# Weakly Supervised Lymph Nodes Segmentation Based on Partial Instance Annotations with Pre-trained Dual-branch Network and Pseudo Label Learning

Litingyu WANG                                                    litingyuwang@gmail.com
University of Electronic Science and Technology of China, Chengdu, China

Yijie QU
University of Electronic Science and Technology of China, Chengdu, China

Xiangde LUO
University of Electronic Science and Technology of China, Chengdu, China
ShangAI Laboratory, Shanghai, China

Wenjun LIAO
University of Electronic Science and Technology of China, Chengdu, China
Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Chengdu, China

Shichuan ZHANG
University of Electronic Science and Technology of China, Chengdu, China
Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Chengdu, China

Guotai WANG HTTPS://ORCID.ORG/0000-0002-8632-158X          guotai.wang@uestc.edu.cn
University of Electronic Science and Technology of China, Chengdu, China
ShangAI Laboratory, Shanghai, China

## Abstract

Assessing the presence of potentially malignant lymph nodes aids in estimating cancer progression, and identifying surrounding benign lymph nodes can assist in determining potential metastatic pathways for cancer. For quantitative analysis, automatic segmentation of lymph nodes is crucial. However, due to the labor-intensive and time-consuming manual annotation process required for a large number of lymph nodes, it is more practical to annotate only a subset of the lymph node instances to reduce annotation costs. In this study, we propose a pre-trained Dual-Branch network with Dynamically Mixed Pseudo label (DBDMP) to learn from partial instance annotations for lymph nodes segmentation. To obtain reliable pseudo labels for lymph nodes that are not annotated, we employ a dual-decoder network to generate different outputs that are then dynamically mixed. We integrate the original weak partial annotations with the mixed pseudo labels to supervise the network. To further leverage the extensive amount of unannotated voxels, we apply a self-supervised pre-training strategy to enhance the model's feature extraction capability. Experiments on the mediastinal Lymph Node Quantification (LNQ) dataset demonstrate that our method, compared to directly learning from partial instance annotations, significantly improves the Dice Similarity Coefficient (DSC) from 11.04% to 54.10% and reduces the Average Symmetric Surface Distance (ASSD) from 20.83 $mm$ to 8.72 $mm$. The code is available at https://github.com/WltyBY/LNQ2023_training_code.git

**Keywords:** Lymph Nodes Segmentation, Label-efficient Learning, Pseudo Labels

## 1. Introduction

Lymph node segmentation is essential in various medical applications, particularly in the diagnosis, staging and treatment planning of diseases such as cancer (Bouget et al., 2023). By segmenting lymph nodes and monitoring factors, such as their size and shape, clinicians can track disease progression and formulate treatment plans (Li and Xia, 2020). Deep learning methods have shown promising results in medical image segmentation (Chen et al., 2021; Isensee et al., 2021). However, obtaining the necessary pixel-level annotations is extremely time-consuming and labor-intensive (Pathak et al., 2015). Therefore, leveraging weak annotations to train a deep learning model is highly desirable for reducing annotation costs.

Weakly supervised learning can be extremely beneficial for training deep learning models in lymph node segmentation. Lymph nodes are distributed throughout the human body, often small in size, making it impractical to label all lymph nodes in a given area. In traditional weakly supervised tasks, labels for training segmentation models typically include image-level (Fu et al., 2023), box-level (Oh et al., 2021), point-level (Zhai et al., 2023) and scribble-level (Luo et al., 2022a) annotations. Image-level annotations merely indicate the presence of the target in the image and lack detailed information on its shape, intensity, and location, potentially leading to subpar performance. Due to the subtle contrast between lymph nodes and surrounding tissues in Computed Tomography (CT) scans, providing supervision information around the boundary poses a challenge for box-level and point/scribble-level annotations. In contrast, this study adopts a different weak annotation strategy called partial instance annotation, where only a small subset of lymph nodes are annotated in a volume. This annotation method provides the model with more information about the target's size, shape and boundary than the other weak annotations. Additionally, compared to fully supervised segmentation datasets, the annotation cost is significantly reduced.

Some researchers (Bouget et al., 2023; Feulner et al., 2013) have proposed annotating only a subset of lymph nodes for training, for example, only annotating lymph nodes with a high probability of disease based on size (minimum diameter value larger than 10 $mm$). However, some works still use fully supervised training procedures on datasets labeled in this manner. Oda et al. (2018) utilized Fully Convolutional Networks (FCNs) for mediastinal lymph node detection and segmentation, which is trained on annotated lymph nodes and other anatomical structures to address data imbalance. Bouget et al. (2023) introduced anatomical prior knowledge during training to assist the model in distinguishing lymph nodes from similar surrounding structures. However, solely segmenting diseased lymph nodes is inadequate for clinical use, as both diseased and normal lymph nodes are essential for diagnosing and treating diseases. Diseased lymph nodes offer insights into affected areas, while nearby normal lymph nodes can indicate potential metastasis paths for cancer. Therefore, it is imperative to segment both diseased and normal lymph nodes, even if only a subset of instances are annotated for training purpose.

In this work, we propose a novel framework named pre-trained **D**ual-**B**ranch network with **D**ynamically **M**ixed **P**seudo labels (DBDMP) which integrates self- and weakly supervised learning concepts along with noisy label learning to train a segmentation model with partial instance annotations. To better improve the feature extraction capability of

the model, we employ a self-supervised pre-training method called Model Genesis (Zhou et al., 2021), which involves an image reconstruction task. Additionally, within the framework of weakly supervised learning, we utilize a noise-robust loss to enhance learning from partial instance annotations. Furthermore, to effectively leverage unlabeled pixels during training, we introduce a real-time pseudo label learning strategy. We dynamically mix the outputs from two decoders to obtain soft pseudo labels, which are more robust to noise compared to hard pseudo labels (Müller et al., 2019). Subsequently, we merge the original partial annotations with the mixed predictions, leveraging the complementary information between the two kinds of labels for robust learning. The main contributions of this work are summarized as follows:

- We propose a novel pseudo label generation strategy for learning from partial instance annotations for lymph node segmentation. By assigning pseudo labels to unannotated lymph nodes instead of directly treating them as background, our approach effectively enhances the segmentation model's recall and reduces false negatives. Furthermore, the utilization of soft pseudo labels is more noise-tolerant than hard pseudo labels, which makes the training process more robust.

