Weakly-Supervised 3D Medical Image Segmentation Using Geometric Prior and Contrastive Similarity

Hao Du, Qihua Dong, Yan Xu, and Jing Liao, *Member, IEEE*

Abstract—Medical image segmentation is almost the most important pre-processing procedure in computer-aided diagnosis but is also a very challenging task due to the complex shapes of segments and various artifacts caused by medical imaging, (*i.e.*, low-contrast tissues, and non-homogenous textures). In this paper, we propose a simple yet effective segmentation framework that incorporates the geometric prior and contrastive similarity into the weakly-supervised segmentation framework in a loss-based fashion. The proposed geometric prior built on point cloud provides meticulous geometry to the weakly-supervised segmentation proposal, which serves as better supervision than the inherent property of the bounding-box annotation (*i.e.*, height and width). Furthermore, we propose the contrastive similarity to encourage organ pixels to gather around in the contrastive embedding space, which helps better distinguish low-contrast tissues. The proposed contrastive embedding space can make up for the poor representation of the conventionally-used gray space. Extensive experiments are conducted to verify the effectiveness and the robustness of the proposed weakly-supervised segmentation framework. The proposed framework are superior to state-of-the-art weakly-supervised methods on the following publicly accessible datasets: LiTS 2017 Challenge, KiTS 2021 Challenge and LPBA40. We also dissect our method and evaluate the performance of each component.

Index Terms—Weakly-supervised segmentation, medical image segmentation, contrastive similarity, geometric prior, point cloud.
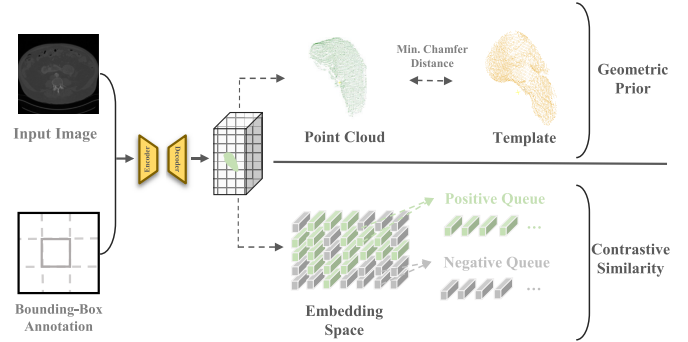
I. Introduction

SEGMENTATION is of fundamental importance for the understanding and interpretation of medical images, as it

Fig. 1. The illustration of the proposed weakly-supervised segmentation framework. As shown in the figure, we propose geometric prior and contrastive similarity for weakly-supervised segmentation. The top row indicates the geometric prior of our method. We first convert the conventionally-used volume representation to point cloud representation and register the template organ to the predicted organ. Then we minimize their Chamfer Distance. The bottom row explains the core idea of the proposed contrastive similarity. By dividing pixels into positive and negative pixels, we encourage organ pixels to gather around in the embedding space to better segment the low-contrast organ.

is essential for the diagnostic, treatment, and follow-up rehabilitation of various diseases. This task has been widely studied with the recent advent of deep convolutional neural networks (CNNs) [1], [2]. Nevertheless, there exists the main limitation that their methods require a large number of training images with pixel-wise annotations. The extremely high cost of collecting and annotating these training images largely hampers the performance and limits the scalability of deep CNNs in the medical domain. A popular paradigm to alleviate the need for pixel-wise annotations is the weakly-supervised segmentation with bounding-box annotations [3], [4], [5], [6], [7]. They employ bounding-box annotations to generate proposals, which are fake labels and thereby mimic full supervision.

Nevertheless, despite the good performances achieved by these works in certain practical scenarios, their applicability might be limited for two reasons: 1) *complex shapes*: in medical image segmentations, some organs have delicate architectures , *i.e.*, intra-kidney variabilities, which are difficult to be precisely segmented without pixel-wise supervision; 2) *imaging artifacts*: as discussed in previous works [8], [9], [10], various medical imaging artifacts caused by technical or physical problems make low-contrast tissues and

non-homogenous textures hard to distinguish, especially in the conventionally widely-used gray space. The complex shapes and imaging artifacts largely limit the applicability of the weakly-supervised segmentation models in many scenarios, especially when segmenting complex structures.

To conquer the challenge of complex organ shapes, especially for the radiological cases (*e.g.*, MRI, CT) with structural imaging, we propose to learn the geometric prior of the organ by a standard organ template. Instead of using volume representation, we first leverage the gridding reverse [11] to convert the segmentation result from volume representation to point cloud representation and then compare it with the template in the point cloud space. The basic unit in point cloud representation is much more fine-grained and flexible than the volume representation (*, i.e.*, flexible point *v.s.* uniform voxel grids), which helps better describe delicate geometric structures. On the other hand, unlike the conventionally-used gray space [8], we leverage the contrastive learning [12] to encode the pixels to high-dimensional embedding space and encourage pixels of the same labels to gather around. This helps alleviate the imaging artifacts for richer expressivity in the embedding space compared to the gray space.

In this paper, we present a novel weakly-supervised segmentation framework, which makes the earliest effort to incorporate geometric prior and contrastive similarity. And the framework is general as well that can be easily applied to improve multiple weakly-supervised segmentation models with bounding-box annotations, *i.e.*, Ai+L [13], BoxInst [7]. By learning geometry prior from the given template and distinguishing low-contrast tissues by the contrastive similarity, our method can generate high-quality results with bounding box supervision only.

Our method consists of two major components. In the geometric prior component, the shapes of proposals are constrained by a given template represented by point cloud. Both the external boundaries and internal structures of the proposal will be optimized by minimizing the distance according to a given template. The second component is the contrastive similarity, addressing the issues raised by medical imaging artifacts. By pre-training a contrastive head, we successfully learn the difference between organ pixels and non-organ pixels. This component can better distinguish low-contrast tissues and non-homogenous texture than conventionally widely-used gray space. Through extensive experiments, we demonstrate that our method can generate a high-quality segments, along with delicate internal details and accurate boundaries. We show that our method outperforms other bounding-box weakly-supervised methods [5], [7] under similar settings. We also conduct extensive experiments to verify the effectiveness of components in our method.

In summary, our major contributions are three folds.

- We propose a simple yet effective weakly-supervised segmentation framework with bounding-box annotations, which can be easily applied to many weakly-supervised segmentation models and improve their performances.
- We propose the geometric prior in point cloud representation to better guide the learning of shapes, especially for those organs with complex structures.

- The proposed contrastive similarity makes up for the poor representation of the conventional gray space and thus can better distinguish tissues with medical imaging artifacts.

Our code and data will be made publicly available for further research.

II. RELATED WORK

In this section, we first review existing weakly-supervised medical segmentation methods with bounding-box annotations in both natural and medical image segmentation, then we discuss recent works with geometric prior and finally present the trends in contrastive similarity.

A. Weakly-Supervised Medical Semantic Segmentation

Generally, methods in weakly-supervised segmentation are classified into four categories by the type of their weak annotations: scribbles [14], points [15], [16], image-level tags [17], [18] and bounding-box annotations [5]. Scribbles and points supervision at least label one scribble or point for each region, and the annotated areas will be directly incorporated into the calculation of segmentation loss. Wang et al. [19] propose to leverage a random walker algorithm [20] to generate initial proposals for the unlabeled regions and then supervise the training of segmentation models by the initial segments. Qu *et al.* [16] uses a similar training pipeline but a different label generation method for label generation which combines K-means clustering and Voronoi partition diagram. Xu et al. [18] enrich the image-level labels to instance-level labels by multiple instance learning (MIL) and segment images using only volume-level labels.

