time, in the context of medical imaging, among those existing state-of-the-art KD methods, the self-ensemble mean-teacher framework [32] is widely explored for image segmentation. Different from the existing methods that separately exploit class probabilities for each voxel, we consider knowledge distillation as a structured prediction problem by matching the relational similarity among all pairs of voxels from the encoded feature maps of the meanteacher model. We have found that our approach significantly improves learning better voxel-wise representations.

## III. Method

In this section, we introduce SimCVD, a semi-supervised segmentation network, which is built from scratch by effectively leveraging scarce labeled data and ample unlabeled data for improving end-to-end voxel-wise representation learning (See Figure 1). We first overview our proposed SimCVD and then describe the task formulation of SimCVD. Finally, we detail each component of SimCVD in the following subsections.

### A. Overview

We aim to construct an end-to-end voxel-wise contrastive distillation algorithm to learn boundary-aware representations in the setting of extremely few annotations for volumetric medical imaging segmentation. Although the accuracy of supervised models is usually higher than that of semi-supervised models, the former requires much more labeled data than the latter. In many clinical situations, we only have few annotated data but a large amount of unlabeled data. This situation necessitates a semi-supervised segmentation algorithm that can utilize the unlabeled data to improve the segmentation performance.

To this end, we propose a novel contrastive distillation framework to advance state-of-the-art voxel-wise representation learning. In particular, our base multi-task segmentation network tackles two tasks simultaneously: classification and regression. Specifically, the segmentation network takes the input volume batch and jointly predicts the probability maps (classification) and the SDMs of the object (regression). To obtain better representations, we propose to perform structured distillation in the latent *feature* space, followed by contrasting the boundary-aware features in the *prediction* space, to learn more effective boundary-aware representations from 3D unlabeled data by regularizing the embedding space and exploring the geometric and spatial context of training voxels. At test time, we remove the mean teacher and two projection heads, and only the student network is deployed for the medical segmentation tasks.

### B. Task Formulation

In this work, we consider a set of training data (3D images) including $N$ labeled data and $M$ unlabeled data, where $N \ll M$. For simplicity, we denote the small set of labeled data as $\mathscr{D}_l = \left\{\left(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Y}_i^{\text{sdm}}\right)\right\}_{i=1}^{N}$, and abundant unlabeled data as $\mathscr{D}_u = \{\mathbf{X}_i\}_{i=N+1}^{N+M}$, where $\mathbf{X}_i \in \mathbb{R}^{H \times W \times D}$ is the volume input, $\mathbf{Y}_i \in \{0,1\}^{H \times W \times D}$ is the ground-truth label, and $\mathbf{Y}_i^{\text{sdm}} \in \mathbb{R}^{H \times W \times D}$ is the computed ground truth SDMs from $\mathbf{Y}_i$, which measures the distance

from each voxel to the object boundary. Every 3D image $\mathbf{X}_i$ consists of a set of 2D image slices $\mathbf{X}_i = [\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,D}]$ where $\mathbf{x}_{i,j} \in \mathbb{R}^{H \times W}$.

Our proposed SimCVD framework consists of a mean-teacher network, $\mathcal{F}_t(\mathbf{X}; \theta_t)$, and a student network, $\mathcal{F}_s(\mathbf{X}; \theta_s)$. Inspired by recent work [14], [32], the optimization of these two networks can be achieved with an exponential moving average (EMA) which uses a weighted combination of the parameters of the student network and the parameters of the teacher network to update the latter. This strategy has been widely shown to improve training stability and the model's final performance. Motivated by this idea, our training strategy is divided into two steps. At each iteration, we first optimize the student network $\mathcal{F}_s$ by stochastic gradient descent. Then we update the teacher weights $\theta_t$ using an exponential moving average of the student weights $\theta_s$.

The inputs to the two networks are perturbed versions of the same image. That is, given a volume input $\mathbf{X}_i$, we first add different perturbations (i.e., affine transformation and random crop) to generate two different images $\mathbf{X}_i^t$ and $\mathbf{X}_i^s$. We then feed $\mathcal{F}_t$ and $\mathcal{F}_s$ with these two corresponding augmented images to obtain two confidence score (probability) maps $\mathbf{Q}_i^t$ and $\mathbf{Q}_i^s$.

Before we present our proposed SimCVD in detail, we first describe our base architecture below.

## C. Base Architecture

Our base architecture adopts V-Net [20] as the network backbone, which consists of an encoder network $e_t : \mathbb{R}^{H \times W \times D} \rightarrow \mathbb{R}^{H' \times W' \times D' \times D_e}$ and a decoder network $d_t : \mathbb{R}^{H' \times W' \times D' \times D_e} \rightarrow [0,1]^{H \times W \times D} \times [-1,1]^{H \times W \times D}$ for the teacher network, and similarly $e_s, d_s$ for the student network, *i.e.*, $\mathcal{F}_t = d_t \circ e_t$ and $\mathcal{F}_s = d_s \circ e_s$. $H', W', D'$ are the size of the hidden pattern and $D_e$ is the encoded feature dimension. Inspired by previous work on medical imaging segmentation [12], [21], we incorporate multi-task learning into $\mathcal{F}$ to jointly perform both classification and regression tasks.

Given input $\mathbf{X}_i$, the classification branch is designed to generate the probability map $\mathbf{Q}_i^s \in [0,1]^{H \times W \times D}$, and the regression branch is designed to predict the SDM $\mathbf{Q}_i^{s,\text{sdm}} \in [-1,1]^{H \times W \times D}$. The design of the regression branch is simple yet effective, *only* including the hyperbolic tangent function. This design brings two clear benefits: (1) we can eventually encode rich geometric structure information to improve segmentation accuracy, and (2) we can implicitly enforce continuity and smoothness terms for better segmentation maps. Similarly, we have outputs $\mathbf{Q}_i^t$ and $\mathbf{Q}_i^{t,\text{sdm}}$ from the teacher network.

**1) Supervised Loss $\mathcal{L}_{\text{sup}}$:** For training on labeled data, we define the supervised loss as:

$$\mathscr{L}_{\text{sup}} = \frac{1}{N}\sum_{i=1}^{N}\mathscr{L}_{\text{seg}}(\mathbf{Q}_i^{s}, \mathbf{Y}_i) + \frac{\alpha}{N}\sum_{i=1}^{N}\mathscr{L}_{\text{mse}}\left(\mathbf{Q}_i^{s,\text{sdm}}, \mathbf{Y}_i^{\text{sdm}}\right), \tag{1}$$

where $\mathscr{L}_{\text{seg}}$ denotes the segmentation loss (Dice and Crossentropy) [20], and $\mathscr{L}_{\text{mse}}$ is the mean squared error loss. $\alpha$ is a hyperparameter. Note that the SDM loss [12] is imposed as geometric constraints in training.