- During the pseudo label learning stage, a consensus-aware Cross-Entropy loss is proposed. The weight of each pixel is determined by the consistency between the two predictions derived from the weakly supervised learning framework. This approach facilitates the gradual learning of newly predicted foreground voxels by the model while mitigating the risk of being misled by incorrect ones.

- We adopt Model Genesis to initialize model parameters, enhancing the model's capability to extract superior features and edge information through the reconstruction of corrupted images.

Our method was validated on the Mediastinal Lymph Node Quantification (LNQ) dataset, and promising results have been achieved. In the LNQ challenge held on MICCAI 2023, we secured the $4^{th}$ position without utilizing any additional datasets for training, while other participants used extra training sets. Furthermore, our final methods attained a Dice Similarity Coefficient (DSC) score of 54.10% on the validation set and 55.44% on the test set, which correspond to an improvement of 43.06 and 36.40 percentage points compared with supervised learning from the annotated instances only, respectively. It is worth noting that our best DSC on the test set is 57.36%.

## 2. Related Works

**Lymph Nodes Segmentation**   Numerous efforts based on traditional vision-based methods have been dedicated to lymph node detection and segmentation, including Marginal Space Learning (MSL) (Barbu et al., 2011) and atlas-based segmentation (Stapleford et al., 2010), etc. However, traditional methods may face challenges such as suboptimal performance or excessive computation time (Zhao et al., 2020). In recent years, deep learning has been applied to lymph node segmentation due to its outstanding performance in tasks such as image classification and segmentation. Nogues et al. (2016) presented a method for automatic segmentation of lymph node clusters in CT images using holistically-nested neural networks and structured optimization. Bouget et al. (2019) proposed a 2D pipeline

that integrates the outputs of U-Net (Ronneberger et al., 2015) and Mask R-CNN (He et al., 2017) for segmentation and improves the performance with the instance detection. Xu et al. (2021) introduced a Cosine-Sine loss function and a multi-scale Atrous Spatial Pyramid Pooling (ASPP) module to the SegNet (Badrinarayanan et al., 2017) architecture to address the voxel class imbalance and enhance multi-scale information. Although these methods have achieved success in lymph node segmentation, they all relied on fully annotated training datasets with high annotation costs.

**Label-efficient Learning** The objective of label-efficient learning is to reduce the cost and time required by the labeling process while achieving performance comparable to fully supervised methods, especially for image segmentation tasks where it is expensive and time-consuming to obtain dense annotations (Shen et al., 2023). Label-efficient learning techniques encompass semi-supervised learning (Luo et al., 2022b), active learning (Settles, 2009), weakly supervised learning (Luo et al., 2022a) and noisy label learning (Wang et al., 2020b), among others. For example, Lin et al. (2016) utilized a graphical model that jointly propagates information from scribbles to unlabeled pixels based on superpixels (Ren and Malik, 2003). Luo et al. (2022a) employed auxiliary branch to generate pseudo labels in real-time and used a specific loss function to expand the scribbled regions. To ensure robustness against inaccurate annotations in segmentation tasks, Liu et al. (2022) enforced multi-scale cross-view consistency, and Wang et al. (2020a) introduced a noise-robust Dice loss. Compared to inaccurate annotations in existing noise-robust methods, partial instance learning has a larger degree of errors due to that most instances have been erroneously taken as the background. Furthermore, as the annotation type is different from the above weak annotations, existing weakly supervised segmentation methods cannot be directly used for learning from partial instance annotations.

**Self-supervised Learning** Self-supervised learning serves as a mechanism for models to learn rich feature representations from unlabeled data, thereby reducing reliance on large labeled datasets. This is commonly achieved by designing a pretext task. Gidaris et al. (2018) designed a classification-based pretext task to predict discretized rotation angles of an input image. Nogues et al. (2016) implemented the self-supervised task by solving jigsaw puzzles. Zhou et al. (2021) introduced Model Genesis which reconstructs a corrupted input to its original state. Designs like Model Genesis allow the model to extract universal image features effectively. Lei et al. (2021) proposed a novel contrastive learning approach, estimating the relative 3D offset between any pair of patches within the same volume. This method can perform well with just one-shot fine-tuning, while most other methods require fully supervised fine-tuning in the downstream task. However, in existing works, models trained by self-supervised learning are mainly fine-tuned with a small set of fully annotated images in downstream tasks, while applying them to weakly supervised learning has rarely been investigated.

