Versatile Medical Image Segmentation Learned from Multi-Source Datasets via Model Self-Disambiguation

Xiaoyang Chen Hao Zheng Yuemeng Li Yuncong Ma Liang Ma Hongming Li
Yong Fan
University of Pennsylvania
{xiaoyang.chen,yong.fan}@pennmedicine.upenn.edu

Abstract

A versatile medical image segmentation model applicable to images acquired with diverse equipment and protocols can facilitate model deployment and maintenance. However, building such a model typically demands a large, diverse, and fully annotated dataset, which is challenging to obtain due to the labor-intensive nature of data curation. To address this challenge, we propose a cost-effective alternative that harnesses multi-source data with only partial or sparse segmentation labels for training, substantially reducing the cost of developing a versatile model. We devise strategies for model self-disambiguation, prior knowledge incorporation, and imbalance mitigation to tackle challenges associated with inconsistently labeled multi-source data, including label ambiguity and modality, dataset, and class imbalances. Experimental results on a multi-modal dataset compiled from eight different sources for abdominal structure segmentation have demonstrated the effectiveness and superior performance of our method compared to state-of-the-art alternative approaches. We anticipate that its cost-saving features, which optimize the utilization of existing annotated data and reduce annotation efforts for new data, will have a significant impact in the field.

1. Introduction

Medical image segmentation plays a pivotal role in disease diagnosis [1, 8, 25, 42], treatment planning [36], and biomedical research [5, 6, 12, 23]. Models, tailored for specific applications, imaging modalities, and distinct anatomical regions, are ubiquitous and attract considerable attention [37]. Nonetheless, they often exhibit limited robustness and generalizability, largely due to insufficient training data. Additionally, the necessity to develop separate segmentation models for different organs and modalities poses scalability challenges, causing inefficient resource utilization and escalating development and maintenance costs.

Versatile image segmentation models show potential in overcoming the limitations of their specialized counterparts. However, their training typically requires a large, diverse, and fully annotated dataset, incurring high costs in data curation and annotation. As a result, only small-scale datasets are usually available, with annotations covering only a portion of anatomical structures or image slices, resulting in partial or sparse segmentation labels [38]. These datasets are typically curated by annotators focusing on labeling specific structures of interest while treating others as background. However, such selective annotation introduces ambiguity when interpreting unannotated regions, impeding the efficacy of image segmentation methods that rely on complete annotations. Specialized strategies are thus essential to effectively utilize datasets with ambiguous labels.

Multi-head segmentation models obviate the issue of labeling ambiguity by designing a distinct decoder for different datasets [4, 18]. However, their inefficient memory usage hampers scalability. Dynamic models, such as the class-conditioned model [9], DoDNet [41], and CLIP-driven [24], adeptly address partial labeling through conditional segmentation, enabling adjustments in their outputs for specific tasks. Nevertheless, dynamic models face their own challenges, such as training complexities, inefficiencies in inference due to multiple forward passes, and limitations in fully exploiting the benefits of fine-grained annotations, as class-specific parameters are optimized separately. Semi-supervised segmentation methods generate pseudo-labels for unannotated anatomical structures to facilitate conventional loss computation [15, 43]. However, these methods require fully annotated data for initial fully supervised training, and the incorporation of inaccurate pseudo-labels in later training stage may degrade the model performance. Background modeling methods [11, 35] dynamically compute losses for unannotated voxels to mitigate semantic drifts in partial annotations. Nevertheless, their requirement for fully annotated data limits their applicability in challenging scenarios. Notably, all these existing methods are unable to utilize sparsely labeled data (cf. Fig. 1(c)).

In this study, we present a weakly-supervised approach for medical image segmentation, utilizing a large and diverse dataset with incomplete labeling from multiple sources. Our method utilizes a model self-disambiguation mechanism to tackle labeling ambiguity in both partially and sparsely annotated data. This is achieved by introducing two ambiguity-aware loss functions. Additionally, by leveraging prior knowledge of optimal predictions, we integrate a regularization term into the objective function. This helps reduce uncertainty in model predictions, particularly for challenging and unannotated voxels, thereby expediting convergence. To address imbalances in multi-source data, we propose a hierarchical sampling strategy. Our approach facilitates training a single versatile model using multi-source datasets and enables efficient inference in a single forward pass, predicting all anatomical structures simultaneously. Our contributions are three-fold:

- We propose a weakly-supervised approach that leverages partially and sparsely labeled data to address data limitations in medical image segmentation. Remarkably, our approach exhibits impressive versatility and self-disambiguation capabilities, holding great promise for enhancing label efficiency and reducing the costs associated with model development, deployment, and maintenance.
- We employ hierarchical sampling to account for the imbalance issues in multi-source datasets and incorporate prior knowledge to improve the model performance.
- We showcase the proposed method's effectiveness on a multi-modality dataset of $2,960$ scans from eight distinct sources for abdominal organ segmentation. Our approach demonstrates substantially improved efficiency and effectiveness compared to state-of-the-art alternative methods.

2. Related Work

Category-specific models. Developing separate models for different anatomical structures using annotated data from various sources is a straightforward strategy for leveraging multi-source datasets. However, this method is computationally complex and inefficient, as it requires training multiple models and processing test images through each model during inference. Additionally, it fails to capitalize on the benefits of fine-grained segmentation, which could improve feature representations and overall performance [22].

Multi-head models. Multi-head models [4, 18] share an encoder but have separate decoders for each dataset. Yet, redundant structures in multiple decoders hinder scalability, and training decoders with limited and less diverse data may degrade model generalization.

Dynamic models. Dynamic models like the class-conditioned model [9], DoDNet [41], CLIP-driven model [24], and Hermes [13] utilize a unified model with task-adjustable outputs via a controller. However, they handle only one segmentation task at a time, causing inefficiencies during inference and limiting their exploitation of fine-grained segmentation benefits.