## D Boundary-Aware Contrastive Distillation

Many prior methods distill knowledge merely in the shared prediction space by delivering the student network that matches the accuracy of the teacher network. However, this strategy is not robust for the following reasons: (1) the learned voxel-wise representations from the mean-teacher model are usually not robust due to the lack of geometric information; (2) the segmentation model still suffers from *generalization* issues; and (3) the network performance needs to be further improved. Therefore, we propose to perform boundary-aware contrastive distillation to train our model for better segmentation accuracy.

Our method differs from previous state-of-the-art methods in three aspects: (1) SimCVD imposes the global consistency in object boundary contours to capture more effective *geometric information*; (2) previous methods follow the standard setting in considering the relations of local patches, while SimCVD aims to exploit correlations among all pairs of voxels to improve robustness; and (3) due to the computational cost, SimCVD does not use a large memory bank. SimCVD trains the contrastive objective as an auxiliary loss during the volume batch updates. To specify our voxel-wise contrastive distillation algorithm on unlabeled sets, we define two discrimination terms: boundary-aware contrastive loss and pair-wise distillation loss.

**1) Boundary-Aware Contrastive Loss $\mathscr{L}_{\text{contrast:}}$** We describe our unsupervised boundary-aware contrastive objective as follows. Our key idea is to make use of "boundary-aware" knowledge by a contrastive learning objective that enforces the consistency of the predicted SDM outputs on the unlabeled set during training. The key ingredient to working with two views of input images is to apply *dropout* as mask. Specifically, given the collection of an input volume $\mathbf{X}_i$, the student SDM $\mathbf{Q}_i^{s,\text{sdm}}$, the teacher SDM $\mathbf{Q}_i^{t,\text{sdm}}$, we first directly add them up to build two boundary-aware features: $\mathbf{Q}_i^{s,\text{ba}} = \mathbf{X}_i + \mathbf{Q}_i^{s,\text{sdm}}$ and $\mathbf{Q}_i^{t,\text{ba}} = \mathbf{X}_i + \mathbf{Q}_i^{t,\text{sdm}}$. Then, we feed them into the projection heads with two independent dropout masks $z_i^{s}$, $z_i^{t}$, and contrast *positives* and *negatives* by using the InfoNCE loss. We denote the same slice from the two boundary-aware features as *positive*, and slices at different locations or from different inputs as *negative*.

The boundary-aware features are created by adding the original 3D volume to the SDM because we want to fuse both the distance and the intensity information. Another way to achieve this is concatenation — adding another dimension to the feature tensor — requires a more complex projection head which is more prone to over-fitting. Thus, the projection head $\mathscr{H}: \mathbb{R}^{H\times W\times D} \rightarrow \mathbb{R}^{D_h\times D}$ encodes each 2D slice to a $D_h$-dimensional feature vector. The implementation is simple, which includes an alpha dropout [71], an adaptive average

pooling, and a 3-layer multilayer perceptron (MLP). Here the MLP is designed to convert each 2D slice to a vector.

Denoting the output of the projection head as $\mathbf{H}_i^s = \mathscr{H}(\mathbf{Q}_i^{s,\,\mathrm{ba}}; z_i^s), \mathbf{H}_i^t = \mathscr{H}(\mathbf{Q}_i^{t,\,\mathrm{ba}}; z_i^t)$, and the $j$th row of $\mathbf{H}_i$ as $\mathbf{h}_{i,j}$, the InfoNCE loss [57] is defined by:

$$\mathscr{L}(\mathbf{h}_{i,j}^t, \mathbf{h}_{i,j}^s) = -\log \frac{\exp(\mathbf{h}_{i,j}^t \cdot \mathbf{h}_{i,j}^s/\tau)}{\sum_{k,l} \exp(\mathbf{h}_{i,j}^t \cdot \mathbf{h}_{k,l}^s/\tau)}, \tag{2}$$

where $\tau$ is a temperature hyperparameter. The indices $k$ and $l$ in the denominator are randomly sampled from a mini-batch of images such that $B$ 2D slices are sampled in total. $i$ and $j$ denote the 3D image index and slice index, respectively. The $\mathbf{h}_{k,l}^s$ 's in the denominator that are not $\mathbf{h}_{i,j}^s$ are called negative samples. Inspired by the recent success [24], our boundary-aware contrastive loss is defined as:

$$\mathscr{L}_{\text{contrast}} = \frac{1}{|\mathscr{N}^+|} \sum_{\forall (i,j) \in \mathscr{N}^+} [\mathscr{L}(\mathbf{h}_{i,j}^t, \mathbf{h}_{i,j}^s) + \mathscr{L}(\mathbf{h}_{i,j}^s, \mathbf{h}_{i,j}^t)], \tag{3}$$

where $\mathscr{N}^+ = \{(i,j): i = N+1, \ldots, N+M, j = 1, \ldots, D\}$ denotes a collection of the positive 2D slice pairs. Note that the index $i$ in $\mathscr{N}^+$ is over all the unlabeled data, hence these data affect the training of $\mathscr{F}_s$ and $\mathscr{F}_t$.

**2) Pair-Wise Distillation Loss $\mathscr{L}_{pd}$:** On one hand, boundarya-ware contrastive objectives uncover distinctive global boundary-aware representations that benefit the training of downstream tasks, *e.g.* object classification, when limited labeled data is available. On the other hand, dense predictive tasks, *e.g.* semantic segmentation, may require more discriminative spatial representations. As complementary to boundary-aware contrastive objectives, a promising local pair-wise strategy is vital for the medical image segmentation tasks. With this insight, we propose to perform voxel-to-voxel pair-wise distillation to explicitly explore structural relationships between voxel samples to improve spatial labeling consistency.

In our implementation, we enforce such a constraint on the hidden patterns from the encoders $e_t$ and $e_s$. Specifically, let $\mathbf{V}_i^t \in \mathbb{R}^{H'W'D' \times D_e}$ and $\mathbf{V}_i^s \in \mathbb{R}^{H'W'D' \times D_e}$ be the first-3-dimension-flattened hidden patterns of $e_t(\mathbf{X}_i)$ and $e_s(\mathbf{X}_i)$ respectively, and $\mathbf{v}_{i,j}$ be the $j$th row of $\mathbf{V}_i$. The pair-wise distillation loss is defined as:

$$\mathscr{L}_{\text{pd}} = -\frac{1}{M} \sum_{i=N+1}^{N+M} \sum_{j=1}^{H'W'D'} \log \frac{\exp(s(\mathbf{v}_{i,j}^s, \mathbf{v}_{i,j}^t))}{\sum_k \exp(s(\mathbf{v}_{i,j}^s, \mathbf{v}_{i,k}^t))}, \tag{4}$$

where $s(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\| \mathbf{v}_1 \| \, \| \mathbf{v}_2 \|}$ measures the cosine of the angle between two $\mathbf{v}$ 's as their similarity. Note again that this loss also involves all the unlabeled data.