## 3. Methods

Fig. 1 illustrates the proposed partial instance annotation learning framework named DB-DMP, which consists of a self-supervised pre-training stage and a pseudo label learning stage to deal with partial instance annotations. To achieve more stable predictions, we introduce
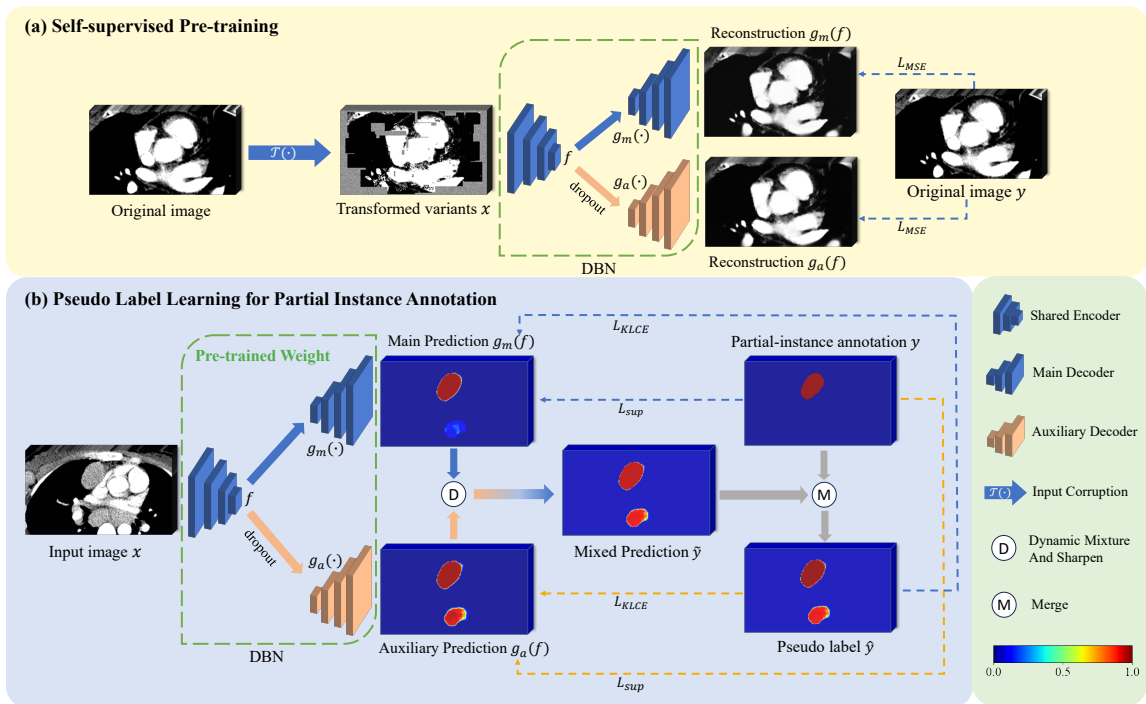
Figure 1: An overview of the proposed DBDMP which utilizes a dual-branch network with one shared encoder and two decoders. (a) In the self-supervised learning stage, Model Genesis is employed for pre-training. (b) In the downstream learning stage, a mixture of the outputs from the two decoders is combined with the original partial annotation to obtain a pseudo label. We also use a consensus-aware loss $\mathcal{L}_{KLCE}$ to avoid over-fitting noise in the pseudo labels.

a network with one encoder and two decoders to generate pseudo labels for unannotated instances. In self-supervised pre-training stage, as shown in Fig. 1(a), a model with dual branches is trained separately to improve feature extraction capabilities by reconstructing corrupted images. In Fig. 1(b), the outputs of the two decoders are mixed to obtain the pseudo label, aiming to leverage the prediction from the auxiliary branch to complement that from the main branch and supplement weak annotation information. To robustly learn the pseudo label, we introduce a consensus-aware loss function that assigns higher weights to voxels with more reliable pseudo labels. Additionally, to deal with the extreme imbalance between foreground and background voxels, we prioritize learning the foreground voxels with confidence and give lower weight to the learning of background voxels, with the aim of mining more potential lymph nodes that are unannotated.

## 3.1 Dual-branch Network

As shown in Fig. 1, the dual-branch network, which extends from the VNet architecture (Milletari et al., 2016), comprises a shared encoder and two decoders inspired by Luo
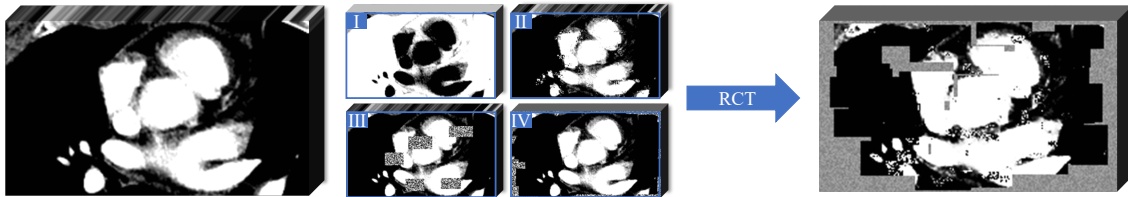
Figure 2: The transformations used to generate the input for self-supervised training: I. Non-linear transformation, II. Local pixel shuffling, III. In-painting, IV. Out-painting. RCT (Randomly Composed Transformation) means that the basic transformations are composed, each with a probability to be used. Note that in-painting and out-painting are not performed together each time.

et al. (2022a). Let's denote the inputs of the encoder as $x$ and the output features as $f$, which include features from the bottleneck layer and skip connections at different resolutions. The two decoders share the same structure, but have different inputs and parameters. The main decoder directly utilizes the features $f$ from the encoder as input, while the auxiliary decoder takes perturbed $f$ through dropout as input. We define the mappings of features from $f$ to the outputs of the main and auxiliary decoders as $g_m(\cdot)$ and $g_a(\cdot)$, respectively. In detail, one convolution block contains two convolution layers and a residual connection with Instance Normalization (IN) and Leaky ReLU activation function. Both the encoder and decoder are symmetrical structures with five different resolutions.

## 3.2 Self-supervised Pre-training

An appropriate pretext task can empower neural networks to learn low-level and high-level features that are conducive to downstream tasks (Jing and Tian, 2020). Zhou et al. (2021) introduced a self-supervised method called Model Genesis, which performs an image reconstruction process, and has shown promising results for downstream supervised segmentation tasks. Therefore, we use Model Genesis to pre-train the dual-branch network. Unlike the original Model Genesis that only trains one decoder, we extend it by training two decoders for the reconstruction process during pre-training.