Weakly-supervised segmentation with bounding box annotations earns increasing interest in medical image segmentation for its simplicity and low-annotation cost. We can define the bounding boxes with two corner coordinates that are easy to store in real scenarios. In addition, the bounding box annotations are location-aware so that they provide the spatial relationship of the target object, which is a popular direction in recent researches [21], [22], [23], [24]. In the early stages, researchers [22], [25] propose to consider pixels within the bounding box as foreground pixels and train the segmentation framework by these noisy labels. Despite the good performance achieved by such a scheme, they may accumulate errors during the alternative generation process. Most recently, researchers [7] tried to directly generate the segmentation result instead of the error-prone alternative way. Generally, they build a *mask head* to produce the segmentation result, and the bounding box annotation is employed to train this mask head. In this work, we follow this segmentation scheme where the segmentation result is directly generated by the mask head. Furthermore, to address the fore-mentioned complex shapes and imaging artifacts we propose geometric prior and contrastive similarity, respectively.

B. Geometric Prior

Different from natural images, there exists obvious anatomical prior (*, i.e.*, atlas prior) in medical images, specifically in

organs of human bodies (*i.e.*, shape and position). Existing works incorporating such anatomical prior mostly fall into two categories: loss-based methods and graph-based methods. Generally, graph-based methods [26], [27], [28], [29], [30] leverage the probability maps of occurring anatomy chances to construct graph models and estimate the foreground probability from the input image gray space. An appearance model of basic forms is employed to improve the segmentation accuracy [29]. Gao et al [30] proposes to apply an appearance ConvNet to characterize the foreground. Despite the high accuracy achieved by these methods, the graphical models bring heavy and expensive computational burdens to the segmentation framework, which makes it infeasible in certain scenarios.

Another popular direction in combining the prior with the segmentation framework is loss-based methods. Researchers [31], [32], [33], [34], [35], [36] mostly minimize the distance between the segmentation network output and the pre-defined anatomical priors. Several works [31], [32] pose regularization terms on the training objective (, *i.e.*, anatomical adjacency or boundary conditions). Distance between predictions and atlas prior are also calculated in latent feature space [35], [36].

Compared with graph-based methods, loss-based methods provide a versatile fashion to incorporate anatomical priors with a wider range of scales while maintaining the computational efficiency of the segmentation framework. However, previous loss-based methods fail to address the aforementioned two issues for two aspects: 1) Previous works generally slice the volume into 2D or 3D patches, which are then processed sequentially to save memory cost. However, such a partitioning method breaks the global geometric relationships, resulting in inferior segmentation performance. Different from them, we learn the geometric prior in 3D embedding space to capture the overall geometry and proposed completeness head to ensure the shape completeness of the proposal. 2) Unlike previous works using a volume representation, we leverage the Gridding Reverse [11] to convert the volume representation to point cloud representation. Compared to the volume representation constrained by uniform voxel grids, point cloud without grids is more flexible in representing delicate structures.

C. Contrastive Learning

Contrastive Learning aims to attract the positive and reverse the negative by dividing the feature space into positive and negative data pairs. As for semantic segmentation, it has been mainly used to learn robust pixel-wise embeddings by the following three class labels [37], [38], [39], [40], [41], [42], [43], [44], [45], [46].: ground truths, pseudo labels, and a mix of both ground truths and pseudo labels. 1) by GTs: Zhao et al. [42] leverage the contrastive learning to improve the intra-class compactness and inter-class separability of the segmentation framework; 2) by pseudo labels: Chaitanya et al. [41] propose a local contrastive loss to learn good pixel-level features from pseudo labels. 3) by a mix of GTs and pseudo labels: Alonso et al. [40] propose to utilize the mix of GTs and pseudo labels to enhance the feature embed-

ding level. Some researchers leverage contrastive learning to address the time-consuming pixel-wise labeling in medical image segmentation. Chaitanya et al. [12] propose a two-stage self-supervised contrastive learning framework to learn the feature matching both in global and local mechanisms from unlabeled data in the pre-training stage. Hu et al. [47] proposes a semi-supervised scheme to learn self-supervised global contrast and supervised local contrast. In our work, we observe that the conventionally-used gray space is not enough to distinguish positive and negative pixels (, *i.e.*, organ pixels and non-organ pixels), especially in Magnetic Resonance Imaging. We thus leverage contrastive learning to calculate the contrastive similarity between pixels. By encoding pixels to high-dimensional features and encouraging pixels of the same label to gather around in the embedding space, it enhances the discriminability and thus alleviates the poor performance of the gray space in handling medical imaging artifacts, *i.e.*, artifacts in ultrasound imaging and similar surrounding tissues.

III. METHOD

A. Overall Framework

Given an input image $I \in \mathbb{R}^{S \times H \times W}$ (S indicates the slice number, H indicates height and W represents width) and its corresponding bounding-box annotation $\mathbb{B}^{1 \times 6}$ (constrained by its upper left coordinates and bottom right coordinates), our weakly-supervised framework $\mathbf{F}(\cdot)$ obtains the pixel-wise segmentation mask $\mathcal{M} = \mathbf{F}(I)$ and the training goal is to minimize the loss function L_{frame}:

$$\min_{\mathbf{F}} L_{\text{frame}}(I, \mathbb{B}, \mathbf{F}) \tag{1}$$

Pipeline Following nnUNet [2], we randomly sample an input patch $p \in \mathbb{R}^{S' \times H' \times W'}$ from the original input image and encode the patches by a ConvNet encoder E. Similar to [1], we adopt a multi-layer ConvNet as the decoder G to obtain the feature maps $P \in \mathbb{R}^{S' \times H' \times W'}$, where the decoder shares the same layer number as the encoder. As shown in Fig. 2, the proposed geometric prior and contrastive learning are incorporated in the training of the *mask head* to address the aforementioned complex shapes and imaging artifacts issues. The training loss L_{frame} is composed of two components: L_{ori} and L_{mask}

$$L_{\text{frame}} = L_{\text{ori}} + L_{\text{mask}} \tag{2}$$

where L_{ori} indicates the original training loss of a standard weakly-supervised framework, e.g., L_{ori} in BoxInst indicates the training loss L_{fcos} of the original FCOS [48] which comprises the training losses of *class head*, *box head*, and *centerness head*. L_{mask} stands for the training loss of the mask head. In the following paragraphs, we mainly discuss the training of the mask head. The training of the mask head can be formulated as Eq.(3).

$$L_{\text{mask}} = L_{\text{geo}} + L_{\text{cons}} \tag{3}$$

The mask head produces binary segmentation masks, which are further optimized by our proposed geometric prior L_{geo} and contrastive similarity L_{cons}. Multiple mask heads are

required for the multiple-class segmentation task. More specifically, for geometric prior loss, we build a *completeness head* to predict the completeness score for every proposal, indicating the conditional probability that the object is complete inside the input. Each complete proposal is converted into point cloud and registered with the point cloud of the template organ. The Chamfer Distance loss is applied to minimize their distance. In the aspect of the contrastive similarity loss, we build a *contrastive head* to obtain the contrastive similarity by the feature maps that the contrastive head assigns positive and negative labels to each position in the feature maps. In the following paragraphs, we first introduce the geometric prior. Then we elaborate on the technical details of the proposed contrastive similarity.

B. Geometric Prior

The proposed geometric prior is applied in 3D point cloud space for two reasons: 1) we observe that 2D slices cannot well-preserve the geometric continuity of 3D organs. Thus, we learn the geometric shape of the organ in 3D embedding space. 2) The conventionally-used volume representation cannot handle the segmentation of meticulous structures. The expressivity of the volume representation is largely limited by the uniform voxel grids. Instead, we leverage gridding reverse [11] to process the segmentation of complex shapes in point cloud embedding space.