Semi-supervised segmentation. Semi-supervised segmentation methods [15, 43] tackle partial labeling by generating pseudo-labels for unannotated anatomies, incorporating additional regularizations like anatomy size [43] and inter-model consistency [15] to stabilize training. However, inaccuracies in pseudo-labels can impair model performance. Moreover, the necessity of a fully labeled dataset for pre-training [43] and training multiple single-anatomy segmentation models [15] may limit practical applicability.

Weakly-supervised segmentation. Weakly-supervised segmentation methods utilize various forms of weak supervision, including image-level labels [16], bounding boxes [19], points [10, 40], scribbles [28, 38], and incomplete annotations [2, 11, 32, 35]. Our work falls within this broad category, focusing on learning from ambiguous data labeled partially and sparsely. However, in contrast to existing methods that utilize image-level labels, bounding boxes, points and scribbles, none of which can attain comparable segmentation performance to voxel-wise supervision, or those tailored for training with partial labels, which struggle to leverage sparse labels, or those trained on data with sparse labels, requiring clear background definitions and annotations, our approach is designed to handle both partial and sparse labels, achieving highly competitive performance to fully supervised segmentation, even in the absence of background annotations.

Background modeling. Background modeling methods [11, 35] address label ambiguity in partially labeled data by dynamically calculating the loss for unannotated voxels. These methods assume complete annotations of all identified organs within the volume and consolidate the predictions for all categories, excluding the annotated ones, into a distinct class during loss computation. Our approach bears similarities to these methods but offers enhanced capability in handling sparsely annotated data by relaxing labeling constraints. Notably, these methods still require fully annotated data during training, limiting their applicability in practical scenarios where such data is unavailable. In contrast, our method maintains effectiveness even when all training images are incompletely annotated.

Segment anything model. The "segment anything" model [20] and its variants, such as MedSAM [29] and SAM-Med2D [7], share the same goal as our work, aiming for a universal segmentation model applicable to various images and objects. However, these models presume the availability of a substantially large labeled dataset and do not attempt to handle practical challenges like label incompleteness and ambiguity. Moreover, these models are typically designed to generate segmentation results automatically without providing their semantic labels and are more suited for interactive use, requiring user input, such as a bounding box.

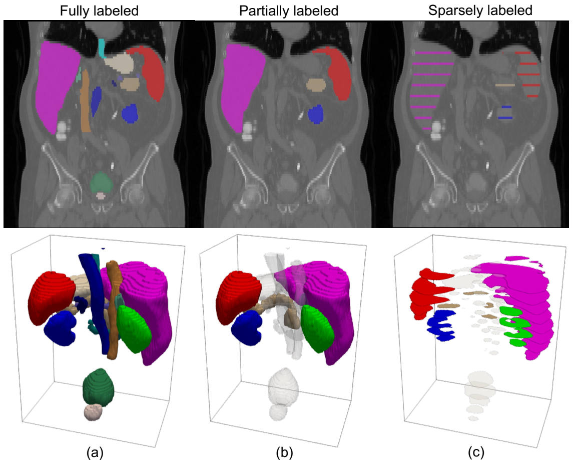


Figure 1. Illustrations of (a) fully labeled, (b) partially labeled, and (c) sparsely labeled images. The fully labeled image contains annotations for all anatomical structures of interest, the partially labeled image includes labels for a subset, and the sparsely labeled image provides annotations for only a fraction of the slices and structures. Note that annotated structures are fully marked within a particular volume (b) or slice (c).

3. Methods

This section begins by introducing the motivation, objective and scope of our study. It then offers an overview of our proposed framework, followed by a detailed explanation of the key strategies employed in this research.

3.1. Motivation, Objective & Scope

Given the challenges of obtaining a large, diverse, and fully annotated dataset for training versatile medical image segmentation models, as well as the increased accessibility and cost-effectiveness of weakly labeled data compared to the fully labeled data, our study pursues a cost-effective alternative utilizing two forms of weak supervision: partially labeled data and sparsely labeled data. Fig. 1 illustrates the differences between these data types, emphasizing variations in their annotation scopes and details.

It is important to clarify that the term "sparsely labeled data" here specifically refers to images with per-voxel annotations, rather than other types of weak annotations such as image-level labels, point annotations, scribbles, and bounding boxes, which are often categorized as sparse annotations in other studies. Nonetheless, our definition allows for flexibility: annotations for different slices and structures are independent, meaning that a structure annotated in one slice does not have to be annotated in other slices. Additionally, it is crucial to emphasize that sparsely labeled data encompasses a broader spectrum, with partially labeled data representing a specific category within it. By using these distinct terminologies, we highlight the differences between existing methods, primarily tailored for partially labeled data, and our approach, which accommodates a wider range of weakly labeled data. Our method excels at better data utilization and has the potential to enhance data accessibility.

3.2. Overview

As illustrated in Fig. 2, our approach employs a hierarchical sampling technique to generate training examples from multi-source multi-modality datasets with ambiguous annotations. A 3D variant of TransUNet [3] (3D TransUNet) is adopted as the base network for extracting per-voxel feature representations from the input. These representations are then processed by a segmentation head to produce multi-channel predictions. To address label ambiguity and ensure effective training, we integrate ambiguity-aware losses. Moreover, we incorporate prior knowledge to regularize the model training.

3.3. Model Self-disambiguation

When annotating medical images, it is a common practice for annotators to focus solely on labeling the anatomical structures of interest. Thus, a significant portion of voxels remains unannotated (with a default value of 0) and is interpreted as background for each image. In a data collection comprising partially and/or sparsely labeled images from diverse sources, unannotated voxels in different images may contain various anatomical structures, leading to semantic ambiguity/drift within the background class. This semantic ambiguity presents significant challenges for fully supervised approaches due to conflicting supervision.

