GeoCoBox: Box-supervised 3D Tumor Segmentation via Geometric Co-embedding

Tianzhong Lan[1], **Zhang Yi**[1], **Xiuyuan Xu**[1*], **Min Zhu**[1*]

[1]Sichuan University
{lantianzhong1@stu., zhangyi@, xuxiuyuan@, zhumin@}scu.edu.cn

Abstract

Data economics drives AI by optimizing data usage, reducing costs, and enhancing efficiency. In 3D tumor segmentation, efficiency is crucial due to the high demand for labor-intensive manual annotations. Box-supervised segmentation offers a promising alternative but is constrained by tumor morphology complexity and boundary ambiguity. In this paper, we propose a novel 3D tumor segmentation model that integrates both positional and embedding features to facilitate inter-task collaboration. We introduce an Anatomical-Driven Class Activation Map to predefine the complex tumor morphology prior, which is further refined by our Geometric Pixel Co-embedding Learner. This learner utilizes contrastive learning to encode semantic information between center and edge pixels, enhancing pixel clustering and progressively refining tumor boundary segmentation in a coarse-to-fine manner. Our approach outperforms existing box-supervised methods in segmentation performance, with extensive experiments on four tumor datasets demonstrating significant improvements. This work provides a cost-effective and efficient solution for tumor segmentation, advancing the application of data economics in medical imaging.

Introduction

A large volume of precise data is essential for effectively training deep neural networks (DNNs). However, achieving data economic efficiency by reducing annotation costs offers even greater scientific and technological value in DNN applications. Tumor segmentation in 3D medical images is more time-consuming and labor-intensive than computer vision tasks on natural images, as the precise annotations required for tumor segmentation are more challenging. Due to the advent of label-efficient learning (Jin et al. 2023; Yi et al. 2025), a common approach in natural image research to reducing the reliance on pixel-level annotations is box-supervised segmentation. Bounding boxes provide the most supervision signals compared to other weak labels because they offer approximate information about the shape and size of the object. Many methods employ bounding boxes for coarse supervision and optimize the segmentation results (Lu, Deng, and Zhang 2024; Tian et al. 2021; Li et al. 2024; Kulharia et al. 2020).

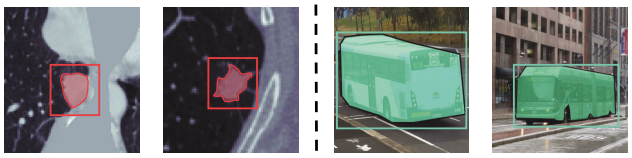

Figure 1: Tumors vary greatly in size and shape with vague boundaries, unlike natural objects with regular shapes and clear edges.

However, as shown in Fig. 1, methods that perform well on natural images may not apply directly to medical images due to several factors: 1) organs and tumors vary significantly in size and shape; 2) identifying low-contrast tissues and nonhomogeneous textures in medical images is challenging, as they only contain grayscale information, unlike natural images, which include color for additional cues. As indicated in Fig. 2, existing methods (Du et al. 2023; Zhao et al. 2024; Zhong et al. 2024; Shao et al. 2025; Li et al. 2022; Wang and Xia 2021; Cheng et al. 2023) can be classified into three types of pipelines: The first type (Fig. 2 (a)) assigns pseudo-labels for mask supervision; however, these pseudo-labels may be coarse and unreliable, leading to significant performance degradation (Wu et al. 2023). The second type (Fig. 2 (b)) relies on pixel pairwise learning, which is simple and effective but not suitable for grayscale medical images(Du et al. 2023). The third type (Fig. 2 (c)) utilizes organ templates as explicit priors, combined with pixel embeddings, to achieve good results. In contrast to organ segmentation that relies on consistent anatomical structures, tumor segmentation faces greater challenges due to small lesion sizes and highly variable morphological characteristics. Additionally, existing methods fail to capture inter-task correlations, resulting in isolated branches during the learning process.

In this paper, we reconsider the class activation map (CAM) method, a popular technique in weakly supervised research (Zhou et al. 2016; Chen et al. 2022). Under box constraints, we introduce an Anatomical-Driven CAM (AD-CAM) to provide an anatomical prior for characterizing and localizing tumor structures. Furthermore, the grayscale information motivates us to experiment with contrastive learn-

*Corresponding Authors.

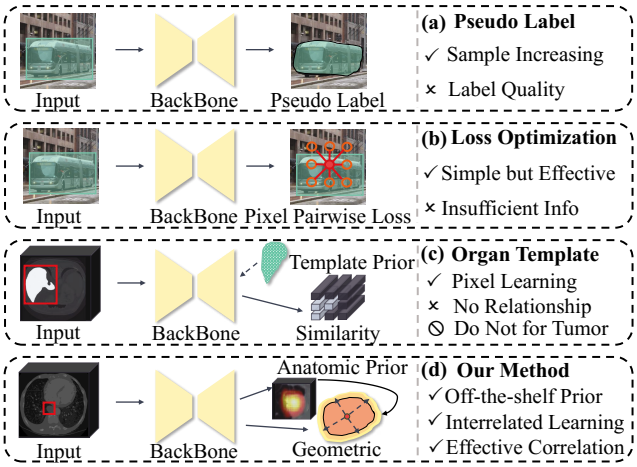

Figure 2: The current box-supervised segmentation pipelines. The comparison between (a) pseudo-label generation models, (b) loss-based models, (c) organ template learning models, and (d) our approach.

ing to encode high-dimensional semantic information for each pixel, promoting the clustering of pixels from the same labels. We propose a geometric pixel co-embedding learner to fix the results from AD-CAM in a coarse-to-fine manner. As stated in (Zhai et al. 2023; Hu et al. 2024), the center points of box-shaped masks have the **highest probability** of being positive samples, whereas the boundary information provided by tumor priors is prone to classification errors. In contrast to previous work (Du et al. 2023; Hu et al. 2021; Zhai et al. 2023; Hu et al. 2024), we explicitly integrate the pre-trained contrastive head with the positional information provided by tumor priors, focusing solely on training the similarity between center and boundary points. This approach not only enhances segmentation performance but also reduces computational cost. This idea aligns with the notion that human annotators tend to pay discriminative attention to the center and the boundary regions. We also propose a novel pretraining strategy for the contrastive head. It involves supervised contrastive learning using only positive center samples and pre-determined negative samples from outside the bounding box. This method similarly reduces computational overhead.