**3) Consistency Loss $\mathscr{L}_{\text{con:}}$:** Inspired by recent work [14], [32], consistency is designed to further encourage training stability and performance improvements on the unlabeled set. In our implementations, we first perform different perturbation operations on the unlabeled input volume $\mathbf{X}_i$, *i.e.*, adding noise $\eta_i$, and then define the consistency loss as:

$$\mathscr{L}_{\text{con}} = \frac{1}{M}\sum_{i=N+1}^{N+M} \mathscr{L}_{\text{mse}}\left(\mathscr{F}_s(\mathbf{X}_i^s + \eta_i^s), \mathscr{F}_t(\mathbf{X}_i^t + \eta_i^t)\right). \tag{5}$$

**4) Overall Training Objective:** SimCVD is a general semi-supervised framework for combining contrastive distillation with geometric constraints. In our experiments, we train SimCVD with two objective functions — a supervised objective and an unsupervised objective. For the labeled data, we define the supervised loss in Section III-C. For the unlabeled data, the unsupervised training objective consists of the boundary-aware contrastive loss, pair-wise distillation loss, and consistency loss in Section III-D. The overall loss function is:

$$\mathscr{L} = \mathscr{L}_{\text{sup}} + \lambda\mathscr{L}_{\text{contrast}} + \beta\mathscr{L}_{\text{pd}} + \gamma\mathscr{L}_{\text{con}}, \tag{6}$$

where $\lambda, \beta, \gamma$ are hyperparameters that balance each term.

# IV. Experimental Setup

## A. Dataset and Pre-Processing

We evaluated our approach on two popular benchmark datasets: the Left Atrium (LA) MR dataset from the Atrial Segmentation Challenge,[1] and the NIH pancreas CT dataset [72]. For the Left Atrium dataset, it comprises 100 3D gadoliniumenhanced MR imaging scans (GE-MRIs) with expert annotations, with an isotropic resolution of 0.625 × 0.625 × 0.625mm$^3$. Following the experimental setting in [20], we use 80 scans for training, and 20 scans for evaluation. We employ the same pre-processing methods by cropping all the scans at the heart region and normalizing the intensities to zero mean and unit variance. All the training sub-volumes are augmented by random cropping to 112 × 112 × 80mm$^3$. For the pancreas dataset, it contains 82 contrast-enhanced abdominal CT scans. Following the experimental settings in [23], we randomly select 62 scans for training, and 20 scans for evaluation. In the pre-processing, we first truncate the intensities of the CT images into the window [−125, 275] HU [73], and then resample all the data into a fixed isotropic resolution of 1.0 × 1.0 × 1.0mm$^3$. Finally, we crop all the scans centering at the pancreas region, and normalize the intensities to zero mean and unit variance. All the training sub-volumes are augmented by random cropping to 96 × 96 × 96mm$^3$. In this study, we compare all the methods on LA and the pancreas dataset with respect to 20% labeled ratio. To emphasize the effectiveness of SimCVD, we further validate all the methods with respect to 10% labeled ratio on LA dataset.

[1] http://atriaseg2018.cardiacatlas.org/

### B. Implementation Details

In this study, all evaluated methods are implemented in PyTorch, and trained for 6000 iterations on an NVIDIA 1080Ti GPU with a batch size of 4. For data augmentation, we use standard data augmentation techniques (*i.e.*, random rotation, flipping, and cropping). We set the hyper-parameters $\alpha, \lambda, \beta, \gamma, \tau$ as 0.1, 0.5, 0.1, 0.1, 0.5, respectively. For the projection head, we set $p = 0.1$ in the *AlphaDropout* layer, and output size 128 × 128 for *AdaptiveAvgPool2d*. We use SGD optimizer with a momentum of 0.9 and a weight decay of 0.0005 to optimize network parameters. The initial learning rate is set as 0.01 and divided by 10 every 3000 iterations. For EMA updates, we follow the experimental setting in [20], where the EMA decay rate $\alpha$ is set to 0.999. We use the time-dependent Gaussian warming-up function $\Psi_{\text{con}}(t) = \exp\left(-5(1 - t/t_{\max})^2\right)$ to ramp up parameters, where $t$ and $t_{\max}$ denote the current and the maximum training step, respectively. For fairness, we do not adopt any post-processing step.

In the testing stage, we adopt four metrics to evaluate the segmentation performance: Dice coefficient (Dice), Jaccard Index (Jaccard), 95% Hausdorff Distance (95HD), and Average Symmetric Surface Distance (ASD). Following [20] and [23], we adopt a sliding window strategy, which uses a stride with 18 × 18 × 4 for the LA and 16 × 16 × 16 for the pancreas.

## V. Results

### A. Experiments: Left Atrium

We compare SimCVD with published results from previous state-of-the-art semi-supervised segmentation methods, including V-Net [7], MT [32], DAN [15], CPS [68], Entropy Mini [69], UA-MT [20], ICT [70], SASSNet [21], DCT [23], and Chaitanya *et al.* [24] on the LA dataset in two labeled ratio settings (*i.e.*, 10% and 20%).

The quantitative results on the LA dataset are shown in Table I. SimCVD substantially improved the segmentation accuracy in both 10% and 20% labeled cases. The results are visualized in Fig 2. Specifically, in the setting of 20% labeled ratio, our proposed SimCVD raises the previous best average results from 89.94% to 90.85% and from 81.82% to 83.80% in terms of Dice and Jaccard, even achieving comparable performance to the fully supervised baseline. Using the 10% labeled ratio, SimCVD further advances the state-of-the-art results from 87.49% to 89.03% in Dice. The gains in Jaccard, ASD, and 95HD are also substantial, achieving 80.34%, 2.59, and 8.34, respectively. This suggests that: (1) taking voxel samples with a contrastive objective yields better voxel embeddings; (2) incorporating pair-wise spatial labeling consistency can boost the performance by accessing more structural knowledge; and (3) utilizing a geometric constraint (*i.e.*, SDM) is capable of helping identify more accurate boundaries. Leveraging all these aspects, we can observe consistent performance gains.

### B. Experiments: Pancreas

To further evaluate the effectiveness of SimCVD, we compare our model on the pancreas CT dataset. Experimental results on the pancreas CT dataset are summarized in Table II. We observe that our model consistently outperforms all previous methods, achieving up to

6.72% absolute improvements in Dice. As shown in Figs. 2 and 3, our method is capable of predicting high-quality object segmentation, considering the fact that the improvement in such a setting is difficult. This demonstrates: (1) the necessity of comprehensively considering both boundary-aware contrast and pair-wise distillation; and (2) the efficacy of global shape information. Compared to the previous strong models, our approach achieves large improvements on all the datasets, demonstrating its effectiveness.