In this study, we tackle the challenges posed by semantic drift by computing the loss for unannotated voxels adaptively, considering both the possible categories for the unannotated voxels and the label type (*i.e.*, partially labeled or sparsely labeled). Without loss of generality, let's consider a scenario where there are a total of N anatomical structures of interest, denoted by Φ_N, and each training example has only a fraction of its slices annotated. In each annotated slice, the annotation may cover only a subset of M out of N structures, where $1 \leq M \leq N$. Generally, this subset can comprise any combination and be denoted as $\Phi_M = \{i_1, i_2, \dots, i_M\}$, where $1 \leq i_p < i_q \leq N$ for any $p < q$. For the annotated voxels in a given slice, their labels are definitive and offer clear supervision. However, the true labels for the unannotated voxels in the same slice are unknown. It is only certain that these voxels may belong to either the "real" background category or any class in $\Phi_N \backslash \Phi_M$, representing the difference between sets Φ_N and Φ_M, *i.e.*, $\{x \mid x \in \Phi_N \text{ and } x \notin \Phi_M\}$. Therefore, we adaptively adjust the loss calculation for the unannotated voxels to accommodate label ambiguity. We adopt the following ambiguity-aware focal cross-entropy loss ($\mathcal{L}_{\text{focal\_ce}}$)

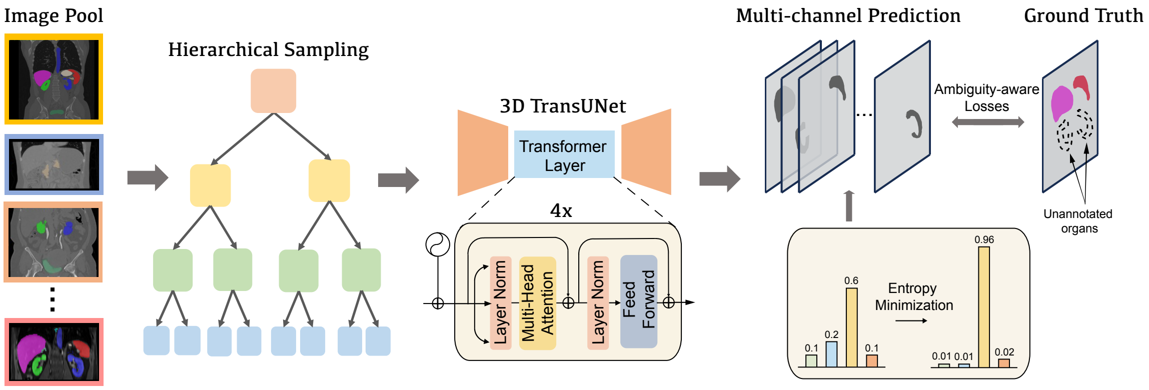


Figure 2. Overview of our approach. It trains a model by using hierarchical sampling for training example generation, 3D TransUNet as its base network, two ambiguity-aware losses and a prior knowledge-based entropy minimization regularization term for guidance.

and dice loss ($\mathcal{L}_{\text{dice}}$), calculated slice-wise, as the objective:

$$\mathcal{L}_{\text{focal\_ce}} = \frac{1}{N_v} \sum_{c\in\{0\}\cup\Phi_M} \sum_{i=1}^{N_v} \mathbb{1}_{y_i=c}(1-\tilde{p}_{ic})^2 \log \tilde{p}_{ic}, \quad (1)$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{1}{|\Phi_M|+1} \sum_{c\in\{0\}\cup\Phi_M} \frac{2\cdot \text{TP}_c + \epsilon}{2\cdot \text{TP}_c + \text{FP}_c + \text{FN}_c + \epsilon}, \quad (2)$$

where $\mathbb{1}$ denotes an indicator function, $|\cdot|$ is the cardinality, N_v represents the number of pixels in the slice, $\text{TP}_c = \sum_{i=1}^{N_v} \tilde{p}_{ic}\tilde{y}_{ic}$, $\text{FP}_c = \sum_{i=1}^{N_v} \tilde{p}_{ic}(1-\tilde{y}_{ic})$ and $\text{FN}_c = \sum_{i=1}^{N_v}(1-\tilde{p}_{ic})\tilde{y}_{ic}$ are the soft values for the true positive, false positive and false negative respectively, ϵ is set to 1 to avoid division by 0,

$$\tilde{p}_{ic} = \begin{cases} p_{ic}, & c \in \Phi_M, \\ \sum_{j\notin\Phi_M} p_{ij}, & c = 0, \end{cases} \quad (3)$$

$$\tilde{y}_{ic} = \begin{cases} y_{ic}, & c \in \Phi_M, \\ \sum_{j\notin\Phi_M} y_{ij}, & c = 0, \end{cases} \quad (4)$$

p_{ic} and y_{ic} represent the c-th element of the probability vector and the one-hot encoded vector for the expert label for the i-th pixel, respectively.

Unlike methods that simply treat unannotated voxels as background, potentially misleading the model, or those that overlook voxels with ambiguous labels in the loss calculation, leading to incorrect predictions for voxels not belonging to any specific structure, our ambiguity-aware losses enable the model to self-disambiguate during training and infer the correct labels for all voxels.

It is noteworthy that for partially labeled data, the loss computation can be simplified. The ambiguity-aware losses can be calculated for each volume rather than slicewise, with N_v representing the number of voxels, and Φ_M denoting the set of annotated structures within the volume.

3.4. Prior Knowledge Incorporation

We have the prior knowledge that each voxel/pixel corresponds to a single label in multi-class medical image segmentation tasks. When confronted with challenging and unannotated voxels, the model encounters difficulty in determining their classes, leading to high-entropy predictions. Our hypothesis is that reducing this uncertainty can improve the differentiation between categories and accelerate a more reliable convergence during the optimization process. We thus incorporate this prior knowledge into the training process, encouraging the model to produce more confident and informative predictions. This is achieved by regularizing the model training to minimize the Shannon entropy below.

$$\mathcal{L}_{\text{reg}} = -\frac{1}{N_v} \sum_{i=1}^{N_v} \sum_{c=0}^{N} p_{ic} \log p_{ic}, \quad (5)$$

Notably, this regularizer is class-agnostic and can be applied to both annotated and unannotated voxels.