Model Genesis (Zhou et al., 2021) employed three types of transformations on the original images, as detailed in Fig. 2: 1) Non-linear transformation integrates the Bézier Curve (Mortenson, 1999) to assign a unique determined value to each pixel, encouraging the model to focus on the information of image appearance and intensity distribution. 2) Local pixel shuffling samples a window smaller than the model's receptive field in the patch and rearranges the internal pixels to encourage the model to learn the local texture and boundary. 3) Out-painting or In-painting: Out-painting sets the outer pixels of a shape to random values, while the inner pixels retain their original intensities. In-painting follows the opposite way. The network learns visual features of images by reconstructing the original images from the corrupted version. In the self-supervised pre-training as shown in Fig. 1(a), the outputs from the main and auxiliary decoders compute the Mean Squared Error (MSE)

$\mathcal{L}_{MSE}$ separately with the original image $y$. The reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \mathcal{L}_{MSE}(p^m, y) + \mathcal{L}_{MSE}(p^a, y) \tag{1}$$

where $p^m = g_m(f)$ and $p^a = g_a(f)$ are the predictions of the main and auxiliary decoder, respectively. And $y$ denotes the reconstruction target, i.e., the original input image.

### 3.3 Supervised Loss for Partial Instance Annotations

Partial instance annotations can be considered as a noisy label learning problem, where some foreground regions are incorrectly labeled as the background. However, compared to conventional noisy label learning scenarios, the noise in the background of partial instance annotations can be excessive, with a large amount of false negatives. Therefore, we integrate weakly supervised learning and noisy label learning methods to effectively learn from partial instance annotations and generate reliable supervision signals. This combination enables us to mitigate the impact of excessive noise in the background and produce more effective supervision signals for the learning process.

Firstly, in our quest for more reliable supervision signals, we employ a noisy label learning technique to learn from partial instance annotation. Given a large amount of false negatives in the labels, we utilize the Symmetric Cross-Entropy (SCE) loss to balance the confidence between the partial instance annotation and the model's prediction, as proposed by Wang et al. (2019):

$$\mathcal{L}_{SCE}(p, y) = \gamma \times \mathcal{L}_{CE}(p, y) + \mathcal{L}_{CE}(y, p) \tag{2}$$

where $\mathcal{L}_{CE}$ is the widely used Cross-Entropy loss. $p$ and $y$ are the predicted probability map in a certain branch and partial instance annotation. The relationship between $\gamma$ and 1 indicates which one is more trustworthy between the prediction and the label. In this work we set $\gamma$ to 0.8 due to that the partial instance annotation is less credible than the model's prediction.

Secondly, we introduce the Partial Cross-Entropy (PCE) loss (Lee and Jeong, 2020) to ensure that the foreground in the partial instance label can be reliably learned. Unlike the PCE loss used in scribble-level annotations (Luo et al., 2022a), which supervises all the labeled voxels (including all target categories as well as the background) and does not calculate on unlabeled voxels, for partial instance annotations, we compute the cross-entropy only for the foreground voxels:

$$\mathcal{L}_{PCE}(p, y) = -\sum_{i \in \Omega} y_i \log p_i \tag{3}$$

where $\Omega$ is the foreground voxels in partial instance annotation. $p_i$ and $y_i$ denote the predicted probability and partial instance annotation of voxel $i$.

Finally, despite that $\mathcal{L}_{PCE}$ helps to improve recall of lymph nodes, it increases the risk of false positives. To deal with this problem, we additionally introduce a Tversky loss (Salehi et al., 2017) for supervision. Unlike Dice loss, which treats False Positives (FPs) and False Negatives (FNs) samples equally, the Tversky loss can balance the importance of both with different weights and mitigate class imbalance simultaneously:

$$\mathcal{L}_{Tversky}(p, y) = \frac{TP}{TP + \alpha \times FP + (1 - \alpha) \times FN} \tag{4}$$

where $N$ is the number of voxels, $TP = \sum_{i=1}^{N} p_i y_i$, $FP = \sum_{i=1}^{N} p_i(1 - y_i)$ and $FN = \sum_{i=1}^{N}(1-p_i)y_i$. By adjusting the hyper-parameter $\alpha$, we can control the importance between FPs and FNs. To predict more foreground voxels (false positive samples relative to partial instance annotation), $\alpha$ is set to 0.4 based on experiments.

For partial instance annotation, the supervised loss for each decoder is a combination of $\mathcal{L}_{PCE}$, $\mathcal{L}_{SCE}$ and $\mathcal{L}_{Tversky}$:

$$\mathcal{L}_{sup}(p, y) = \mathcal{L}_{SCE}(p, y) + \mathcal{L}_{PCE}(p, y) + \mathcal{L}_{Tversky}(p, y) \tag{5}$$

### 3.4 Online Pseudo Label Learning

Due to the presence of incorrectly labeled background voxels in partial instance annotations, it is unreliable to directly extract supervisory signals from them. Inspired by pseudo label learning for scribble annotations (Luo et al., 2022a), we first dynamically mix the predictions from the two decoders:

$$\tilde{y}_{mix} = \theta \times p^m + (1.0 - \theta) \times p^a \tag{6}$$

where $\theta$ is randomly generated from a uniform distribution between 0 and 1 at each iteration, enhancing the diversity of the pseudo label and compelling the model to continually update its predictions (Huo et al., 2021).