## VI. Ablation Study

In this section, we conduct extensive studies to better understand SimCVD. We justify the inner working of SimCVD from two perspectives: (1) boundary-aware contrastive distillation (Section VI-A), and (2) the projection head (Section VI-B). In these studies, we evaluate our proposed method on the LA dataset with 10% labeled ratio (8 labeled and 72 unlabeled).

### A. Analysis on Boundary-Aware Contrastive Distillation

**1) Ablation on Model Component:** In the model formulation, our motivation is to advance state-of-the-art voxel-wise representations by capturing the geometric and semantic information in 3D space. Rather than transferring knowledge across confidence score maps directly, our SimCVD distills "boundary-aware" knowledge from the teacher network. To validate the idea of boundary-aware contrastive distillation, we compare SimCVD to an ablative baseline (*i.e.*, SimCVD w/o SDM). Table III (a) compares each component of SimCVD in the 10% labeled setting. First, we observe that removing the SDMs in training hurts the segmentation performance by −0.79%,−1.27%,−1.6, and −3.09 absolute differences in terms of Dice, Jaccard, ASD, and 95HD. This confirms our intuition that the learned boundary-aware representations provide a good prior for improving segmentation accuracy. We also find that using adaptive max pooling strategy (*i.e.*, SimCVD w/ adaptive max pooling) largely degrades the segmentation performance. Our segmentation results demonstrate that SimCVD is an effective approach, outperforming the best previous method with +0.71%,+1.08%, +1.20, and +6.35 absolute differences in terms of Dice, Jaccard, ASD, and 95HD. We hypothesize that it is because "w/ adaptive max pooling" leads to information loss during training.

**2) Ablation on Loss Formulation:** In the loss formulation, our main idea is to pull closer similar (*positive*) pairs upon the same threshold, while pushing apart dissimilar (*negative*) pairs. Our learning objective is designed to jointly exploit effective correlations in the *prediction* and *feature* space in an informative way. To evaluate the effectiveness of each objective term, we conduct ablation studies by removing each term separately. As shown in Table III (b), we observe that SimCVD outperforms the ablative baseline "SimCVD w/o $\mathscr{L}_{\text{contrast}}$" by a large margin and achieves a +3.90% increase in Dice. It clearly demonstrates it can effectively capture global context structure and local cues of 3D shapes. Next, we study whether enforcing similarity constraints in the *feature* space is effective enough to exploit structured knowledge in practice. We find that, under the 10% labeled setting, removing $\mathscr{L}_{\text{pd}}$ leads to a performance drop, and the accuracy measures decreases by −0.92% and −1.45% in Dice and Jaccard, respectively. Our qualitative results confirm that *pair-wise distillation*

is effective for SimCVD in improving the network performance. When we remove $\mathscr{L}_{\text{sdm}}$, the network performance does not drop significantly. We speculate that our boundary-aware contrastive distillation framework is capable of eliciting "boundary-aware" knowledge from the teacher model with high accuracy. This further highlights the effectiveness of our proposed SimCVD. To further verify the robustness of SimCVD, we perform pair-wise t-test between SimCVD and the other methods, using the per-case test Dice score as samples. The null hypothesis states that the test scores come from the same distribution, and thus the methods do not differ; whereas the alternative hypothesis is that SimCVD yields higher Dice score. For all the alternative methods in the Table below, the p-values are close to or less than 0.05. Small p-values indicates that we are confident in rejecting the null hypothesis and conclude that SimCVD indeed outperforms the baseline models.

## B. Analysis on Projection Head

To further understand how different aspects of our projection head contribute to the superior model performance, we conduct extensive experiments and discuss our findings below.

**1) How to Interpret Dropout?:** Our experimental results have shown that SimCVD is an effective approach. In the following, we aim to answer two questions. First, how can we interpret SimCVD's *dropout* training strategy? Can we view *dropout* as a form of data augmentation? Second, is it capable of exploiting additional informative cues in practice?

First, we examine whether removing *dropout* during training can achieve comparable performance. Table IV shows the ablation result of our *dropout* on LA. As shown in Table IV, we observe that using *dropout* achieves a much better result on LA dataset. Compared to the setting $p = 0.1$, we find that "no dropout" ($p = 0$) leads to a dramatic performance degradation by −1.34%,−2.11%,−1.71,−2.69 absolute differences in terms of Dice, Jaccard, ASD, and 95HD, respectively. While in the case of $p = 0.5$, it also significantly hurts the network performance. On the other hand, we observe slight improvements on the other $p$ settings, compared to "no dropout", but eventually underperform SimCVD. This clearly demonstrates the superiority of our *dropout* strategy to learn better representations with respect to different pairs of augmented images. We speculate that adding *dropout* can be interpreted as a minimal form of data augmentation, in which the positive pair takes two views of the same images, and their representations make a clear difference in dropout masks.

**2) Effect of Augmentation Techniques:** To further examine our hypothesis, we compare common data augmentation techniques (*i.e.*, local shuffle pixel, non-linear transformation, in-painting, out-painting) in Table V. As is shown, the quantitative results reveal interesting behavior of different data augmentation: adding more data augmentation does not further contribute to the good model performance. We note that, somewhat surprisingly, it hurts the final prediction performance, and none of them outperforms the basic *dropout* mask. This suggests that by including these data augmentation techniques, it is possible to introduce additional noise during training, which leads to the representation collapse.

**3) Effect of Pooling Size:** In Table III, we demonstrate the network improvements from using adaptive mean pooling instead of adaptive max pooling. We investigate the effects of different pooling sizes in Table IV. Empirically, we observe that using a larger pooling size clearly improves performance consistently. However, we find that the results can not be improved further by increasing the pooling size to 256. In our implementation, we set the pooling size as 128.

## VII. Conclusion

In this work, we propose SimCVD, a simple contrastive distillation learning framework, which largely advances state-of-the-art voxel-wise representation learning on medical segmentation tasks. Specifically, we present an unsupervised training strategy, which takes two views of an input volume and predicts their signed distance maps of their object boundaries in a contrastive objective, with only two different dropout masks. We further conduct extensive analyses to understand the state-of-the-art performance of our approach, and demonstrate the importance of learning distinct boundary-aware representations and using dropout as the minimal data augmentation technique. We also propose to perform structural distillation by distilling pair-wise similarities, which achieves good performance improvements. Our experimental results show that SimCVD obtained new state-of-the-art results on two benchmarks in an extreme few-annotation setting.

We believe that our unsupervised training framework provides a new perspective on data augmentation along with unlabeled 3D medical data. We also plan to extend our method to solve multi-class medical image segmentation tasks.

## References

[1]. Staib LH and Duncan JS, “Model-based deformable surface finding for medical images,” IEEE Trans. Med. Imag, vol. 15, no. 5, pp. 720–731, Oct. 1996.

[2]. Yang J, Staib LH, and Duncan JS, “Neighbor-constrained segmentation with level set based 3-D deformable models,” IEEE Trans. Med. Imag, vol. 23, no. 8, pp. 940–948, Aug. 2004.