3.5. Imbalance Mitigation

Medical image segmentation models often encounter challenges in maintaining consistent performance across diverse domains due to variations in imaging modalities, equipment, imaging protocols, and patient demographics. While aggregating data from multiple sources can enrich training data diversity and bolster model robustness, it may also introduce imbalances at the modality, dataset, and class levels. Neglecting these issues during model training could result in inferior performance, particularly on underrepresented modalities, datasets, and categories.

However, such imbalance issues have not been well addressed in existing methods that utilize multi-source partially labeled data [11, 15, 35, 43]. To tackle these challenges, we propose a hierarchical sampling approach. During training, we initiate the sampling by selecting images based on the type of anatomical structure, thereby narrowing down the number of eligible images. The chosen structure determines the location of the training image patch center. Subsequently, we conduct a secondary sampling based on the modality of the medical images within the subset of images from the first level, enabling us to focus on images that belong to the chosen modality. Next, we draw a sample from the candidate pool based on the dataset of origin for each image, ensuring equitable treatment for images from various sources. Finally, we select an image from the chosen dataset for training. The proposed strategy enables us to account for the variations across domains, ultimately ensuring a balanced representation of the training data.

3.6. Overall Objective

The overall objective for training ($\mathcal{L}$) is a weighted combination of the uncertainty-aware focal cross-entropy loss ($\mathcal{L}_{\text{focal\_ce}}$), uncertainty-aware dice loss ($\mathcal{L}_{\text{dice}}$), and the Shannon entropy minimization regularization term ($\mathcal{L}_{\text{reg}}$):

$$\mathcal{L} = \mathcal{L}_{\text{focal\_ce}} + \mathcal{L}_{\text{dice}} + \lambda\mathcal{L}_{\text{reg}}. \tag{6}$$

where the hyper-parameter λ is set to 3 for training examples without annotations to mitigate the null effects of $\mathcal{L}_{\text{focal\_ce}}$ and $\mathcal{L}_{\text{dice}}$, and 1 otherwise.

4. Experiments and Results

4.1. Experiment Setup

Dataset. We curated a dataset of $2,960$ volumetric images from eight sources, including seven public datasets (AbdomenCT-1K (AbCT-1K) [30], AMOS [17], BTCV [21], FLARE 2022 (FLARE22) [31], NIH pancreas (NIH-Pan)[33], TotalSegmentator (TotalSeg) [39], and WORD [27]) and one private dataset (Urogram). Notably, the AMOS dataset is multi-modality, featuring 300 labeled CT and 60 labeled MRI images, which we segregated into two subsets, AMOS-CT and AMOS-MRI. We demonstrated the model's self-disambiguation capability by segmenting 16 abdominal structures, namely spleen (Sp), right kidney (RK), left kidney (LK), gallbladder (GB), esophagus (Eso), liver (L), stomach (St), aorta (A), postcava (PC), pancreas (Pan), right adrenal gland (RAG), left adrenal gland (LAG), duodenum (Duo), bladder (B), prostate/uterus (PU), and portal vein and splenic vein (PSV). Although TotalSeg and WORD provided masks for anatomical structures beyond our scope, we retained only the pertinent ones. In WORD, RK and RAG are assigned with the same labels as LK and LAG, respectively. We thus partitioned segments for kidney and adrenal gland into connected components and automatically assigned labels to the largest two, assisted by a model trained without WORD. The same was applied to AbCT-1K to separate RK and LK. For cases with only one kidney or adrenal gland, we manually verified and assigned the correct labels. Approximately 90% of the images were selected for training, with the remainder reserved for evaluation purposes. Details about each dataset are outlined in Fig. 3. Note that we have removed corrupted images, and **no** images used in this study are fully annotated.

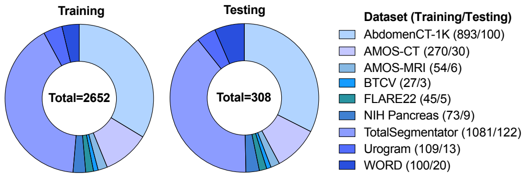
Table 1. Method performance comparison.

Method	**Base**	**DSC** [%]
DoDNet [41]	3D TransUNet	83.5±15.9
CLIP-driven [24]	3D TransUNet	83.3±16.3
Ours	Unet++ [44]	87.4±8.5
	Swin UNETR [14]	86.9±8.5
	MedNeXt [34]	88.4±7.3
	3D TransUNet	**88.7±7.0**

Data preprocessing. To facilitate hierarchical sampling, we added a prefix to each image name indicating the dataset it comes from and its modality. For example, a CT image from AMOS initially named "amos_0001.nii.gz" was renamed "amos_ct_amos_0001.nii.gz" after prefixing. Additionally, we standardized all images to lie in a common coordinate system to ease model training with images in varied orientations. All data were resampled to a uniform spacing of $2 \times 2 \times 2$ mm^3. Intensity values in CT images were clipped at -400 and 400 HU, while for MRI images, the clipping was done at the 1st and 99th percentiles of the intensity distribution. Finally, the intensity values were normalized to the range of $[0, 1]$.

Implementation details. PyTorch was used to implement the proposed method. We employed the AdamW optimizer [26] with an initial learning rate of 0.001 and a polynomial learning rate scheduler with a decay of 0.9. Data augmentations, such as random rotation and scaling, were applied during the training process. Unless otherwise specified, the default patch size and number of iterations were set to $112 \times 112 \times 112$ and $200,000$, respectively. Distributed data parallel was used to enhance training efficiency. All experiments were conducted on a single node with 8 NVIDIA Titan Xp GPUs. The effective batch size was 8 for all experiments, and all models were trained from scratch.