As shown in Fig. 2, we propose the geometric prior to better weakly supervise the training of the mask head. The geometric prior refers to the template organ's boundary shape and internal distribution. More specifically, we introduce gridding reverse [11] to building a bridge between the volume representation and point cloud representation conversion. This helps us to get rid of the representation constraint in the volume representation. After converting both template organ $\mathbb{T}$ and proposal $\mathbb{S}$ into the point cloud representation, we then register the template organ to the proposal by a widely used ICP registration tool [49]. We finally minimize the geometric prior loss of the Chamfer Distance between the template organ and the proposal in the point cloud embedding space. Below we first introduce the conversion and registration of point cloud and then the definition of geometric prior loss.

1) Sparse Registration: To utilize the rich expressivity in point cloud representation, we introduce the gridding reverse [11] to help the transition between the volume representation and the point cloud representation. For each voxel grid, the gridding reverse calculates the weighted sum of the eight vertices of the corresponding grid and assigns the weighted sum to the coordinates of a new point. Unlike uniform voxel grids, the high flexibility of points' coordinates enables the point cloud representation to describe meticulous and complex architectures. This helps better learn the difficult intra-organ variabilities in weakly-supervised segmentation. However, some tiny translations, rotations, or scales of the template greatly impact the calculation of the Chamfer Distance between point clouds, especially when the object structure is complex. Thus we conduct a sparse registration [49] between the general shape of the proposal $\mathbb{S}$ and the template

$\mathbb{T}$ to align their scales, shifts, and rotations before calculating the Chamfer Distance. Specifically, we sample 20% points uniformly across the interval for the template and the proposal, respectively. Then we apply the ICP registration tool [49] iteratively calculate the transform matrix (consisting of scale, shift, and rotation): 1) first calculate the correspondence set from the target point cloud and source point cloud transformed with the current transformation matrix. 2) update the transformation matrix by minimizing an objective function defined over the correspondence set. These two steps are conducted iteratively until a pre-set maximum iteration number. Continuously, we can register the template point cloud by the calculated transform matrix.

2) Alignment Loss: The geometric prior is then applied in the loss function of the *mask head* training that we optimize the Chamfer Distance between the proposal $\mathbb{S}$ and the registered template organ $\mathbb{T}$. This can be formulated as follows:

$$L_{\text{geo}} = \frac{1}{|\mathbb{S}|} \sum_{x \in \mathbb{S}} \min_{y \in \mathbb{T}} ||x - y||_2 + \frac{1}{|\mathbb{T}|} \sum_{y \in \mathbb{T}} \min_{x \in \mathbb{S}} ||y - x||_2. \quad (4)$$

Specifically, the mask head produces binary segmentation masks for each proposal. The proposal is a probabilistic segmentation mask consisting of the segmented instances. This segmentation mask is further processed by the Gumbel-Softmax [50] to obtain the binary voxel. Locations of low probability are assigned to 0 and vice versa. We then adopt the gridding reverse [11] to obtain the point cloud proposal $\mathbb{S}$ from the binary voxel. Finally, we calculate the Chamfer Distance between $\mathbb{S}$ and $\mathbb{T}$ as the geometric prior loss.

C. Contrastive Similarity

Gray space performs poorly for the artifacts in medical imaging and similar surrounding tissues. It is not enough to distinguish between positive and negative pixels (, *i.e.*, organ pixels and non-organ pixels). We thus leverage contrastive learning to calculate the contrastive similarity between pixels. Encoding pixels to high-dimensional features and encouraging pixels of the same label to gather around in the embedding space helps to increase the distinguishability.

The contrastive similarity component comprises two major stages: pre-training and training. The contrastive head is first pre-trained with bounding-box annotation only, and then we leverage the pre-trained contrastive head to better supervise the training of the mask head. Following previous contrastive learning works [47], [51], we build a two-layer point-wise convolution $h(\cdot)$ as the *contrastive head* to extract distinct representations from feature maps P. More specifically, we first pre-train the proposed *contrastive head* in a coarse-to-fine fashion where only bounding box annotations are included in the pre-training stage. We encode pixels into embedding features $\mathcal{C} = h(P)$ and encourage pixels of the same label to gather around in the embedding space. The contrastive similarity between two pixels is defined as the distance in the embedding space. To calculate the contrastive similarity loss for the whole image, an undirected graph is constructed where the vertices correspond to the pixels and edges are links

between neighboring pixels. The contrastive similarity associated with each edge is then summarized for the calculation of the overall contrastive similarity loss. Below we first introduce the pre-training of the contrastive head and then the definition of contrastive similarity loss.

1) Pre-Training: The pre-training of the contrastive head is conducted in a weakly-supervised fashion that only bounding box annotations are included in the pre-training stage. To be more specific, there are two sub-stages in the pre-training stage: *coarse* and *refine*. In the *coarse* stage, we first take pixels within the bounding box as positive labels and pixels outside the bounding box as negative labels. Then we train the contrastive head by such labeling. However, the performance of the contrastive head is largely limited for the noisy labels. Thus, we propose the *refine* stage to further improve the performance of the contrastive head. In the *refine* stage, we first take random K negative pixels as *referring pixels*. And for each pixel within the bounding box, we calculate the distance $\mathcal{D}$ between all K referring pixels:

$$D_{u,v,z} = \sum_{i=1}^{K} \mathbb{1}\{\{\mathcal{C}_{u,v,z} \cdot \mathcal{C}_i\} \geq \tau\} \qquad (5)$$

where $\mathcal{C}_{u,v,z} \in R^{S \times H \times W}$ indicates the feature at the (u, v, z) of the embedding features, $\mathbb{1}$ stands for 1 if the distance is greater than τ and 0 if less, and τ is the threshold to decide whether pixels are positive or negative. If $D_{u,v,z}$ is greater than $K/2$, the pixel at location (u, v, z) is considered positive, and vice versa. This process can be formulated as the voting process that the majority decides whether positive or negative. Then, we train the contrastive head using the same training loss as in the *coarse* stage, which is formulated as:

$$loss = -\frac{1}{|\Omega|} \sum_{(u,v,z) \in \Omega} \frac{1}{|\mathcal{P}(u,v,z)|}$$
$$\cdot \log \frac{\sum_{(u_p,v_p,z_p) \in \mathcal{P}(u,v,z)} exp(\mathcal{C}_{u,v,z} \cdot \mathcal{C}_{u_p,v_p,z_p}/\tau)}{\sum_{(u_n,v_n,z_n) \in \mathcal{N}(u,v,z)} exp(\mathcal{C}_{u,v,z} \cdot \mathcal{C}_{u_n,p_n,z_n}/\tau)}$$
$$(6)$$

where $\mathcal{C}_{u,v,z} \in R^{S \times H \times W}$ indicates the feature at the (u, v, z) of the feature map, and Ω stands for all points inside input. $\mathcal{P}(u, v, z)$ denotes the set of points with the same label as the pixel at (u, v, z) and $\mathcal{N}$ denotes the set of points with different labels. τ is the temperature constant.

2) Similarity Loss: The contrastive similarity loss is to calculate the neighboring similarities that we trade each pixel in the training image as vertices and links between the vertices as the edges. We minimize the sum of the edges to optimize the contrastive similarity. Considering an undirected graph $G = (V, E)$ built on the input image I, where V corresponds pixels and E indicates edges between neighboring pixels, the predicted segmentation mask can be viewed as the probability of pixel (u, v, z) being foreground. Then the probability of pixel (u_1, v_1, z_1) and pixel (u_2, v_2, z_2) being the same label is:

$$Prob(y_e = 1) = \tilde{\mathcal{M}}_{u_1,v_1,z_1} \cdot \tilde{\mathcal{M}}_{u_2,v_2,z_2}$$
$$+ (1 - \tilde{\mathcal{M}}_{u_1,v_1,z_1}) \cdot (1 - \tilde{\mathcal{M}}_{u_2,v_2,z_2}), \quad (7)$$

where $\tilde{\mathcal{M}}$ indicates the foreground probability mask and y_e represents the label of the edge.