[3]. Yang J and Duncan JS, “3D image segmentation of deformable objects with joint shape-intensity prior models using level sets,” Med. Image Anal, vol. 8, no. 3, pp. 285–294, Sep. 2004. [PubMed: 15450223]

[4]. Chakraborty A, Staib LH, and Duncan JS, “Deformable boundary finding in medical images by integrating gradient and region information,” IEEE Trans. Med. Imag, vol. 15, no. 6, pp. 859–870, Dec. 1996.

[5]. Staib LH and Duncan JS, “Boundary finding with parametrically deformable models,” IEEE Trans. Pattern Anal. Mach. Intell, vol. 14, no. 11, pp. 1061–1075, Nov. 1992.

[6]. Ronneberger O, Fischer P, and Brox T, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, Cham, Switzerland: Springer, 2015, pp. 234–241.

[7]. Milletari F, Navab N, and Ahmadi S-A, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 4th Int. Conf. 3D Vis. (3DV), Oct. 2016, pp. 565–571.

[8]. Bai W et al., “Semi-supervised learning for network-based cardiac MR image segmentation,” in Proc. MICCAI, Cham, Switzerland: Springer, 2017, pp. 253–260.

[9]. Ganaye P-A, Sdika M, and Benoit-Cattin H, “Semi-supervised learning for segmentation under semantic constraint,” in Proc. MICCAI, Cham, Switzerland: Springer, 2018, pp. 595–602.

[10]. You C, Yang J, Chapiro J, and Duncan JS, “Unsupervised Wasserstein distance guided domain adaptation for 3D multi-domain liver segmentation,” in Interpretable and Annotation-Efficient Learning for Medical Image Computing. Cham, Switzerland: Springer, 2020, pp. 155–163.

[11]. Wang Y et al. , “Deep distance transform for tubular structure segmentation in CT scans,” in Proc. CVPR, Jun. 2020, pp. 3833–3842.

[12]. Xue Y et al. , “Shape-aware organ segmentation by predicting signed distance maps,” in Proc. AAAI, 2020, pp. 12565–12572.

[13]. Li X, Yu L, Chen H, Fu C-W, Xing L, and Heng P-A, “Transformation-consistent self-ensembling model for semisupervised medical image segmentation,” IEEE Trans. Neural Netw. Learn. Syst, vol. 32, no. 2, pp. 523–534, Feb. 2021 [PubMed: 32479407]

[14]. Laine S and Aila T, “Temporal ensembling for semi-supervised learning,” 2016, arXiv:1610.02242

[15]. Zhang Y, Yang L, Chen J, Fredericksen M, Hughes DP, and Chen DZ, “Deep adversarial networks for biomedical image segmentation utilizing unannotated images,” in Proc. MICCAI, Springer, 2017, pp. 408–416.

[16]. Li X, Yu L, Chen H, Fu C-W, and Heng P-A, “Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model,” 2018, arXiv:1808.03887.

[17]. Nie D, Gao Y, Wang L, and Shen D, “ASDNet: Attention based semi-supervised deep networks for medical image segmentation,” in Proc. MICCAI, Springer, 2018, pp. 370–378.

[18]. Qiao S, Shen W, Zhang Z, Wang B, and Yuille A, “Deep co-training for semi-supervised image recognition,” in Proc. ECCV, Sep. 2018, pp. 135–152.

[19]. Bortsova G, Dubost F, Hogeweg L, Katramados I, and de Bruijne M, “Semi-supervised medical image segmentation via learning consistency under transformations,” in Proc. MICCAI, Springer, 2019, pp. 810–818.

[20]. Yu L, Wang S, Li X, Fu C-W, and Heng P-A, “Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation,” in Proc. MICCAI, Springer, 2019, pp. 605–613.

[21]. Li S, Zhang C, and He X, “Shape-aware semi-supervised 3D semantic segmentation for medical images,” in Proc. MICCAI, Springer, 2020, pp. 552–561.

[22]. Zhu J, Li Y, Hu Y, Ma K, Zhou SK, and Zheng Y, “Rubik’s cube+: A self-supervised feature learning framework for 3D medical image analysis,” Med. Image Anal, vol. 64, Aug. 2020, Art. no. 101746. [PubMed: 32544840]

[23]. Luo X, Chen J, Song T, and Wang G, “Semi-supervised medical image segmentation through dual-task consistency,” in Proc. AAAI, 2020 pp. 1–9.

[24]. Chaitanya K, Erdil E, Karani N, and Konukoglu E, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” in Proc. NeurIPS, 2020, pp. 12546–12558.

[25]. You C, Zhao R, Staib L, and Duncan JS, “Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation,” 2021, arXiv:2105.07059.

[26]. Chen T, Kornblith S, Norouzi M, and Hinton G, “A simple framework for contrastive learning of visual representations,” in Proc. ICML, 2020, pp. 1597–1607.

[27]. Hjelm RD et al. , “Learning deep representations by mutual information estimation and maximization,” in Proc. ICLR, 2019, pp. 1–24.

[28]. Bai W et al., “Self-supervised learning for cardiac MR image segmentation by anatomical position prediction,” in Proc. MICCAI, Springer, 2019, pp. 541–549.

[29]. Peng J, Wang P, Desrosiers C, and Pedersoli M, “Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels,” in Proc. NeurIPS, 2021.

[30]. Yang C, Xie L, Su C, and Yuille AL, “Snapshot distillation: Teacher-student optimization in one generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2859–2868.

[31]. Zhuang J, Cai J, Wang R, Zhang J, and Zheng W-S, “Deep KNN for medical image classification,” in Proc. MICCAI, Springer, 2020, pp. 127–136.

[32]. Tarvainen A and Valpola H, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proc. NeurIPS, 2017, pp. 1195–1204.

[33]. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res, vol. 15, pp. 1929–1958, Jun. 2014.

[34]. Perera S, Barnes N, He X, Izadi S, Kohli P, and Glocker B, “Motion segmentation of truncated signed distance function based volumetric surfaces,” in Proc. IEEE Winter Conf. Appl. Comput. Vis, Jan. 2015, pp. 1046–1053.

[35]. Dangi S, Linte CA, and Yaniv Z, “A distance map regularized CNN for cardiac cine MR image segmentation,” Med. Phys, vol. 46, no. 12, pp. 5637–5651, Dec. 2019. [PubMed: 31598971]

[36]. Park JJ, Florence P, Straub J, Newcombe R, and Lovegrove S, “DeepSDF: Learning continuous signed distance functions for shape representation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 165–174.

[37]. Hinton G, Vinyals O, and Dean J, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.

[38]. Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, and Bengio Y, “FitNets: Hints for thin deep nets,” 2014, arXiv:1412.6550.

[39]. Liu Y, Chen K, Liu C, Qin Z, Luo Z, and Wang J, “Structured knowledge distillation for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2604–2613.