Our main contributions are highlighted as follows:

- We propose a 3D tumor segmentation model for box supervision, namely GeoCoBox. It explicitly embeds the positional information with contrastive features, enabling intertask collaboration.
- We introduce an Anatomical-Driven Class Activation Map for predefining the tumor morphology that can provide an anatomical prior for pixel-wise learning.
- We propose a Geometric Pixel Co-embedding Learner for refining the tumor boundaries. This method, along with our proposed contrastive head pretraining strategy, effectively utilizes the tumor center information, thereby reducing computational overhead.
- We benchmark our proposed GeoCoBox on four widely used public tumor segmentation datasets. The results demonstrate the effectiveness of our method.

Related Works

Box-supervised Semantic Segmentation

Box-supervised segmentation (Li et al. 2024; Du et al. 2023) has gained increasing attention in medical image segmentation due to its simplicity and low annotation cost. It precisely serves as the fully supervised label for object detection tasks. Therefore, labels collected for detection can be directly applied to weakly supervised segmentation tasks. As a weak label, the bounding box is location-aware and provides a relatively tight prior, particularly in medical images where object boundaries are not always clearly defined. In the domain of natural images, studies often treat box-supervised segmentation as a denoising task, where pixels within the bounding box are considered foreground, and the segmentation model is trained based on pixel relationships(Hsu et al. 2019; Lan et al. 2021). However, grayscale data in the medical domain often lacks sufficient pixel information in most medical images. Many studies incorporate additional prior knowledge to assist with pixel association tasks.

Prior Generation

To reduce annotation costs, box-supervised segmentation methods typically follow two strategies: pseudo-label generation and multiple-instance learning (MIL). Early works such as BoxSup (Dai, He, and Sun 2015) and Box2Seg (Kulharia et al. 2020) generate pseudo masks from ground-truth boxes. More recent MIL-based approaches (Tian et al. 2021; Cheng et al. 2023; Shao and Fang 2025; Yi et al. 2022) leverage box compactness, treating projections along box-aligned axes as positive or negative samples. However, these methods struggle in medical imaging due to low contrast and textural complexity.

In the medical domain, some studies incorporate anatomical priors to guide learning. For example, (Zhou et al. 2019) introduces size-aware losses, and (Du et al. 2023) combines organ templates with point cloud embeddings. Others add regularization based on organ adjacency or boundary conditions (Ganaye, Sdika, and Benoit-Cattin 2018; Liu et al. 2025). Yet, such priors are often inapplicable to tumors with high variability.

Contrastive Learning

Contrastive learning learns discriminative representations by pulling similar samples closer and pushing dissimilar ones apart without extra annotations (Chen et al. 2020; Wang et al. 2024). In segmentation, it improves intra-class compactness and inter-class separability (Wu, Zhuang, and Chen 2024; Zhao et al. 2021; Zhang et al. 2025; Zhou et al. 2021), and is particularly useful for grayscale medical images that lack strong texture cues (Hu et al. 2021; Alonso et al. 2021). Existing methods often adopt a two-stage pipeline—first pretraining with unlabeled or weak labels, then refining with additional supervision—yet they treat branches independently and ignore inter-task synergy. In contrast, we perform one-stage pretraining by contrasting edge and center