Then, we apply a sharpening function to adjust the entropy of the label distribution. The predicted probability of class $k$ can be define as:

$$\tilde{y}^k = \frac{e^{\tilde{y}_{mix}^k/\tau}}{\sum_{j \in C} e^{\tilde{y}_{mix}^j/\tau}} \tag{7}$$

where $\tilde{y}_{mix}^k$ is the mixed output of class $k$ and $C$ is the set of all categories. $\tau$ is a temperature that is normally set to 1 (Hinton et al., 2015). When $\tau > 1$, the labels become smoother, leading to increased entropy within the labels. Consequently, the information carried by negative labels is relatively amplified, directing the model training to pay more attention to negative labels. Conversely, when $\tau < 1$, the labels become sharper. Properly sharpening the labels can enhance their robustness to noise while also maintaining the differences between classes. We set $\tau$ to 0.3 in our implementation.

Finally, we integrate the mixed pseudo label with the partial instance annotation to obtain the final pseudo label, ensuring that the pseudo label complements the partial annotation. The final pseudo label is denoted as $\hat{y}$, and its $i^{th}$ element is defined as $\hat{y}_i = y_i + (1.0 - y_i)\tilde{y}_i$, i.e., the zero region in $y$ is replaced by the corresponding values from $\tilde{y}$.

The perturbation introduced in the auxiliary decoder may lead to uncontrollable effects. Ideally, predictions for background voxels near the classification boundary should shift towards the foreground space. However, foreground voxels in partial instance annotations may be predicted as background, leading to misleading effects in the model's training. To mitigate such adverse effects, we only learn from pixels with minor discrepancies based on the consistency of the two outputs, ensuring a smooth and gradual learning process. We utilize Kullback-Leibler (KL) divergence to estimate the consistency of the two outputs and

use it to generate voxel-wise weights for the Cross-Entropy loss. Following the approach outlined in Zheng and Yang (2021), the weight for voxel $i$ is defined as:

$$\mathcal{W}_i = e^{-KL(p_i^m, p_i^a)} \tag{8}$$

where $KL(p_i^m, p_i^a) = p_i^a log(p_i^a/p_i^m)$ is the KL divergence loss calculated from the $i^{th}$ voxel of the two probability maps $p^m$ and $p^a$. When the predictions of a certain voxel from the main decoder and the auxiliary decoder are highly dissimilar, Eq. (8) will lead to a lower value of $\mathcal{W}_i$. Based on this observation, the learning loss of the main branch for pseudo label can be formulated as:

$$\mathcal{L}_p(p^m, p^a, \hat{y}) = \frac{1}{W} \sum_i [\mathcal{W}_i(-\hat{y}_i \log p_i^m) + KL(p_i^m, p_i^a)] \tag{9}$$

where $W$ is the sum of $\mathcal{W}_i$ for all voxels. The introduction of $KL$ can avoid excessive discrepancies between the predictions of the two decoders.

The proposed DBDMP framework learns from both partial instance annotation and pseudo label by minimizing the following combined objective function:

$$\begin{aligned} \mathcal{L}_{total} =& \mathcal{L}_{sup}(p^m, y) + \mathcal{L}_{sup}(p^a, y) \\ &+ \lambda_p[\mathcal{L}_p(p^m, p^a, \hat{y}) + \mathcal{L}_p(p^a, p^m, \hat{y})] \end{aligned} \tag{10}$$

where $y$ and $\hat{y}$ are partial instance annotation and the generated pseudo label, respectively. $\lambda_p$ is the trade-off weight that schedules with an epoch-dependent sigmoid-like ramp-up function in the first 100 epochs as the pseudo labels in the early training stage can be in poor equality:

$$\lambda_p = \lambda \times e^{-5 \times (1 - t/t_{max})^2} \tag{11}$$

where $\lambda$ is a hyper-parameter that represents the final value of $\lambda_p$. $t_{max}$ is set to 99 which means the maximal epoch for ramp-up and $t$ is the current epoch.

## 4. Experiments

### 4.1 Dataset

**LNQ2023 Challenge Dataset** The Mediastinal Lymph Node Quantification (LNQ): Segmentation of Heterogeneous CT Data Challenge dataset includes 513 CT volumes. Each volume contains 48 to 656 slices with slice thickness ranging from 2.0 to 5.5 $mm$ and pixel size 1.0 $mm \times$ 1.0 $mm$. The matrix size in the axial plane is 512$\times$512. The images were split at patient level into 393, 20, and 100 for training, validation, and testing, respectively. In the training set, cases are partially annotated, meaning only one or several positive lymph nodes in the volumes are labeled, while all diseased lymph nodes in the validation and test sets are fully annotated.

### 4.2 Implementation Details

Our method was implemented in nnUNet (Isensee et al., 2021), which is a Pytorch-based (Paszke et al., 2019) toolkit for image computing with deep learning. The implementation
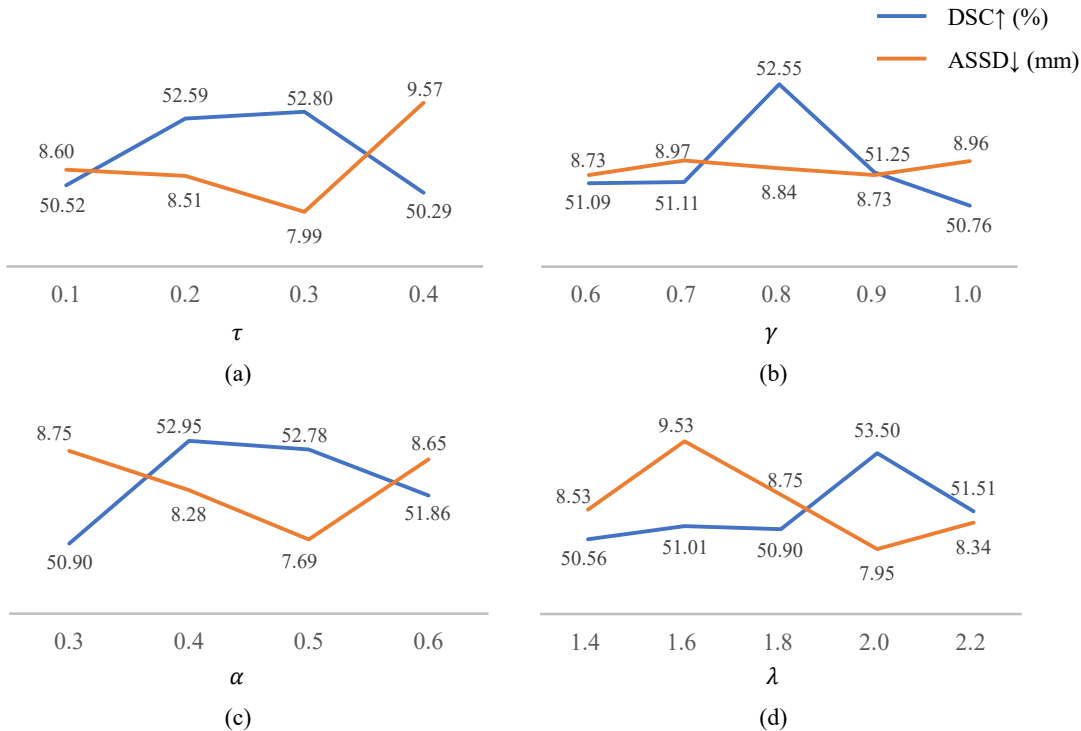
Figure 3: Sensitivity analysis of hyper-parameters $\tau$, $\gamma$, $\alpha$ and $\lambda$, respectively.