Evaluation metric. The Dice Similarity Coefficient (DSC, %) was used for performance evaluation. Since annotations were limited to a subset of anatomical structures in each image, and the region of interest (ROI) varied across images, only those annotated structures within the ROI were included for quantitative evaluation. Notably, the segmentation difficulty varied across different structures, and the number of annotated structures differed among datasets, leading to significant variation in the quantitative values across datasets. It is noteworthy that all the reported performances were obtained with a single model, not through ensemble learning techniques.



(a)

Dataset	Sp	RK	LK	GB	Eso	L	St	A	PC	Pan	RAG	LAG	Duo	B	PU	PSV
AbCT-1K	✓	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
AMOS-CT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
AMOS-MRI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
BTCV	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓
FLARE22	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
NIH-Pan	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
TotalSeg	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
Urogram	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
WORD	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✗

(b)

Figure 3. (a): Training and testing image composition. (b): Annotated anatomical structures in different datasets.

Table 2. Comparison of overall and PU segmentation performance (DSC, %) with varying numbers of datasets used for training. The "3 Sets" experiment exclusively involves AMOS, BTCV, and FLARE22 datasets.

Setting	Overall	PU
3 Sets	83.5±13.9	75.8±21.6
8 Sets	**88.7**±**7.0**	**79.2**±**17.6**

Table 3. Performance with different sampling methods.

Sampling Approach	DSC [%]
Random	87.5±7.7
MDC	87.9±7.7
CMD (Default)	**88.7**±**7.0**

Table 4. Performance (DSC, %) comparison among DoDNet, CLIP-driven, and the proposed method on each anatomical structure.

Method	Base	Sp	RK	LK	GB	Eso	L	St	A	PC	Pan	RAG	LAG	Duo	B	PU	PSV	Average
DoDNet	3D TransUNet	92.6 ±11.9	90.6 ±13.2	89.8 ±13.9	73.3 ±28.3	76.6 ±14.6	93.2 ±15.1	85.8 ±19.8	84.8 ±22.5	82.5 ±19.7	81.6 ±14.3	69.6 ±18.9	72.3 ±17.9	69.6 ±21.3	83.6 ±17.3	59.2 ±29.5	75.7 ±20.5	80.0 ±18.7
CLIP-driven	3D TransUNet	92.5 ±11.8	89.0 ±16.6	90.5 ±13.0	74.5 ±27.7	76.8 ±14.6	93.1 ±15.2	86.2 ±20.0	84.4 ±21.9	81.7 ±21.0	81.6 ±14.1	70.7 ±17.7	73.4 ±17.4	69.3 ±22.7	85.9 ±16.5	68.2 ±30.3	74.5 ±21.7	80.7 ±18.9
Ours	UNet++	94.9 ±6.9	93.1 ±9.0	93.5 ±4.9	78.2 ±23.2	81.5 ±8.6	96.2 ±6.8	90.3 ±13.0	90.7 ±11.2	85.4 ±12.7	84.1 ±9.4	74.1 ±12.5	75.2 ±**13.3**	73.9 ±17.9	88.2 ±15.0	76.0 ±23.6	75.8 ±17.0	84.5 ±12.8
	Swin UNETR	94.7 ±7.3	92.8 ±9.7	93.0 ±6.4	76.4 ±**22.9**	80.0 ±9.0	96.1 ±6.6	90.0 ±12.5	90.4 ±10.1	85.4 ±11.6	83.0 ±9.9	73.1 ±13.1	74.0 ±14.9	72.6 ±17.1	88.4 ±14.3	74.8 ±22.2	74.8 ±16.1	83.7 ±12.7
	MedNeXt	95.0 ±8.9	93.3 ±9.5	93.7 ±6.9	**78.4** ±23.9	**83.2** ±**7.5**	**96.5** ±**6.4**	91.3 ±**12.1**	91.9 ±8.3	86.7 ±12.3	**85.1** ±**8.9**	**75.5** ±**11.3**	76.4 ±13.4	76.3 ±**15.9**	88.9 ±14.2	77.5 ±23.7	**77.6** ±**15.5**	84.5 ±12.4
	3D TransUNet	**95.3** ±**6.6**	**93.8** ±**7.3**	**94.0** ±**5.2**	77.7 ±25.0	82.4 ±9.0	**96.5** ±**6.4**	**91.5** ±12.3	**92.5** ±**6.5**	**87.0** ±**11.2**	**85.1** ±9.3	75.4 ±12.0	**76.5** ±14.6	**76.4** ±16.6	**90.3** ±**13.3**	**79.2** ±**17.6**	**77.6** ±16.5	**85.7** ±**11.9**

4.2. Results on Partially Labeled Data

Current methods are limited to utilizing partially labeled data for training. Therefore, we compared our method and state-of-the-art approaches, DoDNet and CLIP-driven, using exclusively partially labeled data. For fair comparisons, we replaced the base network in their original frameworks with the 3D TransUNet and adopted the same hierarchical sampling approach as ours. Notably, we adjusted DoDNet to produce a single-channel output, since DoDNet was originally designed to predict both anatomical structures and associated tumors concurrently whereas our study focused on the former. Furthermore, both DoDNet and CLIP-driven were unable to utilize images lacking annotations, such as those from TotalSegmentator that contain no relevant anatomical structures under study within their ROIs. Consequently, we excluded those images for training DoDNet and CLIP-driven models.

Main results. Table 1 summarizes the segmentation performance assessed *on a per-subject basis* for DoDNet, CLIP-driven, and the proposed method, which employed different base networks. This evaluation method computed the average DSC of all annotated classes for each testing subject and subsequently averaged them across subjects. During training DoDNet and CLIP-driven, we observed that they exhibited significantly slower convergence rates and thus doubled the training time compared to our proposed method. Our experimental results indicated that, using the same 3D TransUNet as the base network, our approach achieved an impressive average DSC of 88.7% on the testing set, surpassing both DoDNet and CLIP-driven, which attained average DSCs of 83.5% and 83.3%, respectively.