Thus, we can define an indicator on each edge to indicate whether they belongs to the same label. If the contrastive similarity between two neighboring pixels is above the pre-defined threshold τ, the indicator on the edge linking them is assigned to 1, and 0 vice versa. We discard the edges with 0 and further summarize the contrastive similarity loss of positive edges, which can be formulated as:

$$L_{\text{cons}} = -\frac{1}{N} \sum_{e \in E_{in}} \mathbb{1}_{\{\mathcal{C}_{e_{start}} \cdot \mathcal{C}_{e_{end}} \geq \tau\}} \log Prob(y_e = 1). \quad (8)$$

This serves as the contrastive similarity loss of the whole image.

IV. EXPERIMENTS

A. Datasets

The scope of the proposed framework is organ segmentation, that we select three organ datasets for evluation.

- LiTS: The public liver LiTS [55] dataset comprises 201 CT scans from various CT scanners and devices. The resolution of images in this dataset is from 0.56mm to 1.0mm in axial and 0.45mm to 6.0mm in z direction. Slices in z range from 42 to 1026. We split 131 cases into training and evaluation sets by a ratio of 4:1.
- KiTS21: The publicly accessible KiTS21 [56] dataset consists of 300 cases during the period from 2010 to 2020. Each CT scan in the dataset is annotated by three expert annotators for the following semantic classes: Kidney, Tumor, and Cyst. We split the provided CT scans into training and validation sets by a ratio of 4:1.
- LPBA40: This dataset consists of 40 T1-weighted image volumes from randomly selected cases (among 40 scans: 20 males, 20 females, and 29.2 ± 6.3 years). The scans were acquired with a spatial resolution of $0.86 \times 1.5 \times 0.86$ mm^3. Here, we conduct experiments on the subset of the hippocampus in LPBA40..

B. Implementation Details and Evaluation Metrics

The proposed framework with geometric prior and contrastive similarity can be easily incorporated with any weakly-supervised segmentation models with bounding box annotations. We take the latest BoxInst [7] as our baseline, and experiments are conducted based on this model unless otherwise specified. The training setting is mostly based on BoxInst's training settings and we obtain the optimal hyper-parameters by extensive experiments: The basic learning rate is 0.01 with weight decay 1e-4, patch size $128 \times 128 \times 128$ and a MultiStepLR scheduler with warmup is adopted. Our framework is trained on GeForce RTX 3090 GPU. During the inference stage, we set the threshold of the completeness head to 0.6 by extensive experiments and the threshold of the class head to 0.5 as the default value in BoxInst. The weights of geometric loss and contrastive loss are set to be 0.4 and 0.6, respectively, from extensive experiments of different loss
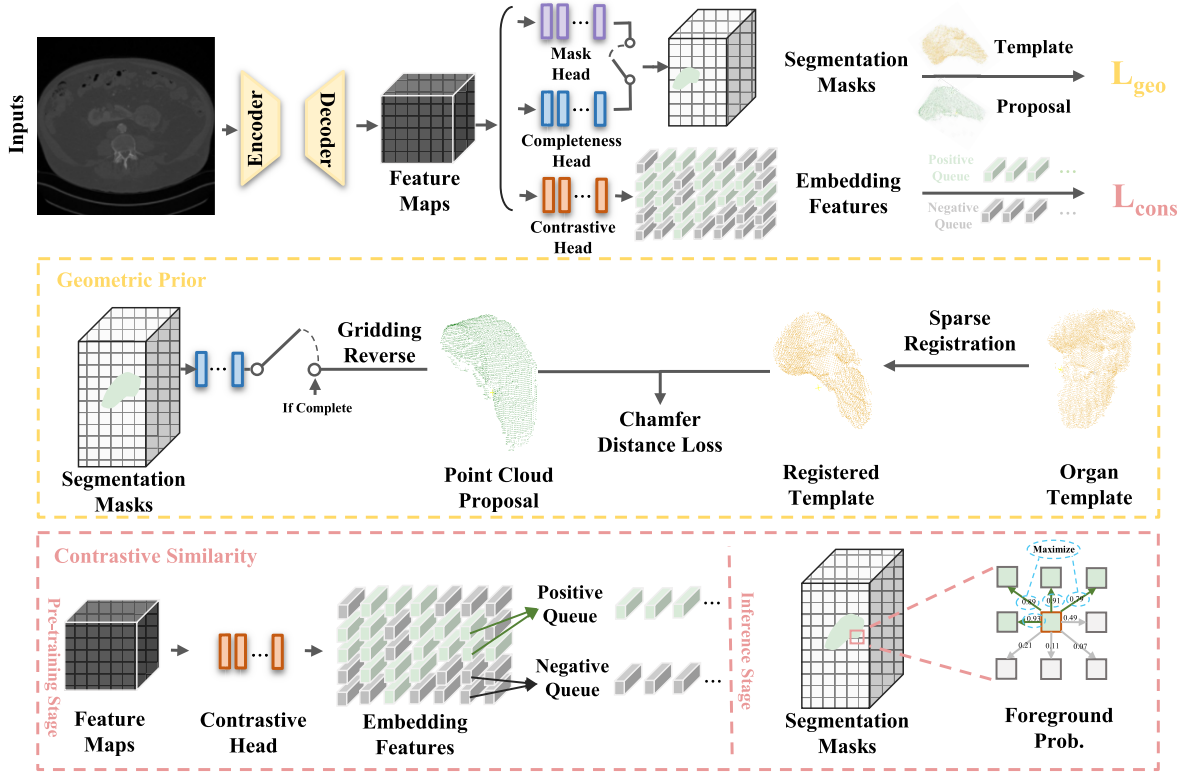
Fig. 2. The overall framework of our weakly-supervised segmentation framework with geometric prior and contrastive similarity. The training of our framework is organized as follows: We first pre-process the input images and obtain the embedding feature maps. Then we jointly supervise the mask head's training by the proposed geometric prior and contrastive similarity. In the aspect of geometric prior, we first convert the segmentation result to point cloud if complete. Then we calculate the Chamfer Distance between the result and the registered template. In the aspect of contrastive similarity, we first pre-train a contrastive head using only bounding box annotations. Then we minimize the distance between pixels of the same labels by the pre-trained contrastive head.

TABLE I

PERFORMANCE COMPARISON WITH OTHER METHODS. WE EVALUATE THE PERFORMANCE USING TWO METRICS: DICE SCORE AND HD95. †
WE OBTAIN THEIR DICE SCORE BY THEIR OPEN SOURCE CODE. ‡: WE REBUILD THE 3D FRAMEWORK BASED ON ITS 2D OPEN SOURCE CODE.
Ours w/o geometric INDICATES WE REMOVE THE GEOMETRIC PRIOR COMPONENT. *Ours w/o contrastive* INDICATES WE REMOVE THE
CONTRASTIVE SIMILARITY COMPONENT

Method	Backbone	LiTS17		KITS21		LPBA40	
		↑DSC(%)	↓HD95	↑DSC(%)	↓HD95	↑DSC(%)	↓HD95
Fully Supervised	UNet [1]	95.5	5.3	96.0	3.2	83.7	2.1
DeepCut (2016) [25]	-	37.1	15.2	36.2	14.7	-	-
SDI (2017) [23]	VGG-16 [52]	49.2	11.7	-	-	38.7	9.0
MIL (2020) [53]†	ENet [54]	69.4	9.4	72.3	8.7	-	-
GMIL (2021) [47]†	ENet [54]	71.1	8.8	71.7	5.9	58.2	4.7
BoxInst (2021) [7]‡	UNet [1]	47.1	11.3	48.4	11.6	37.9	8.9
Ours w/o geometric	UNet [1]	52.9	10.9	54.2	10.7	44.9	7.1
Ours w/o contrastive	UNet [1]	69.7	9.4	68.5	9.1	57.3	4.9
Ours	UNet [1]	**79.8**	**8.7**	**80.2**	**5.3**	**65.4**	**4.2**

weights in Tab. V. We randomly select a training sample as the template for each specific dataset unless otherwise specified. For contrastive similarity, the training epochs of coarse and refine stages are both 60 epochs and we set K to 10 in our experiments. The referring pixels are uniformly randomly sampled. We evaluate our results by two widely adopted metrics: the Dice score (DSC) in percentages and the Hausdorff Distance (95%). 1) DSC score is calculated as

the overlap area of two masks divided by their summation. A higher DSC score corresponds to better overlap with GT. 2) The Hausdorff Distance mainly measures the boundary distance between the segmentation result and the pixel-wise segmentation masks. Better segmentation results are of a smaller value than inferior results.