was carried out on a single NVIDIA 2080Ti GPU with 11GB VRAM. We utilized a VNet-like (Milletari et al., 2016) network as the backbone for all experiments, and extended it to two decoders, as detailed in Section 3.1.

For preprocessing, we first cropped the volumes to the lung region based on intensity. Subsequently, we resampled each volume into the resolution of 3.0 $mm\times$0.8 $mm\times$0.8 $mm$. Finally, we normalized each volume to have zero mean and unit variance. Our networks were trained using a patch-based approach with a patch size of $224 \times 128 \times 64$ and a batch size of 2. We employed the polynomial learning rate strategy to decay the learning rate in each epoch.

For self-supervised pre-training in Fig. 1(a), we used Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.99, an initial learning rate of 0.01, and weight decay of $3 \times 10^{-5}$ to minimize the reconstruction loss in Eq. (1). The training process lasted for 1000 epochs, with 250 iterations in each epoch. During weakly supervised training, the segmentation model was initialized with the weights obtained from self-supervised pre-training. We minimized the loss functions in Eq. (10) using SGD optimizer with a momentum of 0.9, while keeping the other parameters the same as those in the self-supervised pre-training stage. The training epoch was 300 with 250 iterations in each epoch.

During the inference stage, we loaded the weights from the final epoch and only utilized predictions from the main decoder as the final outputs. All inference processes were conducted using a sliding window strategy. We also applied a specific post-processing method,

Table 1: Ablation study of our proposed method on LNQ2023 Challenge Dataset for both validation set and test set. DBN is the Dual-branch Network detailed in Section 3.1. DBN† means using the pre-training weight generated by Model Genesis (Zhou et al., 2021) in Section 3.2. $\mathcal{L}_{Dice}$ is the widely used Dice loss and $\mathcal{L}_{KLCE}$ is explained in detail in Eq. (9). Training VNet with one encoder and one decoder by $\mathcal{L}_{CE}$ is served as the baseline method.

| | Network | $\mathcal{L}_{sup}$ | $\mathcal{L}_p$ | Validation Set | | Test Set | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | | | | DSC↑(%) | ASSD↓(mm) | DSC↑(%) | ASSD↓(mm) |
| (a) | DBN | $\mathcal{L}_{CE}$ | $\mathcal{L}_{Dice}$ | $15.12_{\pm14.75}$ | $26.94_{\pm11.78}$ | $23.22_{\pm17.50}$ | $25.41_{\pm15.33}$ |
| (b) | DBN | $\mathcal{L}_{CE} + \mathcal{L}_{Tversky}$ | $\mathcal{L}_{Dice}$ | $34.99_{\pm25.82}$ | $16.39_{\pm12.73}$ | $45.10_{\pm21.58}$ | $13.54_{\pm12.05}$ |
| (c) | DBN | $\mathcal{L}_{CE} + \mathcal{L}_{Tversky}$ | $\mathcal{L}_{KLCE}$ | $32.10_{\pm24.29}$ | $15.88_{\pm9.09}$ | $40.26_{\pm22.39}$ | $15.19_{\pm10.94}$ |
| (d) | DBN | $\mathcal{L}_{SCE} + \mathcal{L}_{Tversky}$ | $\mathcal{L}_{KLCE}$ | $31.48_{\pm25.68}$ | $17.63_{\pm11.84}$ | $39.34_{\pm22.58}$ | $16.54_{\pm15.75}$ |
| (e) | DBN | $\mathcal{L}_{PCE} + \mathcal{L}_{Tversky}$ | $\mathcal{L}_{KLCE}$ | $52.53_{\pm22.29}$ | $\underline{8.39}_{\pm6.84}$ | $\mathbf{57.36}_{\pm17.09}$ | $\underline{9.85}_{\pm13.25}$ |
| (f) | DBN | $\mathcal{L}_{PCE} + \mathcal{L}_{SCE} + \mathcal{L}_{Tversky}$ | $\mathcal{L}_{KLCE}$ | $\underline{53.31}_{\pm20.40}$ | $\mathbf{8.07}_{\pm6.55}$ | $\underline{56.10}_{\pm17.58}$ | $10.28_{\pm12.75}$ |
| (g) | DBN† | $\mathcal{L}_{PCE} + \mathcal{L}_{SCE} + \mathcal{L}_{Tversky}$ | $\mathcal{L}_{KLCE}$ | $\mathbf{54.10}_{\pm21.92}$ | $8.72_{\pm7.71}$ | $55.44_{\pm18.98}$ | $\mathbf{9.35}_{\pm7.69}$ |
| Baseline | VNet | $\mathcal{L}_{CE}$ | | $11.04_{\pm17.86}$ | $20.83_{\pm8.64}$ | $19.04_{\pm19.40}$ | $18.23_{\pm10.10}$ |
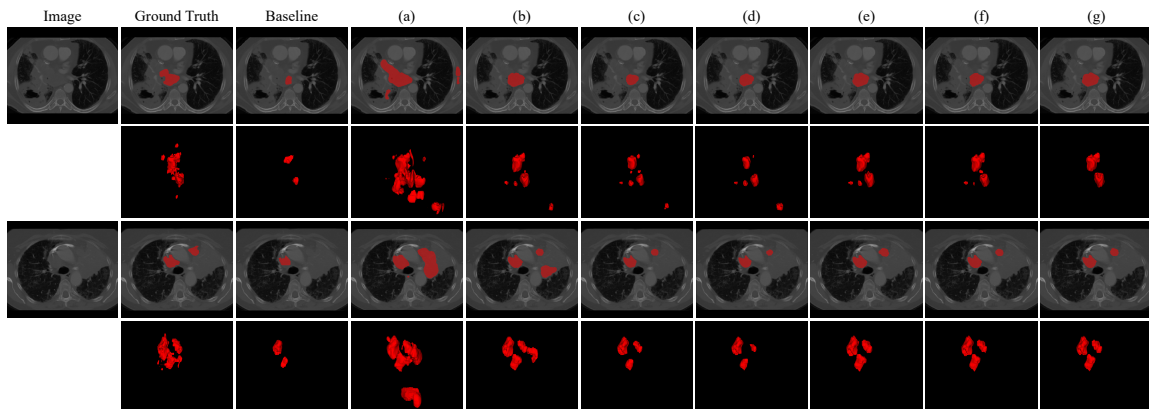