Further insights into performance were gleaned by examining segmentation performance *on each anatomical structure*, as depicted in Table 4. This analysis involved averaging DSCs across individual images with specific structures annotated, revealing the superior segmentation performance of the proposed method over DoDNet and CLIP-driven. Remarkably, our proposed method, employing 3D TransUNet as the base network, achieved an average DSC of 85.7% for individual structures, outperforming DoDNet and CLIP-driven by 5.7% and 5.0%, respectively.

Moreover, we conducted a comparative evaluation of their performance and undertook a visual comparison across each dataset, as illustrated in Table 5 and Fig. 4, which accentuated the consistently superior performance of the proposed method across all datasets.

To evaluate model generalizability to unseen datasets, we additionally trained a model using only AMOS, BTCV, and FLARE22 as training data. As summarized in Table 2,

Table 5. Performance (DSC, %) comparison among DoDNet, CLIP-driven and the proposed method on each dataset.

Method	Base	AbCT-1K	AMOS-CT	AMOS-MRI	BTCV	FLARE22	NIH-Pan	TotalSeg	Urogram	WORD
DoDNet	3D TransUNet	91.3 ±2.9	81.5 ±9.0	80.5 ±9.5	76.0 ±5.4	89.0 ±1.5	82.4 ±4.5	77.6 ±22.5	90.9 ±2.7	79.3 ±4.9
CLIP-driven	3D TransUNet	91.3 ±3.2	82.0 ±9.3	80.8 ±9.3	76.6 ±5.3	89.3 ±1.2	81.9 ±4.0	76.8 ±23.1	90.3 ±3.7	80.6 ±4.2
Ours	UNet++	92.9 ±2.3	84.9 ±6.9	**81.6** ±8.8	**79.4** ±5.9	90.5 ±0.9	83.6 ±5.2	84.2 ±10.5	93.1 ±1.8	83.5 ±4.5
	Swin UNETR	92.8 ±2.0	83.6 ±7.9	80.6 ±6.5	78.8 ±5.9	90.2 ±1.3	82.7 ±5.2	83.4 ±10.1	93.1 ±**1.5**	83.1 ±4.3
	MedNeXt	93.3 ±2.4	**85.6** ±7.2	**81.6** ±**8.1**	78.8 ±6.7	90.8 ±**0.6**	84.2 ±4.2	85.9 ±8.5	**93.8** ±**1.5**	**84.2** ±3.9
	3D TransUNet	**93.5** ±**1.9**	85.5 ±**6.6**	**80.5** ±9.3	78.7 ±**5.6**	**90.9** ±0.7	**84.7** ±**4.4**	**86.4** ±**7.9**	93.6 ±**2.3**	84.1 ±**3.8**

Table 6. Performance (DSC, %) comparison among different sampling methods on each anatomical structure.

Sampling Method	Sp	RK	LK	GB	Eso	L	St	A	PC	Pan	RAG	LAG	Duo	B	PU	PSV	Average
Random	94.8 ±6.7	92.7 ±8.0	93.0 ±6.2	75.0 ±23.7	80.9 ±9.5	96.0 ±6.6	90.0 ±12.0	92.0 ±5.3	85.5 ±11.0	83.6 ±9.0	71.7 ±14.1	73.3 ±16.0	72.5 ±17.4	89.5 ±13.5	74.4 ±24.6	73.8 ±15.9	83.7 ±12.5
MDC	95.0 ±7.8	93.6 ±8.0	93.5 ±6.7	77.8 ±23.2	81.3 ±8.9	96.5 ±6.4	91.1 ±11.2	91.3 ±9.8	86.1 ±12.1	84.6 ±9.0	73.7 ±12.9	74.4 ±15.1	73.2 ±18.0	89.0 ±14.4	75.7 ±23.1	74.9 ±15.3	84.5 ±12.6
CMD (Default)	**95.3** **±6.6**	**93.8** **±7.3**	**94.0** **±5.2**	**77.7** **±25.0**	**82.4** **±9.0**	**96.5** **±6.4**	**91.5** **±12.3**	**92.5** **±6.5**	**87.0** **±11.2**	**85.1** **±9.3**	**75.4** **±12.0**	**76.5** **±14.6**	**76.4** **±16.6**	**90.3** **±13.3**	**79.2** **±17.6**	**77.6** **±16.5**	**85.7** **±11.9**