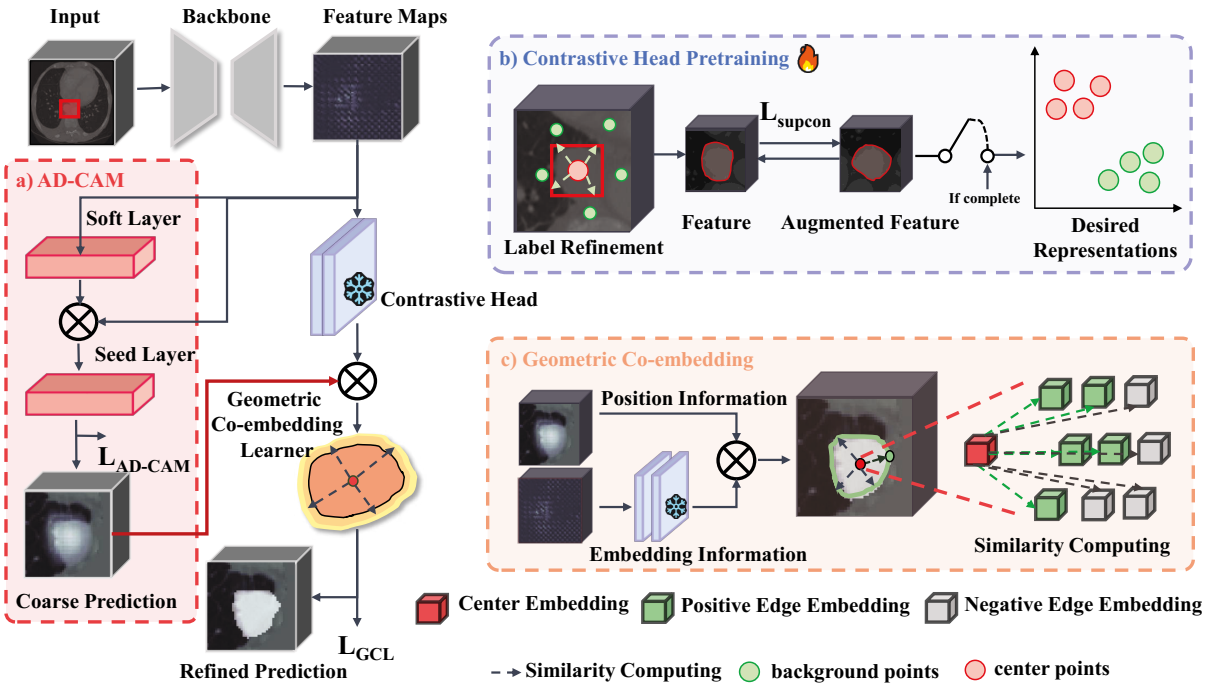

Figure 3: The GeoCoBox framework uses 3D-UNet as the backbone. It includes two branches: AD-CAM and GCL. The AD-CAM branch generates coarse masks and provides edge positions. The GCL branch is employed to compute the similarity between the center embedding and the edge embeddings provided by the pre-trained contrastive head. The red line indicates that GCL explicitly integrates the positional information provided by AD-CAM and the embedding information from the contrastive head.

pixels within box-shaped regions. We further refine results using predictions from the AD-CAM branch, reducing computation and ensuring label continuity without the need for handcrafted weight matrices.

GeoCoBox

As shown in Fig. 3, our GeoCoBox consists of two key components: AD-CAM and Geometric Co-embedding Learner (GCL). The framework trains in an end-to-end manner. This paper focuses on the binary classification task of determining the presence or absence of tumors, and therefore does not involve discussions on multi-class tumor classification.

Preliminaries

Given an input image $\mathcal{X} \in \mathbb{R}^{D\times H\times W}$ (D, H, and W indicate the depth, height, and width, respectively), and the corresponding GT box annotations $\mathcal{B} \in \mathbb{R}^{b\times 6}$ (corner coordinates of the upper left and bottom right points of b boxes), the box-filled mask $B \in \{0,1\}^{D\times H\times W}$ is created by assigning 1 within $\mathcal{B}$ and 0 outside $\mathcal{B}$. The goal of the proposed framework $\mathcal{F}$ is to minimize the loss function $\mathcal{L}_{total}$: $\min_{\mathcal{F}} \mathcal{L}_{total}(\mathcal{X}, B, \mathcal{F})$, where the $\mathcal{L}_{total}$ is composed of the loss function of the AD-CAM and the GCL.

Anatomical-Driven CAM

CAM is typically used in weakly supervised segmentation tasks to generate finer pseudo masks. The standard CAM pipeline initiates with training a classification network, where prediction errors for individual training samples are quantified using binary cross-entropy loss. Upon model convergence, the patch x is processed through the network to derive its corresponding map:

$$CAM(x) = \frac{ReLU\left(Wf(x)\right)}{\max\left(ReLU\left(Wf(x)\right)\right)}, \tag{1}$$

where W means the convolutional layer for classification, and $f(x)$ represents the feature maps of the input x. Inspired by (Chen et al. 2022), which is tailored for weakly-supervised tasks. We employ two fully connected (FC) layers for anatomic prior. Specifically, we use the first FC layer to make the prediction, its logits can be denoted as:

$$z = FC_{soft}(f(x)), \tag{2}$$

we then compute the element-wise multiplication between z and each channel of $f(x)$ to get the global context feature as follows:

$$f^m(x) = z \otimes f(x), \tag{3}$$

where $f(x)$ and $f^m(x)$ indicate features before and after the multiplication. The multiplied feature map $f^m(x)$ is then fed to another FC layer to learn the relationship between foreground and background, so we obtain a new prediction for x:

$$z^m = FC_{seed}(f^m(x)), \tag{4}$$

where the FC_{seed} has the same architecture as FC_{soft}. Due to the sum-over-class pooling nature of BCE, each pixel in

CAM may be responsive to multiple classes co-occurring in the same receptive field. Unlike ReCAM (Chen et al. 2022), we use Soft Margin (SM) loss and Dice loss to reactivate the model and generate results, which is more suitable for mask generation. The SM loss is formulated as:

$$\mathcal{L}_{SM} = \frac{1}{C}\sum_{i}^{C}\log(1+\exp(-B[i]\times z^m[i])), \quad (5)$$

where i in $\mathbf{B}[i]$ and $z^m[i]$ denotes the i-th elements. C indicates the total number of elements. The overall objective function for the AD-CAM branch is formulated as:

$$\mathcal{L}_{AD-CAM} = \mathcal{L}_{SM} + \mathcal{L}_{Dice}, \quad (6)$$

where $\mathcal{L}_{Dice}$ is defined as follow:

$$\mathcal{L}_{Dice} = -\frac{2\,|B\times z^m|}{|B|+|z^m|}. \quad (7)$$

Geometric Co-embedding Learner

Grayscale medical images alone do not provide sufficient information for pixel-wise learning, leading to inaccuracies in the edge information predicted solely by AD-CAM. Following prior works (Zhai et al. 2023; Hu et al. 2024), we assume the box center as a positive seed, as tumors are often annotated with center-focused boxes by radiologists, especially under 3D patch cropping. We propose GCL to compute the similarity between center pixels and edge pixels for refining the tumor edge textures provided by AD-CAM. GCL consists of two components: contrastive head pertaining (Fig.3 (b)) and geometric co-embedding application (Fig.3 (c)).