Figure 4: Visualization of segmentation results of the ablation study in Table 1.

which involved removing lymph node regions at the boundaries of an image and eliminating a portion of the lymph nodes based on voxel intensity and the actual volume. For quantitative evaluation, we calculated the Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASSD) between a segmentation result and the ground truth. In light of the potential for samples with failed predictions, the ASSD for these samples is missing, so we fill them with the maximum value of the successfully calculated ASSD and then average them. This may result in larger mean ASSD values.

### 4.3 Results

**Sensitivity Analysis of Some Hyper-parameters** We conducted experiments to evaluate the sensitivity of the hyper-parameters $\tau$ in Eq. (7), $\gamma$ in Eq. (2), $\alpha$ in Eq. (4) and $\lambda$ in Eq. (11). Fig. 3 presents the results obtained on the validation set.

The hyper-parameter $\tau$ governs the extent of sharpening applied to the soft pseudo labels. We investigated the segmentation performance of the proposed framework by setting $\tau$ to 0.1, 0.2, 0.3, and 0.4, respectively. As illustrated in Fig. 3(a), when $\tau$ increases from 0.1 to 0.3, the DSC score improves. However, the performance is decreased when $\tau$ increases to 0.4, showing that the best value of $\tau$ is 0.3.

The hyper-parameter $\gamma$ signifies the degree of confidence between the model's predictions and partial instance annotations. The results depicted in Fig. 3(b) indicate that, considering both DSC and ASSD, the model achieved the best result when $\gamma = 0.8$. This suggests that the model deems its own predictions more reliable than the partial instance annotations.

The hyper-parameter $\alpha$ in Eq. (4) balances the penalty imposed on False Positives (FPs) and False Negatives (FNs), and was tested with different values in $\{0.3, 0.4, 0.5, 0.6\}$. When $\alpha$ is smaller, the model imposes less penalty on FPs, thus encouraging the model to predict more positive results than those in the partial instance annotations. However, a weaker penalty on FPs may lead to an over-prediction of foreground voxels. In Fig. 3(c), $\alpha = 0.4$ and $\alpha = 0.5$ achieved very close DSC scores, and $\alpha = 0.5$ has a lower ASSD values. When $\alpha = 0.3$, the performance was much lower. The increased ASSD may be due to that the model predicted some foreground voxels that are far from the actual lymph nodes.

The value of $\lambda$ represents the confidence in the quality of the generated pseudo labels during the training process. We conducted experimental tests with the set $\{1.4, 1.6, 1.8, 2.0, 2.2\}$. As illustrated in Fig. 3(d), the result is notably superior when $\lambda = 2.0$ compared to other settings.

**Ablation Study**   We conducted additional experiments to validate the effectiveness of Dual-Branch Network (DBN) and the modifications made to adapt the work of Luo et al. (2022a) for learning from partial instance annotations. The quantitative results on the validation set and test set are presented in Table. 1, where the baseline method was taking the partial annotations as full ones to train a VNet with cross entropy loss.

Table 1 shows that the baseline method only achieved an average Dice of 11.04% on the validation set, indicating insufficient supervision from the partial annotations. By leveraging pseudo labels from the dual-branch network with $\mathcal{L}_{Dice}$, it was improved to 15.12%. By additionally introducing the Tversky loss for the original partial annotations, the average Dice was 34.99%. By combing $\mathcal{L}_{PCE} + \mathcal{L}_{SCE} + \mathcal{L}_{Tversky}$ for partial instance annotations and $L_{KLCE}$ for pseudo labels, the average Dice was 53.31%, and leveraging a pretrained model based on Model Genesis further improved it to 54.10%, showing the effectiveness the loss design and pretraining strategy of our method.