[40]. Yu F et al. , “Annotation-free cardiac vessel segmentation via knowledge transfer from retinal images,” in Proc. MICCAI, 2019, pp. 714–722.

[41]. Li K, Wang S, Yu L, and Heng PA, “Dual-teacher++: Exploiting intra-domain and inter-domain knowledge with reliable transfer for cardiac segmentation,” IEEE Trans. Med. Imag, vol. 40, no. 10 pp. 2771–2782, Oct. 2021.

[42]. He Y et al., “DPA-DenseBiasNet: Semi-supervised 3D fine renal artery segmentation with dense biased network and deep priori anatomy,” in Proc. MICCAI, Springer, 2019, pp. 139–147.

[43]. Zhou Z et al., “Models genesis: Generic autodidactic models for 3D medical image analysis,” in Proc. MICCAI, Springer, 2019, pp. 384–393.

[44]. Yang L et al. , “NuSeT: A deep learning tool for reliably separating and analyzing crowded cells,” PLOS Comput. Biol, vol. 16, no. 9, Sep. 2020, Art. no. e1008193. [PubMed: 32925919]

[45]. Zheng H et al., “Semi-supervised segmentation of liver using adversarial learning with deep atlas prior,” in Proc. MICCAI, Cham, Switzerland: Springer, 2019, pp. 148–156.

[46]. Zhuang X, Li Y, Hu Y, Ma K, Yang Y, and Zheng Y, “Self-supervised feature learning for 3D medical images by playing a Rubik’s cube,” in Proc. MICCAI, Cham, Switzerland: Springer, 2019, pp. 420–428.

[47]. Taleb A et al. , “3D self-supervised methods for medical imaging,” in Proc. NeurIPS, 2020, pp. 18158–18172.

[48]. Hadsell R, Chopra S, and LeCun Y, “Dimensionality reduction by learning an invariant mapping,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 2, Jun. 2006, pp. 1735–1742

[49]. Doersch C, Gupta A, and Efros AA, “Unsupervised visual representation learning by context prediction,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1422–1430.

[50]. Noroozi M and Favaro P, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in Proc. ECCV, Springer, 2016, pp. 69–84

[51]. Wu Z, Xiong Y, Yu SX, and Lin D, “Unsupervised feature learning via non-parametric instance discrimination,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, Jun. 2018, pp. 3733–3742.

[52]. Tian Y, Krishnan D, and Isola P, “Contrastive multiview coding,” 2019 arXiv:1906.05849

[53]. Misra I and van der Maaten L, “Self-supervised learning of pretext-invariant representations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6707–6717.

[54]. Federici M, Dutta A, Forré P, Kushman N, and Akata Z, “Learning robust representations via multi-view information bottleneck,” in Proc. ICLR, 2020, pp. 1–26.

[55]. Chen N, You C, and Zou Y, “Self-supervised dialogue learning for spoken conversational question answering,” in Proc. Interspeech, Aug. 2021, pp. 1–5.

[56]. You C, Chen N, and Zou Y, “Self-supervised contrastive cross-modality representation learning for spoken question answering,” 2021 arXiv:2109.03381

[57]. van den Oord A, Li Y, and Vinyals O, “Representation learning with contrastive predictive coding,” 2018, arXiv:1807.03748.

[58]. Wang X and Gupta A, “Unsupervised learning of visual representations using videos,’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2794–2802.

[59]. You C, Chen N, Liu F, Yang D, and Zou Y, “Towards data distillation for end-to-end spoken conversational question answering,” 2020, arXiv:2010.08923.

[60]. You C, Chen N, and Zou Y, “Contextualized attention-based knowledge transfer for spoken conversational question answering,” in Proc. Interspeech, Aug. 2021, pp. 1–5.

[61]. You C, Chen N, and Zou Y, “MRD-Net: Multi-modal residual knowledge distillation for spoken question answering,” in Proc. 13th Int. Joint Conf. Artif. Intell, Aug. 2021, pp. 1–7.

[62]. You C, Chen N, and Zou Y, “Knowledge distillation for improved accuracy in spoken question answering,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2021, pp. 7793–7797.

[63]. Furlanello T, Lipton Z, Tschannen M, Itti L, and Anandkumar A, “Born again neural networks,” in Proc. ICML, 2018, pp. 1607–1616.

[64]. Yang C, Xie L, Qiao S, and Yuille A, “Knowledge distillation in generations: More tolerant teachers educate better students,” 2018, arXiv:1805.05551.

[65]. Li Y, Yang J, Song Y, Cao L, Luo J, and Li L-J, “Learning from noisy labels with distillation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1910–1918.

[66]. Xie J, Shuai B, Hu J-F, Lin J, and Zheng W-S, “Improving fast segmentation with teacher-student learning,” 2018, arXiv:1810.08476.

[67]. Zhang Y, Xiang T, Hospedales TM, and Lu H, “Deep mutual learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, Jun. 2018, pp. 4320–4328.

[68]. Chen X, Yuan Y, Zeng G, and Wang J, “Semi-supervised semantic segmentation with cross pseudo supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2613–2622.

[69]. Vu T-H, Jain H, Bucher M, Cord M, and Perez P, “ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2517–2526.

[70]. Verma V, Lamb A, Kannala J, Bengio Y, and Lopez-Paz D, “Interpolation consistency training for semi-supervised learning,” in Proc. 28 th Int. Joint Conf. Artif. Intell, Aug. 2019, pp. 3635–3641.

[71]. Klambauer G, Unterthiner T, Mayr A, and Hochreiter S, “Self-normalizing neural networks,” in Proc. NeurIPS, 2017, pp. 972–981.

[72]. Roth HR, Farag A, Turkbey E, Lu L, Liu J, and Summers RM, “Data from pancreas-ct. the cancer imaging archive,” Tech. Rep, 2016

[73]. Zhou Y et al. , “Prior-aware neural network for partially-supervised multi-organ segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 10672–10681.