1) Pre-Processing & Post-Processing: Following nnUNet [2], the pre-processing includes downsampling,

patching, and data augmentation. Here, we downsample the data to reduce memory use and ensure the existence of complete instances inside one patch. Then in the post-processing stage, the pipeline contains resampling, patching, and patch-NMS (Non-Maximum Suppression between patches). We first adjust the spacing and patch the data to the same size as training, with a certain step size. Then we predict a segmentation mask for each patch and return all the predictions to the original space. Since there are overlapping areas between patches, we use patch-NMS in these areas to eliminate the duplicates, which brings an improvement of 0.6% on the segmentation performance.

2) Bounding Box Annotation: We conduct experiments on LiTS, KiTS, and LPBA40 datasets. All these three datasets have pixel-wise segmentation annotations. We utilize the corresponding pixel-wise annotations to obtain the corresponding bounding-box annotations. In the training stage, only bounding-box annotations are included. During the inference stage, we evaluate the performance by pixel-wise annotations.

C. Results

1) Quantitative Results: We report the quantitative results of three datasets (LiTS17, KiTS21, and LPBA40) in Tab. I. Our method aims to integrate the geometric prior and contrastive similarity to give better supervision in the training of the weakly-supervised segmentation framework. Compared with BoxInst [7], which is the baseline of our method, the segmentation performance of our method is largely improved over all three datasets. This is because the geometric prior can supervise the learning of both outer shape and the inner structure, and the contrastive similarity better distinguishes organ and non-organ pixels in the embedding space. Similarly, ours outperforms the most recent two weakly-supervised methods [47], [53] supervised by bounding-box annotations, which are based on the Multiple Instance Learning (*i.e.*, MIL). The reason behind this is that MIL mainly focuses on the boundary regression of the proposal, but additionally, we learn the internal details of organs from the geometric prior. We also evaluate the upper bound of our method in which we supervise the framework training with ground truth.

2) Qualitative Results: We present the qualitative comparison with two state-of-the-art weakly-supervised segmentation methods in Fig. 3. Compared with BoxInst [7], which is the baseline of our framework, we well preserve the proposal's inner structure detail and outer shape. By comparing Col.#4 and Col.#7, the segmentation accuracy is largely improved compared to the baseline in Fig. 3, which refers to the performance comparison (47.1% *v.s.* 79.8%) in Tab. I. This is because the proposed geometric prior and the contrastive similarity help better segment organs. Additionally, the quality of our result is much better than other state-of-the-art methods, especially in terms of internal details. Their methods are based on Multiple Instance Learning (*i.e.*, MIL, GMIL), which mainly focuses on the outer shape of the proposal ignoring the hollows inside the organs (Col.#5 and Col.#6 *v.s.* Col.#7). In contrast to their methods, we learn the inner structure and geometric details from the given template organ along with the outer shape that boosts the training of the mask head. Continuously, to better compare the effect of geometric

TABLE II
ROBUSTNESS ANALYSIS. WE MEASURE THE CORRESPONDING DICE SCORE ON THE KiTS21 DATASET. WE IMPLEMENT THE 3D VERSION OF BOXINST [7] BASED ON ITS 2D VERSION PUBLIC CODE

#	Backbone	Model	Without Our Loss	With Our Loss
1	ResNet [57]	BoxInst 3D [7]	47.8	**77.1**
2	UNet [1]	BoxInst 3D [7]	49.1	**80.2**
3	ResNet [57]	Ai+L [13]	67.2	**72.1**

prior, we conduct the connected component analysis on the segmentation results that there exists only one item in the bounding box. As in Fig. 4, BoxInst and MIL-based methods are improved substantially, while the enhancement in our method is marginal. This is because the geometric prior ensures the integrity and uniqueness of the segment in the bounding box, that no further post-processing step is required in the proposed framework.

D. Ablation Study

In this subsection, we first present the component-wise analysis of our framework and then discuss the effectiveness of the proposed geometric prior and contrastive similarity.

1) Robustness: In this subsection, we mainly discuss the robustness of the proposed weakly-supervised segmentation framework. Our framework is built on BoxInst [7], composed of Backbone and FPN. We analyze the impact of the backbone on the weakly-supervised segmentation performance. As shown in Tab. II, our method is robust to other backbones (*i.e.*, ResNet [57] and U-Net [1]). Furthermore, we verify the generality of the proposed framework by extending it to other weakly-supervised segmentation models. Following Chu et al. [13], we use the Ai+L [13] model with a ResNet50 [57] backbone that is widely used in both non-medical [58] and medical [59] segmentation tasks. We concatenate the geometric prior branch and contrastive similarity branch after the decoder. As shown in Tab. II, the baseline model obtains 67.2% in the KiTS dataset. The segmentation performance is largely improved by 4.9% after adding the proposed geometric prior and contrastive similarity losses.

2) Geometric Prior: In this subsection, we conduct several ablation studies to validate the effectiveness of the proposed geometric priors.