We build a two-layer point-wise convolution as the contrastive head. We obtain the central point pixels from the initial box-shaped mask to pre-train the contrastive head. As shown in Fig. 3 (b), we treat the positive anchor set by including voxels within a small intensity range δ and distance range τ around the center, enabling more diverse representation while maintaining center consistency, where τ is a hyperparameter determined by the scale of the box. Moreover, to mitigate the potential confusion caused by negative samples with similar intensities, we filter out those whose intensity differences from the center region fall within δ. This significantly improves the contrastive margin and avoids ambiguity near box boundaries. Finally, we use supervised local contrastive learning to train a two-layer point convolution head for extracting distinct representations from feature maps. The local contrastive loss is defined as:

$$\mathcal{L}_{supcon} = -\frac{1}{|\Omega|}\sum_{(d,h,w)\in\Omega}\frac{1}{|P|} \cdot \log\frac{\sum_{(d_p,h_p,w_p)\in P}\exp\left(f_{d,h,w}(x)\cdot f_{d_p,h_p,w_p}(x)/\tau\right)}{\sum_{(d_n,h_n,w_n)\in N}\exp\left(f_{d,h,w}(x)\cdot f_{d_n,h_n,w_n}(x)/\tau\right)}, \quad (8)$$

where $f_{d,h,w}(x)$ indicates the pixel feature at the (d,h,w) of the feature map. Ω is the set of points in $f_{d,h,w}(x)$ that are used to compute loss. P and N represent the positive and negative sets of $f_{d,h,w}(x)$, respectively. τ is the temperature constant. Upon completion of the pre-training phase, as illustrated in 3(c), we utilize both the tumor location information generated by AD-CAM and the embedded features extracted from the frozen contrastive head. Specifically, we designate the centroid of the box-shaped mask as an anchor point and calculate the similarity loss based on the edge pixels identified by AD-CAM. We select all pixels within a distance of less than τ from the edges and compute their similarity to the center point. We calculate the cosine similarity between every chosen pixel f_{d_e,h_e,w_e} and the centroid feature f_{d_c,h_c,w_c}:

$$s_{ce} = \frac{f_{d_c,h_c,w_c}\cdot f_{d_e,h_e,w_e}}{\|f_{d_c,h_c,w_c}\|\,\|f_{d_e,h_e,w_e}\|}, c\neq e, \quad (9)$$

then the probability of the center pixel and chosen pixel being the same label is:

$$Prob(edge=1) = f_{d_c,h_c,w_c}\cdot f_{d_e,h_e,w_e} + (1-f_{d_c,h_c,w_c})\cdot(1-f_{d_e,h_e,w_e}), \quad (10)$$

thus we can define the geometrical-refined loss as follows:

$$\mathcal{L}_{GCL} = -\frac{1}{|\Theta|}\sum_{e\in\Theta}\mathbb{1}_{(s_{ce}>0)}\log\ Prob(edge=1). \quad (11)$$

This serves as the loss function for the geometric co-embedding branch. Θ denotes the set of chosen pixels, and $\mathbb{1}$ indicates an indicator on each edge to control computation.

Experiments

Datasets

We select four public medical datasets to evaluate the effectiveness of our model. All four datasets are split into training and test sets at a 4:1 ratio.

- LIDC-IDRI: The LIDC-IDRI(Armato III et al. 2011) dataset comprises lung cancer screening thoracic CT scans with marked-up annotated tumors. Following the previous work(Tang, Zhang, and Xie 2019), we include only the CT scans that meet the selection criteria of LUNA16, resulting in 586 CT scans with a total of 1,131 nodules.
- KiTS19: The KiTS19(Heller et al. 2021) dataset comprises 210 abdominal CT volumes with manually delineated kidney and tumor labels. Three expert annotators annotate each CT scan in the dataset. We only use tumor labels.
- MSD-Lung & Pancreas: These two datasets are part of the MSD(Antonelli et al. 2022) collection. MSD-Lung contains 3D CT scans from 63 patients, which are used for lung cancer segmentation. MSD-Pancreas focuses on pancreatic tumor segmentation, including 281 CT scans.

Implementation and Training Details

The networks are implemented using Pytorch 2.1.1, The experiments are performed on two Nvidia GeForce RTX 4090 GPUs with 24 GB. The basic learning rate is 0.01, with a stochastic gradient descent (SGD) optimizer. GeoCoBox is trained with a batch size of 32 for 80 epochs; the first 40

Dataset	LIDC-IDRI		KiTS19		MSD-Lung		MSD-Pancreas	
Metrics	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
Fully Supervised	79.97	1.84	65.15	10.40	57.93	7.34	64.91	8.96
Box2Mask(Li et al. 2024)	68.25	3.32	63.65	11.16	48.06	17.01	60.75	11.61
BoxInst(Tian et al. 2021)	74.02	2.35	64.78	<u>10.70</u>	49.15	16.92	<u>64.26</u>	<u>9.15</u>
BBTP(Wang and Xia 2021)	74.00	2.31	<u>64.94</u>	10.92	49.64	15.68	62.47	9.45
GPCS(Du et al. 2023)	<u>74.59</u>	2.19	64.10	11.72	48.80	15.09	64.01	9.16
RTNet(Lu et al. 2024)	72.15	2.62	62.64	12.15	46.49	17.74	63.62	9.27
GazeMedSeg(Zhong et al. 2024)	73.94	<u>2.16</u>	63.53	11.01	<u>50.35</u>	15.28	63.70	9.30
AGMM(Wu et al. 2023)	71.87	2.53	64.30	10.84	49.60	**14.49**	61.23	9.52
GeoCoBox (Ours)	**76.09**	**2.05**	**65.11**	**10.55**	**50.99**	<u>14.61</u>	**64.65**	**9.07**

Table 1: Performance comparison with other state-of-the-art methods. All methods below use 3D UNet as the backbone. We evaluate the performance using two metrics: DSC (%) and HD95. Bold numbers indicate the best results, while underlined numbers represent the second-best results.