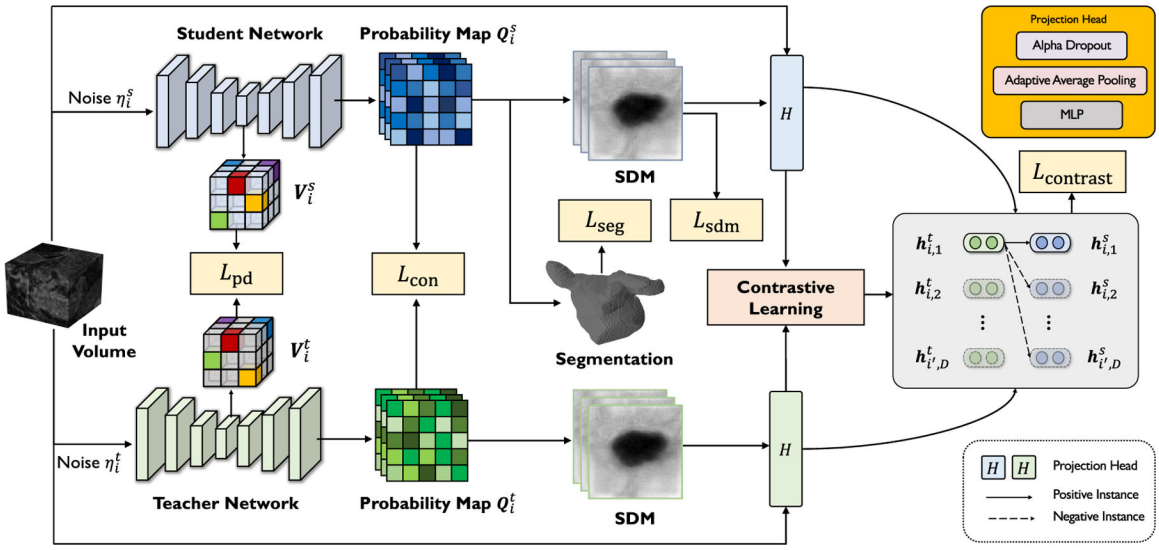


**Fig. 1.**

Overview of our SimCVD in training. Given a 3D input volume, our SimCVD jointly predicts the 3D probability maps and the SDMs of the object using a student network and a teacher network. The student network is trained by stochastic gradient descent using two supervised losses ($\mathscr{L}_{\text{seg}}, \mathscr{L}_{\text{sdm}}$) and three unsupervised terms ($\mathscr{L}_{\text{contrast}}, \mathscr{L}_{\text{pd}}, \mathscr{L}_{\text{con}}$). Specifically, $\mathscr{L}_{\text{contrast}}$ is designed to distill "boundary-aware" knowledge by contrastive learning in the shared latent space, and $\mathscr{L}_{\text{pd}}$ is designed to exploit structural relationships among location-paired voxel representations from the encoders. The teacher network's weights are updated with a "momentum update" (exponential moving average) of the student network's weights.

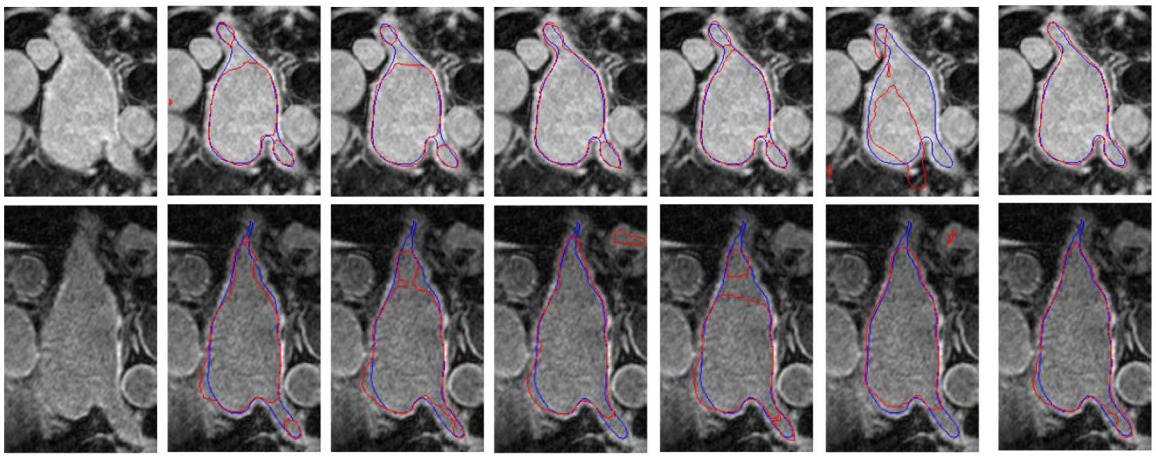


**Fig. 2.**

Visual comparisons with other methods on LA dataset. As observed, SimCVD achieves superior performance with more accurate borders and shapes. We train all the evaluated methods in the setting of 8 annotated images. Red and blue denote the predictions and ground truths, respectively.

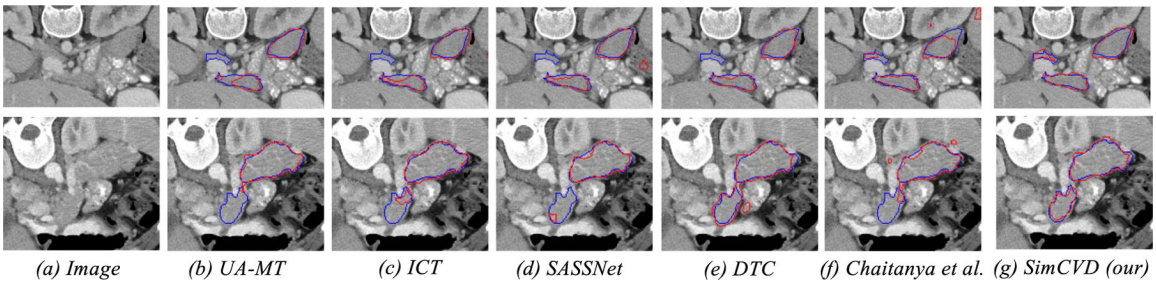


**Fig. 3.**

Visual comparisons with other methods on the pancreas dataset. We train all the evaluated methods in the setting of 12 annotated images. Red and blue denotes the predictions and ground truths, respectively.

**TABLE I**

Quantitative Segmentation Results on the LA Dataset. The Backbone Network of All Evaluated Methods Are V-Net

| Method | # scans used | | Metrics | | | |
|---|---|---|---|---|---|---|
| | Labeled | Unlabeled | Dice[%] | Jaccard[%] | ASD[voxel] | 95HD[voxel] |
| V-Net [7] | 80 | 0 | 91.14 | 83.82 | 1.52 | 5.75 |
| V-Net | 16 | 0 | 86.03 | 76.06 | 3.51 | 14.26 |
| DAN [15] | 16 | 64 | 87.52 | 78.29 | 2.42 | 9.01 |
| CPS [68] | 16 | 64 | 87.87 | 78.61 | 2.16 | 12.87 |
| MT [32] | 16 | 64 | 88.42 | 79.45 | 2.73 | 13.07 |
| Entropy Mini [69] | 16 | 64 | 88.45 | 79.51 | 3.72 | 14.14 |
| UA-MT [20] | 16 | 64 | 88.88 | 80.21 | 2.26 | 7.32 |
| ICT [70] | 16 | 64 | 89.02 | 80.34 | 1.97 | 10.38 |
| SASSNet [21] | 16 | 64 | 89.27 | 80.82 | 3.13 | 8.83 |
| DTC [23] | 16 | 64 | 89.42 | 80.98 | 2.10 | 7.32 |
| Chaitanya *et al.* [24] | 16 | 64 | 89.94 | 81.82 | 2.66 | 7.23 |
| SimCVD (ours) | 16 | 64 | **90.85** | **83.80** | **1.86** | **6.03** |
| V-Net [7] | 8 | 0 | 79.99 | 68.12 | 5.48 | 21.11 |
| DAN [15] | 8 | 72 | 75.11 | 63.47 | 3.57 | 19.04 |
| CPS [68] | 8 | 72 | 84.09 | 73.17 | 2.41 | 22.55 |
| MT [32] | 8 | 72 | 84.24 | 73.26 | 2.71 | 19.41 |
| Entropy Mini [69] | 8 | 72 | 85.90 | 75.60 | 2.74 | 18.65 |
| UA-MT [20] | 8 | 72 | 84.25 | 73.48 | 3.36 | 13.84 |
| ICT [70] | 8 | 72 | 85.39 | 74.84 | 2.88 | 17.45 |
| SASSNet [21] | 8 | 72 | 86.81 | 76.92 | 3.94 | 12.54 |
| DTC [23] | 8 | 72 | 87.49 | 78.03 | 2.37 | 9.06 |
| Chaitanya *et al.* [24] | 8 | 72 | 84.95 | 74.77 | 3.70 | 10.68 |
| SimCVD (ours) | 8 | 72 | **89.03** | **80.34** | **2.59** | **8.34** |