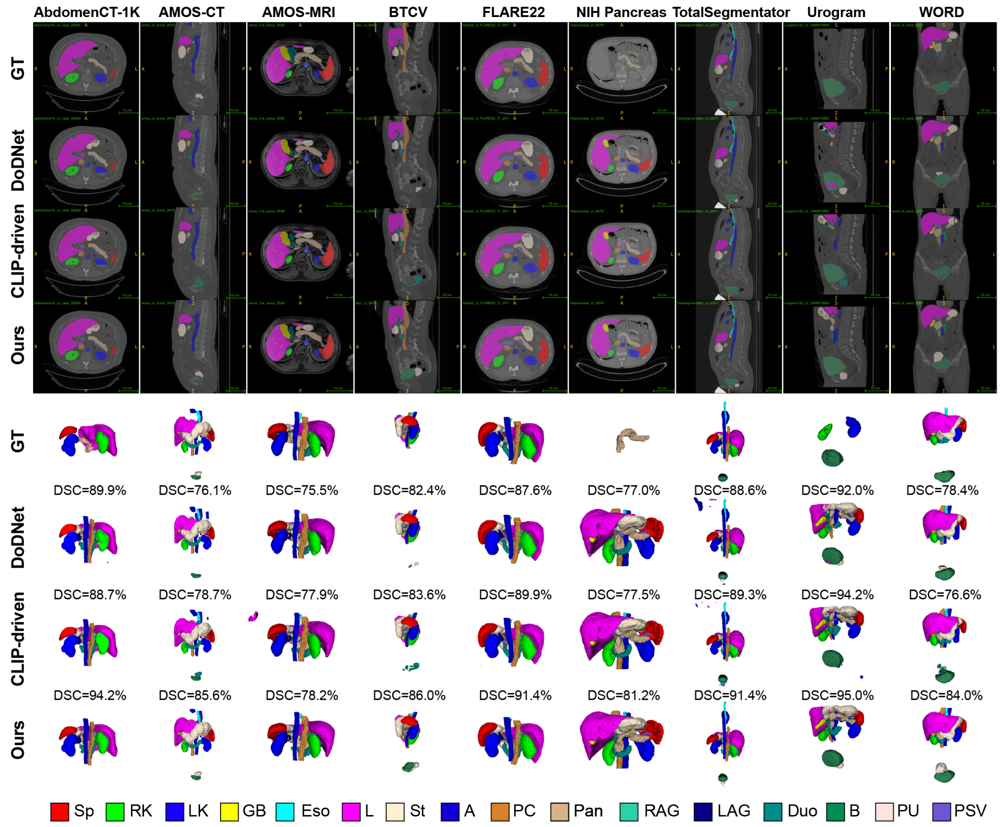


Figure 4. Visual comparison between the ground truth and the predictions generated by DoDNet, CLIP-driven and the proposed method on subjects from different datasets. For a clearer view of detailed differences, zoom in to closely examine the results.

this model achieved an average DSC of 83.5% on the testing set, lower than the one trained with all data, which was expected. Notably, using all eight datasets for training yielded a model with superior prostate/uterus segmentation performance compared to the one trained using only three datasets, despite the additional datasets lacking annotations for the prostate/uterus. This highlights the advantages of fine-grained segmentation.

Table 7. Performance trained with varying numbers of partially labeled images with and without entropy minimization.

Setting	**DSC** [%]
3 Sets w/o reg	82.6±15.2
3 Sets w/ reg	83.5±13.9
8 Sets w/o reg	88.2±**7.0**
8 Sets w/ reg	**88.7**±**7.0**

Table 8. Performance (DSC, %) on sparsely labeled data with different portions of annotated slices.

Setting	**20%**	**100%**
8 Sets (axial)	85.1±11.8	87.8±8.0
8 Sets (sagittal)	86.2±9.0	87.8±7.9
8 Sets (coronal)	86.1±10.3	87.8±7.9

Table 9. Performance with mixed training.

Setting	**DSC** [%]
8 Sets (axial)	87.6±8.3
8 Sets (sagittal)	87.7±8.5
8 Sets (coronal)	87.6±7.6

Additionally, it should be noted that both DoDNet and CLIP-driven demand 16 forward passes to predict the desired anatomical structures, whereas our proposed method achieved the same with just a single forward pass, indicating our method's substantially improved efficiency.

Effect of base network. We compared 3D TransUNet, our custom design, and the default network with other top performers in medical image segmentation, including Unet++ [44], Swin UNETR [14], and MedNeXt-L [34]. All training configurations were consistent across different networks, with the patch size for Swin UNETR adjusted to 96×96×96 due to memory constraints. Results in Table 1 demonstrated that while MedNeXt achieved comparable performance to 3D TransUNet on a per-subject basis, Unet++ and Swin UNETR exhibited inferior performance by over 1% in terms of DSC. A detailed class-wise comparison in Table 4 highlighted 3D TransUNet's stronger performance compared to all others, including MedNeXt. Notably, our approach was largely not affected by the choices of its base network.

Effect of sampling method. To evaluate the benefits of hierarchical sampling, we compared three strategies: 1) sampling following the class→modality→dataset hierarchy (CMD, the default), 2) sampling following the modality→dataset→class hierarchy (MDC), and 3) random sampling that selects an image randomly from the dataset and then chooses a random location within the image volume as the center for training patches. The results presented in Table 3 indicated that our approach outperformed random sampling by 1.2% in overall DSC. While MDC exhibited a 0.4% improvement in DSC over Random, it lagged 0.8% behind CMD. These advantages were particularly noticeable in the comparison on each anatomical structure, especially for smaller structures, such as LAG and RAG, as emphasized in Table 6.

Effect of regularization term. The comparison between models with and without the entropy minimization regularization term is outlined in Table 7. Despite a decrease in performance gain with larger training datasets, consistent enhancements were observed across various dataset sizes. When all 8 datasets were used for training, the addition of the regularization term resulted in a DSC improvement of 0.5%. Notably, when training the model with 3 datasets, including AMOS, BTCV, and FALRE22, a larger improvement of 0.9% in terms of DSC was achieved.

4.3. Results on Sparsely Labeled Data

Our method stands out from existing ones in its capability of handling sparsely labeled data, a critical feature that enhances its applicability in real-world scenarios where the annotation budget is limited and/or data are sparsely labeled. For demonstration purpose, we conducted experiments in which we selectively chose slices from axial, sagittal, or coronal views for training. The experimental results, summarized in Table 8, revealed that models trained with only 20% of slices (evenly spaced) achieved an impressive average DSC ranging from 85.1% to 86.2%, outperforming baseline methods trained with all slices (cf. Table 1). For comparison, we trained three additional models using the same data as in the partially labeled experiments (i.e., trained with 100% slices), but with slice-by-slice loss calculation to simulate sparsely labeled data conditions. These results served as an upper bound and demonstrated the consistent performance of our method across different views.

4.4. Results on Hybrid Data

We conducted experiments using both partially and sparsely labeled data to mimic real-world scenarios. Our mixed training approach utilized AMOS, BTCV, and FLARE22 entirely, and 20% of evenly spaced slices of the other five datasets from axial, sagittal, or coronal views, respectively. Table 9 demonstrates that this hybrid data approach achieved DSCs of 87.6%, 87.7%, and 87.6% for the respective models. In contrast, the model trained solely with 3 partially labeled datasets attained an average DSC of 83.5% (cf. Table 2). Integrating sparsely labeled data notably improved the performance by approximately 4.1%.