Dataset	LIDC-IDRI		KiTS19		MSD-Lung		MSD-Pancreas	
Metrics	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
Fully Supervised	83.54	1.57	63.87	11.36	58.70	9.37	63.94	9.64
Box2Mask(Li et al. 2024)	68.54	3.18	61.09	12.47	49.82	14.43	61.48	10.16
BoxInst(Tian et al. 2021)	74.45	2.15	<u>62.21</u>	12.41	<u>51.46</u>	<u>13.80</u>	62.70	9.97
BBTP(Wang and Xia 2021)	<u>75.70</u>	<u>2.13</u>	61.72	12.53	50.04	14.75	62.39	9.91
GPCS(Du et al. 2023)	74.32	2.43	60.29	12.70	51.11	14.44	<u>63.04</u>	10.15
RTNet(Lu et al. 2024)	72.20	2.37	61.54	<u>12.24</u>	48.97	14.97	61.57	**9.53**
GazeMedSeg(Zhong et al. 2024)	73.05	2.29	60.70	12.49	47.06	15.04	62.49	10.02
AGMM(Wu et al. 2023)	71.21	2.73	57.79	12.52	46.64	15.77	62.45	10.25
GeoCoBox (Ours)	**76.22**	**2.08**	**62.24**	**11.97**	**52.14**	**13.46**	**63.82**	<u>9.67</u>

Table 2: We replace backbone 3D Unet by 3D Unetr, while maintaining other setting.

epochs train only the AD-CAM branch, and the remaining epochs also train the GCL branch. The contrastive head is pre-trained with a batch size of 2 for 50 epochs.

All these datasets have pixel-wise annotations. We utilize annotations to obtain the corresponding box annotations. We only use box annotations during the training stage, while we evaluate the performance by pixel-wise annotations in the inference stage. For all annotated tumors, following the previous studies (Zhai et al. 2023; Zhao et al. 2024), we crop each tumor region from the entire CT image into a patch of size $96 \times 96 \times 96$ for training and testing. There is no need to consider the scenario of missed tumor diagnoses.

Evaluation Metrics

Following previous studies(Du et al. 2023; Hu et al. 2021), we binarize the segmentation maps using a threshold of 0.5 to obtain binary masks. We use two standard metrics: the Dice Similarity Coefficient (DSC), expressed as a percentage, and the 95% percentile Hausdorff Distance (HD95). The DSC measures region similarity, focusing on object consistency. The HD95 calculates the maximum boundary distance between the segmentation and the pixel-wise mask. A higher DSC indicates better overlap with the ground truth labels, while a lower HD95 signifies higher segmentation accuracy.

Comparison with Other Methods

Quantitative Results As illustrated in Table 1, we first report the quantitative results of four datasets (LIDC-IDRI, KiTS19, MSD-Lung, and MSD-Pancreas). We compare GeoCoBox with other seven state-of-the-art methods, which contain three loss optimization methods in the natural image: Box2Mask(Li et al. 2024), BoxInst(Tian et al. 2021), BBTP(Wang and Xia 2021); one medical contrastive learning method: GPCS(Du et al. 2023); and three pseudo-labeling methods in both natural and medical domains: RTNet(Lu et al. 2024), GazeMedSeg(Zhong et al. 2024), and AGMM(Wu et al. 2023). Our method aims to leverage tumor priors to form auxiliary constraints and fully use geometric contrastive learning techniques to accurately learn the embedding relationships between pixels, which offers an advantage over the previous contrastive learning method. Moreover, loss optimization approaches may encounter the issue of low contrast in medical images. When obtaining pixel relationships before constructing tumor priors, attention may be diverted to non-edge information. On the other hand, pseudo-labeling methods may suffer from the problem of low-quality pseudo-labels during the learning process. Our method incorporates the strengths of previous works while minimizing their limitations, achieving the best performance across all four datasets.