**TABLE II**

Quantitative Segmentation Results on the Pancreas Dataset. The Backbone Network of All Evaluated Methods Are V-Net

| Method | # scans used | | Metrics | | | |
|---|---|---|---|---|---|---|
| | Labeled | Unlabeled | Dice[%] | Jaccard[%] | ASD[voxel] | 95HD[voxel] |
| V-Net [7] | 62 | 0 | 77.84 | 64.78 | 3.73 | 8.92 |
| V-Net | 12 | 0 | 62.42 | 48.06 | 4.77 | 22.34 |
| MT [32] | 12 | 50 | 71.29 | 56.69 | 2.82 | 16.31 |
| DAN [15] | 12 | 50 | 68.67 | 53.97 | 3.07 | 15.78 |
| CPS [68] | 12 | 50 | 69.28 | 54.02 | 3.07 | 15.34 |
| Entropy Mini [69] | 12 | 50 | 69.33 | 54.32 | 2.92 | 15.29 |
| UA-MT [20] | 12 | 50 | 72.43 | 57.91 | 4.25 | 11.01 |
| ICT [70] | 12 | 50 | 70.06 | 55.66 | 2.98 | 13.05 |
| SASSNet [21] | 12 | 50 | 70.47 | 55.74 | 4.26 | 10.95 |
| DTC [23] | 12 | 50 | 74.07 | 60.17 | 2.61 | 10.35 |
| Chaitanya *et al.* [24] | 12 | 50 | 70.79 | 55.76 | 6.08 | 15.35 |
| SimCVD (ours) | 12 | 50 | **75.39** | **61.56** | **2.33** | **9.84** |

*IEEE Trans Med Imaging*. Author manuscript; available in PMC 2023 July 06.

**TABLE III**

Ablation on (A) Model Component: SimCVD w/o SDM; SimCVD w/o Adaptive Max Pooling; (B) Loss Formulation: SimCVD w/o $\mathscr{L}_{\text{CONTRAST}}$; SimCVD w/o $\mathscr{L}_{\text{pd}}$; SimCVD w/o $\mathscr{L}_{\text{SDM}}$, Compared to the Baseline and Our Proposed SimCVD

| | Method | Metrics | | | | p-value (vs. SimCVD, [%]) |
|---|---|---|---|---|---|---|
| | | Dice[%] | Jaccard[%] | ASD[voxel] | 95HD[voxel] | |
| | Baseline (UA-MT) | 84.24 | 73.26 | 2.71 | 1941 | 0.019 |
| (a) | SimCVD w/o SDM | 88.24 | 79.07 | 4.19 | 11.43 | 2.47 |
| | SimCVD w/ Adaptive Max Pooling | 88.32 | 79.26 | 3.79 | 14.69 | 0.33 |
| (b) | SimCVD w/o $\mathscr{L}_{\text{contrast}} + \mathscr{L}_{\text{sdm}}$ | 84.97 | 74.49 | 6.13 | 19.98 | 0.018 |
| | SimCVD w/o $\mathscr{L}_{\text{contrast}}$ | 85.13 | 74.57 | 5.97 | 16.61 | 2.9e-6 |
| | SimCVD w/o $\mathscr{L}_{\text{pd}}$ | 88.11 | 78.89 | 2.89 | 12.58 | 2.02 |
| | SimCVD w/o $\mathscr{L}_{\text{sdm}}$ | 88.85 | 80.03 | 2.71 | 9.02 | 5.14 |
| | SimCVD | **89.03** | **80.34** | **2.59** | **8.34** | - |

**TABLE IV**

Ablation on Dropout Rates $p$ and Pooling Size

| | Method | Metrics | | | |
|---|---|---|---|---|---|
| | | Dice[%] | Jaccard[%] | ASD[voxel] | 95HD[voxel] |
| Dropout | $p = 0.0$ | 87.69 | 78.23 | 2.49 | 11.03 |
| | $p = 0.01$ | 87.98 | 78.69 | 3.08 | 11.17 |
| | $p = 0.02$ | 87.99 | 78.70 | 2.60 | 9.03 |
| | $p = 0.05$ | 88.01 | 78.71 | 2.78 | 10.60 |
| | $p = 0.1$ | 89.03 | 80.34 | 2.59 | 8.34 |
| | $p = 0.2$ | 88.10 | 78.86 | 3.28 | 12.70 |
| | $p = 0.5$ | 86.67 | 76.59 | 4.20 | 14.89 |
| Pooling Size | 16 × 16 | 87.04 | 77.22 | 3.69 | 14.28 |
| | 32 × 32 | 87.53 | 77.98 | 3.13 | 11.56 |
| | 64 × 64 | 88.37 | 79.27 | 2.75 | 8.84 |
| | 128 × 128 | 89.03 | 80.34 | 2.59 | 8.34 |
| | 256 × 256 | 87.99 | 78.71 | 2.82 | 9.97 |

**TABLE V**

Ablation on Different Data Augmentations on the LA Dataset. All of Them Include the Dropout Masks ($p = 0.1$)

| Method | Metrics | | | |
|---|---|---|---|---|
| | Dice[%] | Jaccard[%] | ASD[voxel] | 95HD[voxel] |
| SimCVD | 89.03 | 80.34 | 2.59 | 8.34 |
| + Local Shuffle Purel | 88.15 | 78.97 | 1.96 | 8.66 |
| + Non-linear Intensity Transformation | 88.02 | 78.80 | 2.68 | 10.29 |
| + In-painting | 88.37 | 79.26 | 2.84 | 10.97 |
| + Out-painting | 88.24 | 79.07 | 2.58 | 10.62 |