a) Functions of sparse registration: The sparse registration is conducted for the scale/shift/rotation alignment before measuring the Chamfer distance loss for fine-level alignment. If we remove the registration part from the current framework, the training loss explodes to "Not a Number" (NaN) after a few epochs. The reason behind this is that the lack of sparse registration enlarges the Chamfer distance between the proposal and the template. And therefore, the overwhelmingly large gradients lead to the explosion of the training framework. We further compare the performance between the point cloud and volume representations in this sparse registration step. More specifically, the registration is conducted in the point cloud representation by the widely-used SimpleITK [60] with settings as follows: Mean Square metric, random sampling
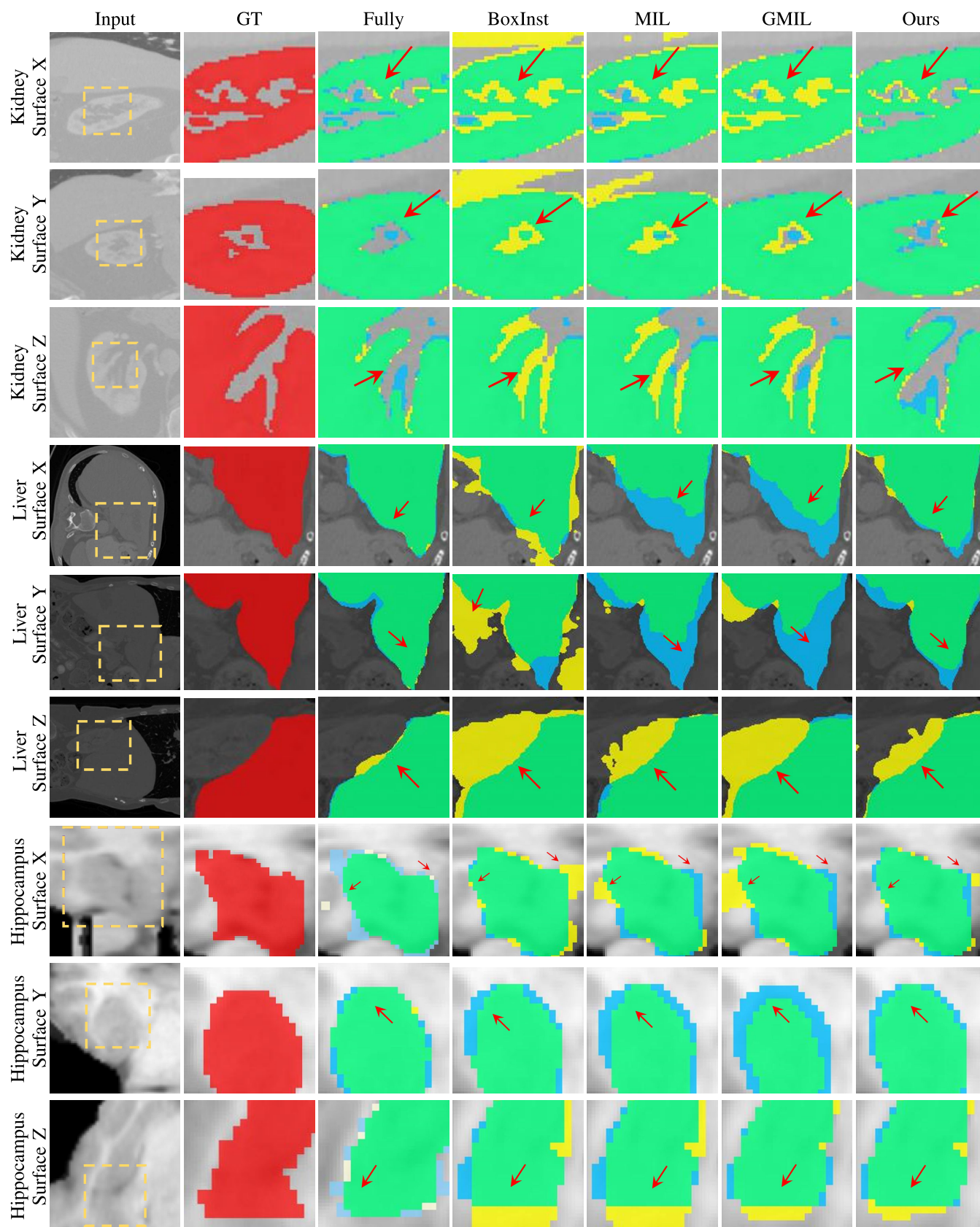
Fig. 3. Qualitative results of our framework over three datasets. As shown in the figure, four colors exist in the segmentation results. *red* indicates ground-truth. *green* indicates correct predictions. *blue* indicates predictions where should have been predicted. *yellow* indicates wrongly-predicted pixels. Our framework is built on BoxInst [7] which is the baseline method. We then compare several methods, including fully-supervised results, the proposed method and demonstrate the high quality of the proposal.
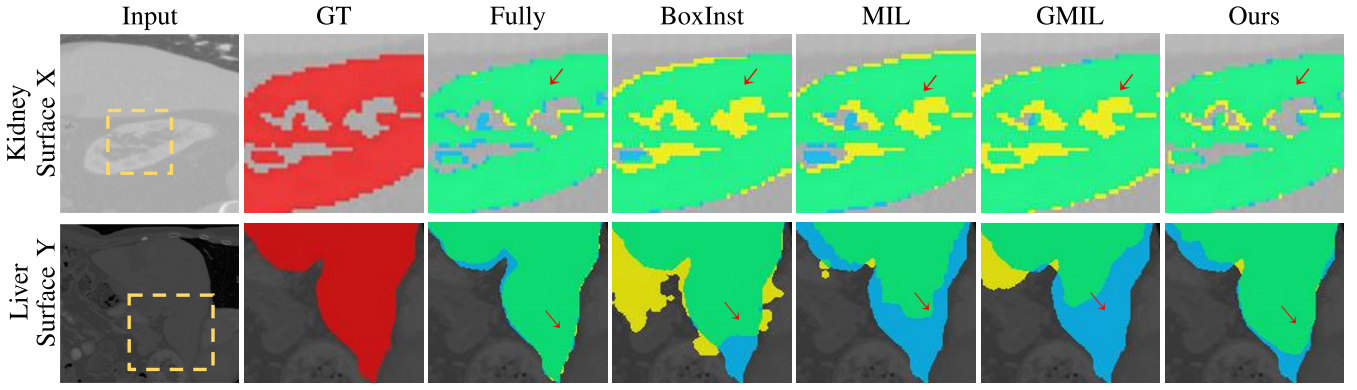
Fig. 4.   Qualitative results with connected component analysis. As shown in the figure, four colors exist in the segmentation results. *red* indicates ground-truth. *green* indicates correct predictions. *blue* indicates predictions where should have been predicted. *yellow* indicates wrongly-predicted pixels. We adopt the same case of 'Kidney - Surface X' and 'Liver - Surface Y' as in Fig. 3.

strategy with a percentage of 0.01, shrink factors [4, 2, 1], smoothing sigmas [2, 1, 0], and 100 iterations. To analyze the difference between point cloud registration and volume registration, we register the organ by SimpleElastix [61] before the conversion to the point cloud representation. The Dice score on KiTS21 is 79.9% compared with 80.2% by the point cloud registration, indicating that our framework remains robust in both representations for sparse registration. We choose the point cloud representation in this sparse registration step to be consistent with the following fine-level alignment step.

b) Point clouds vs. volume representation: For the fine-level alignment, we minimize the Chamfer distance between the template organ and the proposal in the point cloud embedding space. The point cloud representation is much more flexible than the volume representation to align delicate and complex structures. This is because the uniform voxel grids constrain the volume representation. The representative point of each grid is fixed to the center of each corresponding grid. In contrast, there is no such constraint in the point cloud representation. The coordinates of points are much more precise and flexible elastic (*i.e.*, {0.17, 0.18, …} *v.s.* {1.00, 2.00, …}). This helps the point cloud representation to better describe delicate and complex structures than the conventionally-used volume representation. Thus, to validate this point, we conduct an ablation of registration and optimization with the volume representation. More specifically, we replace the Chamfer distance loss defined on point clouds with a Dice loss defined on volumes to optimize the distance between the segmentation result and the geometric prior. As shown in Fig. 5 and Tab. III, we conduct ablation experiments on KiTS, LiTS and LPBA40, respectively. It's clearly noticed that the proposed point cloud representation can handle more complex architectures and delicate geometric shapes by improving 1.7% (78.1% *v.s.* 79.8%), 4.3% (75.9% *v.s.* 80.2%), 1.5% (63.9% *v.s.* 65.4%) in LiTS, KiTS, and LPBA40 in terms of Dice Score Coefficient, respectively.

c) Internal details: We further analyze the impact of the template from the perspective of outer shape and inner structures, respectively. The experiments in this subsection are organized as follows: We first fulfill the inner structure of the template and only retain the outer shape. Then we train the mask head with the deformed template. Normally, there exists intra-organ variabilities, *i.e.*, as shown in Fig. 6,

TABLE III
ANALYSIS OF POINT CLOUD *v.s.* VOLUME. WE COMPARE THE SEGMENTATION PERFORMANCE BY POINT CLOUD REPRESENTATION AND VOLUME REPRESENTATION ON LITS, KITS, AND LPBA40, RESPECTIVELY

#	Representation	DSC on		
		LiTS	KiTS	LPBA40
1	Volume	78.1	75.9	63.9
2	Point Cloud	79.8	80.2	65.4

TABLE IV
ANALYSIS OF GEOMETRIC PRIOR. † INDICATES THE INNER STRUCTURE OF THE TEMPLATE

#	Ablation Setting	DSC on KiTS
1	w/o Completeness Head	79.1
2	w/o Internal Details†	76.3
3	Baseline	80.2

especially meticulous architectures inside the kidney. And the segmentation performance drastically degrades (3.9%) for the lack of internal prior by comparing #2 and #3 in Tab. IV.

d) Template selection: We also investigate the influence of the template utilized as the reference in the geometric prior loss. We first compare the framework performance by different randomly selected templates and then select the worst template (the one with the largest chamfer distance between the mean shape of the training dataset) as the reference to supervise the framework training. Additionally, we utilize a free 3D model[1] as the reference in our method. As shown in Tab. VII, there exists a rare variance in the segmentation performance by the fore-mentioned templates, which strongly indicates the robustness of the template selection in our framework. Despite the presence of tumors or other pathological cases that may alter the organ shape, the shape change generally remains within an acceptable range, and the proposed geometric prior effectively handles these shape changes.

e) Analysis of the completeness head: As discussed in Sec. III-A, following nnUNet [2], for an input image, we first partition it into patches and sequentially process these patches.