To validate the robustness of our method, we replace the

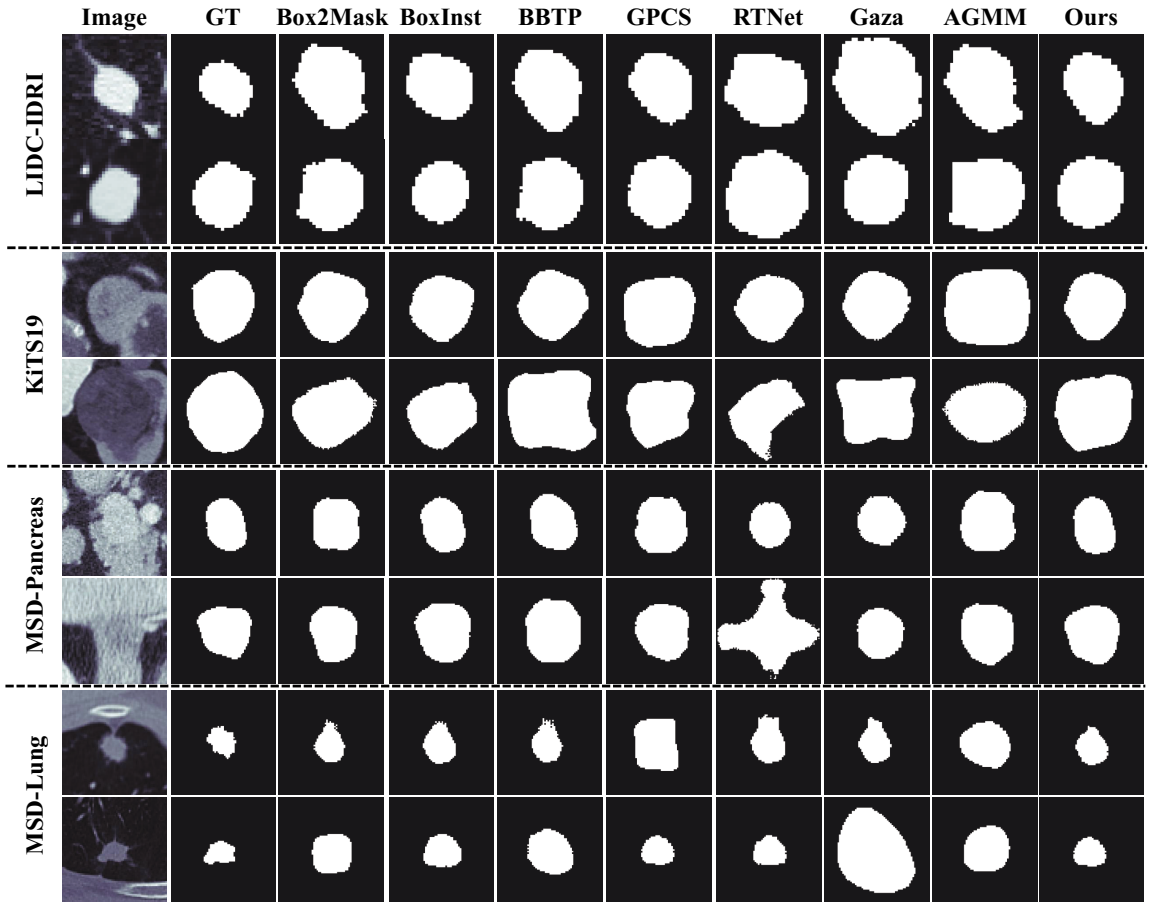

Figure 4: Visualization results of different weakly-supervised methods using 3D U-Net as the backbone on four datasets. It is evident that our method demonstrates superior alignment with the ground truth in terms of both predicted shape and size when compared to other state-of-the-art (SOTA) methods.

AD-CAM				GCL			GCL			GCL		
Soft Layer	Seed Layer	DSC↑	HD95↓	τ	DSC↑	HD95↓	δ	DSC↑	HD95↓	Similarity	DSC↑	HD95↓
SCE	BCE	68.25	3.23	4	73.21	2.73	10HU	72.89	2.73	SSIM	74.10	2.29
SCE	Dice	73.92	2.44	8	**76.09**	**2.05**	20HU	**76.09**	**2.05**	MSE	75.60	2.10
SM	Dice	**76.09**	**2.05**	16	74.32	2.42	40HU	75.22	2.96	CosSim	**76.09**	**2.05**

Table 3: Ablation experiments on the selection of the loss function for the AD-CAM component, τ and δ in the GCL component, and the similarity strategy within the GCL component. The experiments are conducted on the LIDC-IDRI dataset.

backbone from 3D U-Net with the also widely used 3D U-NETR(Hatamizadeh et al. 2022). The results in Table 2 demonstrate that the proposed GeoCoBox outperforms all compared methods in most tasks, except for the HD95 in the MSD-Pancreas dataset, where it slightly underperforms compared to RTNet.

Qualitative Results To further demonstrate the superiority of our method, we present the qualitative comparison with seven other state-of-the-art weakly-supervised segmentation methods in Fig. 4. It can be observed that the proposed GeoCoBox can accurately identify subtle boundaries while remaining consistent with the complex shape of the GT.

Ablation Study

Effectiveness of the Key Components To verify the effectiveness of two key components of GeoCoBox, AD-CAM and GCL, we train three variants of GeoCoBox by disabling AD-CAM or GCL. We also include a baseline with the same basic experimental setting as ours, denoted as Vanilla. The segmentation results on the LIDC-IDRI dataset are presented in Fig. 5. This suggests that removing either component leads to a decrease in performance on both metrics, though the results still outperform the Vanilla

Methods	DSC↑	HD95↓
Box2Mask(Li et al. 2024)	50.31	17.42
BoxInst(Tian et al. 2021)	52.69	15.98
BBTP(Wang and Xia 2021)	50.88	17.84
GPCS(Du et al. 2023)	51.24	16.48
RTNet(Lu et al. 2024)	49.87	17.01
GazeMedSeg(Zhong et al. 2024)	47.25	16.05
AGMM(Wu et al. 2023)	50.80	16.74
Ours	**54.87**	**14.35**

Table 4: Generalization analysis of the proposed model.

model. This finding is reasonable and indicates that the two components complement each other in addressing the box-supervised task.