5. Conclusions

We have developed a novel weakly-supervised medical image segmentation approach that effectively utilizes multi-source partially and sparsely labeled data for training. Our method addresses data limitations of large, diverse, fully annotated datasets, enhancing label efficiency and reducing annotation efforts through the utilization of weakly annotated data. By integrating strategies for model self-disambiguation, prior knowledge incorporation, and imbalance mitigation, our approach establishes a solid foundation for training versatile and reliable segmentation models.

References

[1] Rilwan Babajide, Katerina Lembrikova, Justin Ziemba, James Ding, Yuemeng Li, Antoine Selman Fermin, Yong Fan, and Gregory E Tasian. Automated machine learning segmentation and measurement of urinary stones on ct scan. *Urology*, 169:41–46, 2022. 1

[2] John-Melle Bokhorst, Hans Pinckaers, Peter van Zwam, Iris Nagtegaal, Jeroen van der Laak, and Francesco Ciompi. Learning from sparsely annotated data for semantic segmentation in histopathology images. 2018. 2

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 3

[4] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 1, 2

[5] Xiaoyang Chen, Liangqiong Qu, Yifang Xie, Sahar Ahmad, and Pew-Thian Yap. A paired dataset of t1-and t2-weighted mri at 3 tesla and 7 tesla. *Scientific Data*, 10(1):489, 2023. 1

[6] Xiaoyang Chen, Jinjian Wu, Wenjiao Lyu, Yicheng Zou, Kim-Han Thung, Siyuan Liu, Ye Wu, Sahar Ahmad, and Pew-Thian Yap. Brain tissue segmentation across the human lifespan via supervised contrastive learning. *arXiv preprint arXiv:2301.01369*, 2023. 1

[7] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 2

[8] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24 (9):1342–1350, 2018. 1

[9] Konstantin Dmitriev and Arie E Kaufman. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9501–9511, 2019. 1, 2

[10] Qing En and Yuhong Guo. Annotation by clicks: A point-supervised contrastive variance method for medical semantic segmentation. *arXiv preprint arXiv:2212.08774*, 2022. 2

[11] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. 1, 2, 4

[12] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3): 341–355, 2002. 1

[13] Yunhe Gao, Zhuowei Li, Di Liu, Mu Zhou, Shaoting Zhang, and Dimitris N Meta. Training like a medical resident: Universal medical image segmentation via context prior learning. *arXiv preprint arXiv:2306.02416*, 2023. 2

[14] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021. 5, 8

[15] Rui Huang, Yuanjie Zheng, Zhiqiang Hu, Shaoting Zhang, and Hongsheng Li. Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 146–155. Springer, 2020. 1, 2, 4

[16] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7014–7023, 2018. 2

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

[18] Zhanghexuan Ji, Dazhou Guo, Puyang Wang, Ke Yan, Le Lu, Minfeng Xu, Qifeng Wang, Jia Ge, Mingchen Gao, Xianghua Ye, et al. Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21140–21151, 2023. 1, 2

[19] Hoel Kervadec, Jose Dolz, Shanshan Wang, Eric Granger, and Ismail Ben Ayed. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In *Medical imaging with deep learning*, pages 365–381. PMLR, 2020. 2

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[21] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 5

[22] Mans Larsson, Erik Stenborg, Carl Toft, Lars Hammarstrand, Torsten Sattler, and Fredrik Kahl. Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 31–41, 2019. 2

[23] Yuemeng Li, Hongming Li, and Yong Fan. Acenet: Anatomical context-encoding network for neuroanatomy segmentation. *Medical image analysis*, 70:101991, 2021. 1

[24] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 21152–21164, 2023. 1, 2, 5

[25] Qin Liu, Han Deng, Chunfeng Lian, Xiaoyang Chen, Deqiang Xiao, Lei Ma, Xu Chen, Tianshu Kuang, Jaime Gateno, Pew-Thian Yap, et al. Skullengine: a multi-stage cnn framework for collaborative cbct image segmentation and landmark detection. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 606–614. Springer, 2021. 1

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[27] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *arXiv preprint arXiv:2111.02403*, 2021. 5

[28] Xiangde Luo, Minhao Hu, Wenjun Liao, Shuwei Zhai, Tao Song, Guotai Wang, and Shaoting Zhang. Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–538. Springer, 2022. 2

[29] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[30] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021. 5

[31] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023. 5

[32] Shima Nofallah, Mojgan Mokhtari, Wenjun Wu, Sachin Mehta, Stevan Knezevich, Caitlin J May, Oliver H Chang, Annie C Lee, Joann G Elmore, and Linda G Shapiro. Segmenting skin biopsy images with coarse and sparse annotations using u-net. *Journal of digital imaging*, 35(5):1238–1249, 2022. 2

[33] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pages 556–564. Springer, 2015. 5

[34] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023. 5, 8

[35] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70: 101979, 2021. 1, 2, 4

[36] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, 1(10):480–491, 2019. 1

[37] Ziyang Wang and Congying Ma. Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 870–879, 2023. 1

[38] Ziyang Wang and Chen Yang. Mixsegnet: Fusing multiple mixed-supervisory signals with multiple views of networks for mixed-supervised medical image segmentation. *Engineering Applications of Artificial Intelligence*, 133:108059, 2024. 1, 2

[39] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022. 5

[40] Hongrun Zhang, Liam Burrows, Yanda Meng, Declan Sculthorpe, Abhik Mukherjee, Sarah E Coupland, Ke Chen, and Yalin Zheng. Weakly supervised segmentation with point annotations for histopathology images via contrast-based variational model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2023. 2

[41] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1195–1204, 2021. 1, 2, 5

[42] Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Medical image analysis*, 43:98–111, 2018. 1

[43] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10672–10681, 2019. 1, 2, 4

[44] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019. 5, 8