---

[1] https://www.cgtrader.com/items/741468/download-page

TABLE V
ANALYSIS OF THE LOSS WEIGHT. WE COMPARE THE SEGMENTATION PERFORMANCE BY THE TRAINING OF THE DIFFERENT LOSS WEIGHT

#	Loss Weight		DSC on
	Geometric Prior	Contrastive Similarity	LPBA40
1	1.0	0.0	57.3
2	0.8	0.2	60.1
3	0.4	0.6	65.4
4	0.2	0.8	58.2
5	0.0	1.0	44.9

TABLE VI
ANALYSIS OF THE CONTRASTIVE HEAD. †: C STANDS FOR THE CONTRASTIVE SIMILARITY, S INDICATES THE SSIM SIMILARITY, AND G STANDS FOR THE GRAYSCALE SIMILARITY. THE INFERENCE SPEED INDICATES THE TOTAL TIME OF 1000 RUNS

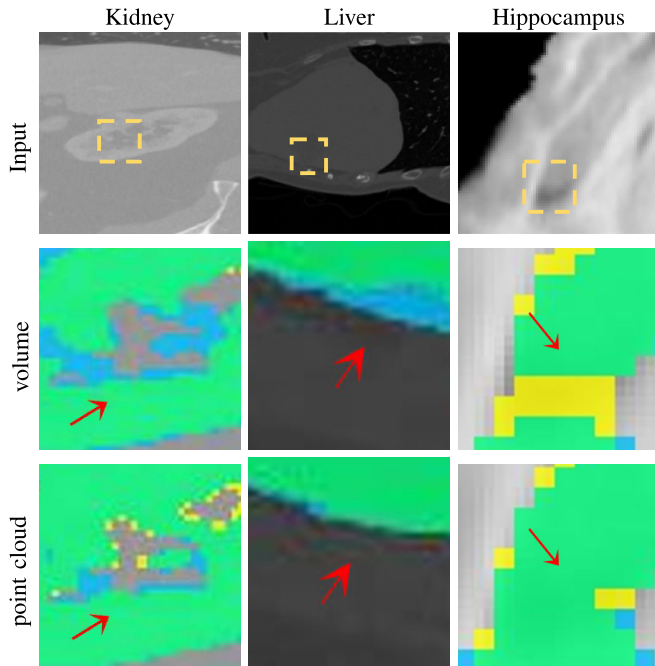| #	Component Setting	Embedding Dimension	Embedding Space†	Inference Speed	DSC (%) on KiTS
1	Contrastive Head	8	C	3s	78.4
2		16	C	5s	79.1
3		32	C	8s	80.2
4		64	C	15s	80.4
5	Similarity Metrics	32	C	8s	80.2
6		1	G	8s	73.4
7		1	S	12s	47.2

Fig. 5. Volume *v.s.* Point cloud. We optimize the learning of the geometric shape in volume representation by Dice Loss. Similarly, *green*, *blue*, and *yellow* indicates the correct, missing and wrong results. We compare the point cloud representation and volume representation on KiTS, LiTS, and LPBA40, respectively. As shown in the figure, the proposed point cloud representation largely improves the segmentation result.

However, during the partitioning, organs may not be complete, which may cause the failure of learning from the template. Thus, to ensure the success of learning from the template, the proposed completeness head detects whether the proposal in the sampled patch is complete. Only complete proposals are further processed. To validate the effectiveness of the completeness head, we conduct an experiment to
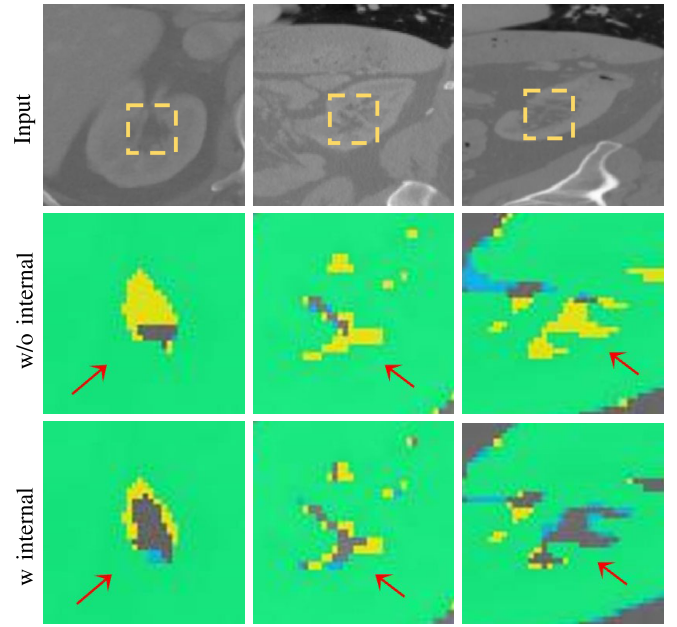

Fig. 6. Internal details. To verify the effectiveness of the internal details of the template, we erase the intra-organ variabilities by fulfilling the template organ. Similarly, *green*, *blue*, and *yellow* indicates the correct, missing and wrong results. As shown in the figure, the segmentation performance drastically degrades, especially for organs with delicate and complex structures. Figures from left to right are more and more delicate and complex.

TABLE VII
ANALYSIS OF THE TEMPLATE SELECTION. † INDICATES THE FREE 3D MODEL FROM THE PUBLIC RESOURCES

#	Ablation Setting	DSC on KiTS
1	random reference 1	79.8
2	random reference 2	79.6
3	worst reference	79.2
4	3D model†	79.3
5	Baseline	80.2

remove the completeness head, as shown in Tab. IV (#1 *v.s.* #3), the performance degrades by 1.1% compared with our framework. The completeness head is trained in a fully-supervised fashion: the completeness label of a patch is determined by whether the bounding box is complete or not in the patch. The completeness head achieves 97.3% accuracy in whether the target object is complete or not. So, theoretically, the performance would degrade if the completeness head's prediction is wrong, but in practice, very few outliers exist, and they cannot influence the segmentation performance.

3) Architectures of the Contrastive Head: We conduct experiments on the design of the proposed contrastive head in this subsection. Following previous works [47], [51], the contrastive head is composed of a two-layer point-wise convolution that helps distinguish organ tissues from non-organ tissues. To further analyze the structure of the contrastive head, we conduct ablation experiments on the dimension of the embedding space. As shown in Tab. VI, we evaluate both the performance and the efficiency of different embedding dimension that by considering the

trade-off between efficiency and performance, we empirically set the embedding dimension to 32.

a) Contrastive similarity: We propose contrastive similarity in the embedding space. To compare the embedding space with the commonly used gray space, we select the following two similarity metrics: *MSE* and *SSIM* [62]. Since medical images are in grayscale, both *MSE* and *SSIM* are calculated on gray values. As shown in Tab. VI, the *MSE* metric fails to distinguish low-contrast tissues (73.4% *v.s.* 80.2%), and the *SSIM* metric is inferior to the contrastive similarity in terms of both segmentation performance and evaluation efficiency (47.2% *v.s.* 80.2% and 12s *v.s.* 8s). The proposed contrastive similarity outperforms other similarity metrics on both accuracy and efficiency.