Hyperparameter and Strategy Selection To further validate the effectiveness of AD-CAM and GCL, we employ three different loss functions for training the AD-CAM, and we choose three refining scales τ and similarity strategies for GCL. As indicated in Table 3, for AD-CAM, we follow (Chen et al. 2022) using SCE and BCE loss functions in the first row. We further consider the imbalance between the foreground and background in medical images. We replace the BCE loss with Binary Dice loss and replace the SCE loss with SM loss, which led to smoother tumor prediction results, thereby achieving an improved performance of 6.7% DSC and 0.8 HD95. For GCL, using $\tau = 8$, $\delta = 20HU$, and the cosine similarity shows the best performance, respectively. Due to the prevalence of small tumors in the LIDC-IDRI dataset, an excessively large correction scale may incorporate more blood vessels with grayscale values similar to those of the tumors, while an overly small scale may overly rely on the coarse results provided by the AD-CAM, leading to decreased segmentation performance. Moreover, cosine similarity, a commonly used similarity metric in contrastive learning, proves more effective than MSE and SSIM in the medical imaging domain.

Analysis of the Data Robustness We also validate the robustness of our method on data from different domains. Specifically, we apply the model trained on the LIDC-IDRI dataset and then validate it on the MSD-Lung dataset. As shown in Table 4, our method demonstrates good data generalization. Notably, the results obtained on the MSD-Lung

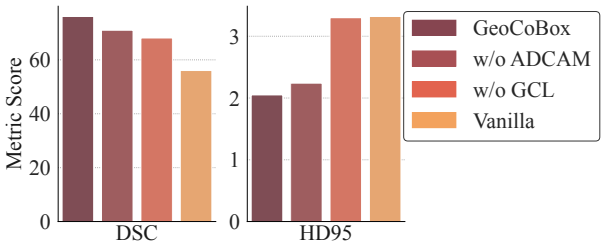

Figure 5: Ablation studies on the proposed components on the LIDC-IDRI dataset.

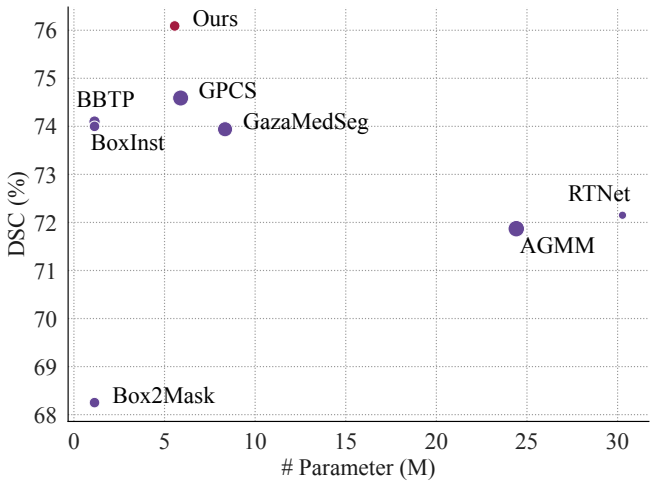

Figure 6: Analysis of the model parameters (Million), training time (epoch), and DSC scores (%). For each data point, the size of the bubble reflects the training time: a larger bubble indicates a longer training time, and vice versa.

dataset are better than those achieved when training directly on the MSD-Lung dataset itself.

Model Efficiency We compare the number of parameters, DSC scores, and training time per epoch among all methods in Fig. 6. It is worth noting that all methods have less than a 40-millisecond difference in inference time for one sample, so we do not compare inference times in this paper. Although our method requires more parameters than loss optimization methods used in natural images, it costs fewer parameters than other medical image methods. Notably, our method requires fewer parameters than the most similar approach, GPCS, due to the computational efficiency of GCL, which only refines the center and boundary points. Additionally, our method achieves the best DSC performance while being one of the methods with the shortest training time.

Conclusion

In this paper, we propose the GeoCoBox that can achieve high-quality 3D tumor segmentation with only box annotations. The core idea of GeoCoBox is to explicitly combine the anatomic prior and geometric pixel embeddings provided by the two proposed components. With the proposed modules, we perform well on four public medical datasets and significantly improve the state-of-the-art. Our method targets small objects in medical images, aiming to offer insights that could benefit research in other domains as well. While our approach demonstrates promising results, further refinement of the contrastive head architecture may yield additional performance gains.

Acknowledgments

This work was supported in part by the Major Program of the National Natural Science Foundation of China under Grant 62172289, in part by the Key Research and

Development Program of the Department of Science and Technology of the Tibet Autonomous Region under Grant XZ202402ZY0003, and in part by the Clinical Medical Research Promotion Program of China Medical Foundation under Grant 2024CMFA10.

References

Alonso, I.; Sabater, A.; Ferstl, D.; Montesano, L.; and Murillo, A. C. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8219–8228.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1): 4128.

Armato III, S. G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M. F.; Meyer, C. R.; Reeves, A. P.; Zhao, B.; Aberle, D. R.; Henschke, C. I.; Hoffman, E. A.; et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 38(2): 915–931.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, Z.; Wang, T.; Wu, X.; Hua, X.-S.; Zhang, H.; and Sun, Q. 2022. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 969–978.

Cheng, T.; Wang, X.; Chen, S.; Zhang, Q.; and Liu, W. 2023. Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3145–3154.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, 1635–1643.

Du, H.; Dong, Q.; Xu, Y.; and Liao, J. 2023. Weakly-supervised 3D medical image segmentation using geometric prior and contrastive similarity. *IEEE Transactions on Medical Imaging*, 42(10): 2936–2947.

Ganaye, P.-A.; Sdika, M.; and Benoit-Cattin, H. 2018. Semi-supervised learning for segmentation under semantic constraint. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, 595–602. Springer.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 574–584.

Heller, N.; Isensee, F.; Maier-Hein, K. H.; Hou, X.; Xie, C.; Li, F.; Nan, Y.; Mu, G.; Lin, Z.; Han, M.; et al. 2021. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical image analysis*, 67: 101821.

Hsu, C.-C.; Hsu, K.-J.; Tsai, C.-C.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in neural information processing systems*, 32.

Hu, Q.; Yi, Z.; Zhou, Y.; Huang, F.; Liu, M.; Li, Q.; and Wang, Z. 2024. MonoBox: Tightness-free Box-supervised Polyp Segmentation using Monotonicity Constraint. *arXiv preprint arXiv:2404.01188*.

Hu, X.; Zeng, D.; Xu, X.; and Shi, Y. 2021. Semi-supervised contrastive learning for label-efficient medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 481–490. Springer.

Jin, C.; Guo, Z.; Lin, Y.; Luo, L.; and Chen, H. 2023. Label-efficient deep learning in medical image analysis: Challenges and future directions. *arXiv preprint arXiv:2303.12484*.

Kulharia, V.; Chandra, S.; Agrawal, A.; Torr, P.; and Tyagi, A. 2020. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, 290–308. Springer.

Lan, S.; Yu, Z.; Choy, C.; Radhakrishnan, S.; Liu, G.; Zhu, Y.; Davis, L. S.; and Anandkumar, A. 2021. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3406–3416.

Li, W.; Liu, W.; Zhu, J.; Cui, M.; Hua, R. Y. X.; and Zhang, L. 2024. Box2mask: Box-supervised instance segmentation via level-set evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Y.; Liu, Y.; Huang, L.; Wang, Z.; and Luo, J. 2022. Deep weakly-supervised breast tumor segmentation in ultrasound images with explicit anatomical constraints. *Medical image analysis*, 76: 102315.

Liu, Y.; Sun, K.; Tang, C.; Qian, Y.; and Li, X. 2025. TPDepth: Leveraging Text Prompts with ControlNet to Boost Diffusion-based Depth Estimation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 4290–4299.

Lu, J.; Deng, J.; and Zhang, T. 2024. BSNet: Box-Supervised Simulation-assisted Mean Teacher for 3D Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20374–20384.

Lu, X.; Cui, Z.; Sun, Y.; Khor, H. G.; Sun, A.; Ma, L.; Chen, F.; Gao, S.; Tian, Y.; Zhou, F.; et al. 2024. Better Rough than Scarce: Proximal Femur Fracture Segmentation with Rough Annotations. *IEEE Transactions on Medical Imaging*.

Shao, Y.; and Fang, T. 2025. Alzheimer's disease detection using co-attention mechanism for acoustic and asr-

transcribed text features. In *Proc. Interspeech 2025*, 5673–5677.

Shao, Y.; Mei, B.; Tan, C.; Huo, H.; and Fang, T. 2025. MoTAS: MoE-Guided Feature Selection from TTS-Augmented Speech for Enhanced Multimodal Alzheimer's Early Screening. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 8497–8505.

Tang, H.; Zhang, C.; and Xie, X. 2019. Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, 266–274. Springer.

Tian, Z.; Shen, C.; Wang, X.; and Chen, H. 2021. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5443–5452.

Wang, F.; Lin, Y.; Yang, L.; Li, H.; Gu, M.; Zhu, M.; and Qu, H. 2024. Outlinespark: Igniting ai-powered presentation slides creation from computational notebooks through outlines. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–16.

Wang, J.; and Xia, B. 2021. Bounding box tightness prior for weakly supervised image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 526–536. Springer.

Wu, L.; Zhong, Z.; Fang, L.; He, X.; Liu, Q.; Ma, J.; and Chen, H. 2023. Sparsely annotated semantic segmentation with adaptive gaussian mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15454–15464.

Wu, L.; Zhuang, J.; and Chen, H. 2024. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22873–22882.

Yi, L.; Zhang, L.; Xu, X.; and Guo, J. 2022. Multi-label softmax networks for pulmonary nodule classification using unbalanced and dependent categories. *IEEE Transactions on Medical Imaging*, 42(1): 317–328.

Yi, L.; Zhang, L.; Zhao, K.; and Xu, X. 2025. Learning from certain regions of interest in medical images via probabilistic positive-unlabeled networks. *Medical Image Analysis*, 103745.

Zhai, S.; Wang, G.; Luo, X.; Yue, Q.; Li, K.; and Zhang, S. 2023. Pa-seg: Learning from point annotations for 3d medical image segmentation using contextual regularization and cross knowledge distillation. *IEEE transactions on medical imaging*, 42(8): 2235–2246.

Zhang, G.; Wang, Y.; Yan, R.; and Fu, X. 2025. MAFM-Gaze: Multi-scale adaptive feature modulation for in-vehicle gaze estimation. *Displays*, 103226.

Zhao, X.; Li, Z.; Luo, X.; Li, P.; Huang, P.; Zhu, J.; Liu, Y.; Zhu, J.; Yang, M.; Chang, S.; et al. 2024. Ultrasound Nodule Segmentation Using Asymmetric Learning with Simple Clinical Annotation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhao, X.; Vemulapalli, R.; Mansfield, P. A.; Gong, B.; Green, B.; Shapira, L.; and Wu, Y. 2021. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10623–10633.

Zhong, Y.; Tang, C.; Yang, Y.; Qi, R.; Zhou, K.; Gong, Y.; Heng, P. A.; Hsiao, J. H.; and Dou, Q. 2024. Weakly-supervised Medical Image Segmentation with Gaze Annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 530–540. Springer.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, Y.; Li, Z.; Bai, S.; Wang, C.; Chen, X.; Han, M.; Fishman, E.; and Yuille, A. L. 2019. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10672–10681.

Zhou, Y.; Xu, H.; Zhang, W.; Gao, B.; and Heng, P.-A. 2021. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7036–7045.