E. Discussion on Different Annotations

In this subsection, we first compare the annotation efforts of weakly-supervised signals and then incorporate point annotations with bounding-box annotations for better segmentation. Firstly, we demonstrate the annotation efforts of point, bounding box, and full supervision. The annotation cost of full supervision is 477 s/scan [63] by experts, while the annotation cost of the extreme points and bounding box are both 37 s/scan [63]. This indicates that weak supervision vastly reduces expert annotation efforts.

We then explore the potential of incorporating bounding-box annotations with point annotations for more accurate segmentation. The bounding-box annotation gives the shape and location of the target segments, and the point annotation provides delicate object boundaries. More specifically, we randomly sample 100 points inside the bounding box and obtain their labels (*e.g.*, 1 for foreground and 0 for background). Then we apply the cross entropy loss to supervise these points. Leveraging the point annotation, the performance is primarily improved by 0.5% (80.7% *v.s.* 80.2%) for the accurate boundaries from the point annotation.

V. CONCLUSION

In this work, we propose a novel weakly-supervised segmentation framework with bounding-box annotations. We introduce the geometric prior and the contrastive similarity to address the challenges of complex shapes and imaging artifacts, enhancing the practicability and robustness of the weakly-supervised segmentation framework. The geometric prior enables the learning of delicate and complex structures, and the contrastive similarity helps better distinguish organ pixels from non-organ pixels.

Our framework is general, which can be easily applied to many weakly-supervised segmentation models and thus improve their performances. It will probably serve as a base for possible future studies on weakly-supervised segmentation, especially for medical image segmentation on organs that are with a specific shape. Extensive experiments are conducted to verify the effectiveness and the superiority of our proposed geometric prior and contrastive similarity.

Limitation and future works The proposed framework requires the target object to be complete in the patch,

which may lead to the performance degradation of the weakly-supervised segmentation for some overwhelmingly large segments (*e.g.*, white matter in the brain, vein). Future works should include the circumstance of incomplete organ segmentation. In addition, our geometric prior is tailored for medical segmentation with shared anatomical structures, such as organ segmentation. However, it is not designed to handle variations in anatomy, such as tumor segmentation, and may even exacerbate the situation. This is due to the absence of standardized anatomical priors for certain segmentation tasks across different patients. For instance, the shape of tumors can differ significantly between patients, and the selected template may not be suitable for all cases, thereby misleading the segmentation learning. Future research could explore the incorporation of deformable templates to account for the variability in anatomical structures encountered during segmentation tasks or explore alternative statistical priors that are better suited to handling anatomical variations.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[2] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2021.

[3] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1796–1804.

[4] Z. Jia, X. Huang, E. I.-C. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2376–2388, Nov. 2017.

[5] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Med. Imag. Anal.*, vol. 54, pp. 88–99, May 2019.

[6] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. B. Ayed, "Constrained domain adaptation for segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 326–334.

[7] Z. Tian, C. Shen, X. Wang, and H. Chen, "BoxInst: High-performance instance segmentation with box annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5443–5452.

[8] T. Budrys, V. Veikutis, S. Lukosevicius, R. Gleizniene, E. Monastyreckiene, and I. Kulakiene, "Artifacts in magnetic resonance imaging: how it can really affect diagnostic image quality and confuse clinical diagnosis?" *J. Vibroeng.*, vol. 20, no. 2, pp. 1202–1213, 2018.

[9] F. E. Boas and D. Fleischmann, "CT artifacts: Causes and reduction techniques," *Imag. Med.*, vol. 4, pp. 229–240, Apr. 2012.

[10] S. N. Sarkar et al., "A subjective and objective comparison of tissue contrast and imaging artifacts present in routine spin echoes and in iterative decomposition of asymmetric spin echoes for soft tissue neck MRI," *Eur. J. Radiol.*, vol. 102, pp. 202–207, May 2018.

[11] H. Xie, H. Yao, S. Zhou, J. Mao, S. Zhang, and W. Sun, "GRNet: Gridding residual network for dense point cloud completion," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 365–381.

[12] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.

[13] T. Chu, X. Li, H. V. Vo, R. M. Summers, and E. Sizikova, "Improving weakly supervised lesion segmentation using multi-task learning," in *Proc. Med. Imag. With Deep Learn.*, 2021, pp. 60–73.

[14] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3159–3167.

[15] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 549–565.

[16] H. Qu et al., "Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3655–3666, Nov. 2020.

[17] G. Patel and J. Dolz, "Weakly supervised segmentation with cross-modality equivariant constraints," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102374.

[18] G. Xu et al., "CAMEL: A weakly supervised learning framework for histopathology image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10681–10690.

[19] X. Wang et al., "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3950–3962, Sep. 2020.

[20] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[21] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1635–1643.

[22] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1742–1750.

[23] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 876–885.

[24] M. Pu, Y. Huang, Q. Guan, and Q. Zou, "GraphNet: Learning image pseudo annotations for weakly-supervised semantic segmentation," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 483–491.

[25] M. Rajchl et al., "DeepCut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 674–683, Feb. 2016.

[26] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson, "A Bayesian model of shape and appearance for subcortical brain segmentation," *NeuroImage*, vol. 56, no. 3, pp. 907–922, Jun. 2011.

[27] M. R. Sabuncu, B. T. Yeo, K. V. Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1714–1729, Oct. 2010.

[28] B. Fischl et al., "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[29] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, 2015.

[30] M. Gao et al., "Segmentation label propagation using deep convolutional neural networks and dense conditional random field," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1265–1268.

[31] P.-A. Ganaye, M. Sdika, and H. Benoit-Cattin, "Semi-supervised learning for segmentation under semantic constraint," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 595–602.

[32] A. BenTaieb and G. Hamarneh, "Topology aware fully convolutional networks for histology gland segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 460–468.

[33] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "DCAN: Deep contour-aware networks for object instance segmentation from histology images," *Med. Image Anal.*, vol. 36, pp. 135–146, Feb. 2017.

[34] Y. Zhou 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.

[35] O. Oktay et al., "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 384–395, Feb. 2018.

[36] A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9290–9299.

[37] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany, "Point-contrast: Unsupervised pre-training for 3D point cloud understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 574–591.

[38] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16684–16693.

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[40] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8219–8228.

[41] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," 2021, *arXiv:2112.09645*.

[42] X. Zhao et al., "Contrastive learning for label efficient semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10623–10633.

[43] J. Peng, M. Pedersoli, and C. Desrosiers, "Boosting semi-supervised image segmentation with global and local mutual information regularization," 2021, *arXiv:2103.04813*.

[44] J. Xiang, Z. Li, W. Wang, Q. Xia, and S. Zhang, "Self-ensembling contrastive learning for semi-supervised medical image segmentation," 2021, *arXiv:2105.12924*.

[45] Y. Zhou, H. Xu, W. Zhang, B. Gao, and P.-A. Heng, "C3-SemiSeg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7036–7045.

[46] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2228–2237, Sep. 2022.

[47] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 481–490.

[48] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[49] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.

[50] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2016, *arXiv:1611.01144*.

[51] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[53] H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. B. Ayed, "Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision," 2020, *arXiv:2004.06816*.

[54] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[55] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*.

[56] N. Heller et al., "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Med. Image Anal.*, vol. 67, Jan. 2020, Art. no. 101821.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[58] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019.

[59] B. M. Anderson et al., "Automated contouring of contrast and non-contrast computed tomography liver images with fully convolutional networks," *Adv. Radiat. Oncol.*, vol. 6, no. 1, Jan. 2021, Art. no. 100464.

[60] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of SimpleITK," *Front. Neuroinform.*, vol. 7, p. 45, Dec. 2013.

[61] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "SimpleElastix: A user-friendly, multi-lingual library for medical image registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 134–142.

[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[63] R. Dorent et al., "Inter extreme points geodesics for end-to-end weakly supervised image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 615–624.