On the testing set, comparison between (c) and (e) in Table 1 shows that replacing $\mathcal{L}_{CE}$ by $\mathcal{L}_{PCE}$ substantially improved the average Dice from 40.26% to 57.36%, showing the effectiveness of reducing the contribution of background voxels in the loss calculation when many lymph nodes are incorrectly labeled as the background in parial instance annotations. The proposed method achieved the lowest ASSD value of 9.35 $mm$. Despite that using $\mathcal{L}_{PCE} + \mathcal{L}_{SCE} + \mathcal{L}_{Tversky}$ for partial instance annotations achieved a lower DSC value than $\mathcal{L}_{PCE} + \mathcal{L}_{Tversky}$, their gap is relatively small considering the performance of the other methods. The different performance between validation set and testing set is mainly from the data distribution shift between the two subsets.

Fig. 4 shows a visual comparison between the compared methods listed in Table 1. It can be observed that the baseline method has obvious under-segmentation, due to taking false negatives in the partial annotation as the background. The naive pseudo label learning method (a) has a lot of over-segmentation, due to that the pseudo labels contain many false positives. By using $L_{KLCE}$ for the pseudo labels and $L_{PCE}$, $L_{SCE}$, $L_{Tversky}$ for partial annotations respectively, the performance continues to improve.

## 5. Discussion and Conclusion

In this study, we explored a weakly supervised learning framework based on partial instance annotations for lymph node segmentation. For such annotations, the key is to identify trustworthy background regions and provide strong foreground signals during training to ensure robust learning of foreground voxels. Our method deals with this problem by generating pseudo labels to mine more potential lymph nodes. As pseudo labels from a single prediction branch may have bias, we propose dynamic mixture of predictions from two branches, leading to more stable pseudo labels and better uncertainty estimation of them based on divergence between the two branches. Loss functions are also carefully designed to highlight the foreground class while reducing the effect of noise in pseudo labels.

This work also has some limitations that could be addressed in the future. First, the segmentation model in our work only learns from the LNQ dataset, and the performance may be further improved by leveraging other existing fully supervised datasets. We believe our approach is adaptable enough for mixed datasets that have both partially and fully annotated cases. Second, the LNQ dataset has only large lymph nodes labeled in the training set, and the distributions of the labeled ones for training and those for testing may be different, making it more challenging to obtain robust performance during testing, especially for small lymph nodes that have not been annotated in the training set. Improving the diversity of the labeled cases under the same annotation budget is a potential solution for this problem, such as making the labeled cases contain lymph nodes with different scales, positions, and shapes. In addition, the loss function in this work has several hyper-parameters, and they are searched manually. In the future, it would be interesting to automate the determination of these hyper-parameters.

In conclusion, we proposed a partial instance annotation learning framework that combines weakly supervised learning and noisy label learning for lymph node segmentation. By introducing a dual-branch network, we dynamically mixed the outputs from the two decoders and fused them with partial instance annotations to obtain reliable pseudo labels. In learning from partial instance annotations, the introduction of multiple loss functions not only provides more reliable foreground and background supervision signals but also facilitates the segmentation of potential lymph nodes that are not labeled out. We conducted experiments using the dataset from the Mediastinal Lymph Node Quantification Challenge, without using any other datasets for pre-training or during the training stage. We finally achieved an average DSC of 54.10% and 55.44%, and average ASSD of 8.72 $mm$ and 9.35 $mm$ on validation set and test set, respectively. In the future, it is of interest to leverage other labeled or unlabeled datasets to assist the learning process, such as using unannotated datasets for self-supervised pre-training, or leveraging a small number of fully labeled images to boost the segmentation performance.

## Acknowledgments

## Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

## Conflicts of Interest

We declare we do not have conflicts of interest.

## Data availability

All data used in the paper are from the Mediastinal Lymph Node Quantification (LNQ): Segmentation of Heterogeneous CT Data competition. The training set data and annotations, as well as the validation set data, can be obtained from the competition website `https://lnq2023.grand-challenge.org/`. However, annotations for the validation set and both the data and annotations for the test set are not available on the website.

## References

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Adrian Barbu, Michael Suehling, Xun Xu, David Liu, S Kevin Zhou, and Dorin Comaniciu. Automatic detection and segmentation of lymph nodes from ct data. *IEEE Transactions on Medical Imaging*, 31(2):240–250, 2011.

David Bouget, Arve Jørgensen, Gabriel Kiss, Haakon Olav Leira, and Thomas Langø. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in ct data for lung cancer staging. *International journal of computer assisted radiology and surgery*, 14:977–986, 2019.

David Bouget, André Pedersen, Johanna Vanel, Haakon O Leira, and Thomas Langø. Mediastinal lymph nodes segmentation using 3d convolutional neural network ensembles and anatomical priors guiding. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(1):44–58, 2023.

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.

Johannes Feulner, S Kevin Zhou, Matthias Hammon, Joachim Hornegger, and Dorin Comaniciu. Lymph node detection and segmentation in chest ct data using discriminative learning and a spatial prior. *Medical image analysis*, 17(2):254–270, 2013.

Jia Fu, Tao Lu, Shaoting Zhang, and Guotai Wang. UM-CAM: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised fetal brain segmentation. In *MICCAI*, pages 315–324, 2023.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xinyue Huo, Lingxi Xie, Jianzhong He, Zijie Yang, Wengang Zhou, Houqiang Li, and Qi Tian. Atso: Asynchronous teacher-student optimization for semi-supervised image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1235–1244, 2021.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43 (11):4037–4058, 2020.

Hyeonsoo Lee and Won-Ki Jeong. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 14–23. Springer, 2020.

Wenhui Lei, Wei Xu, Ran Gu, Hao Fu, Shaoting Zhang, Shichuan Zhang, and Guotai Wang. Contrastive learning of relative position regression for one-shot object localization in 3d medical images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 155–165. Springer, 2021.

Zhe Li and Yong Xia. Deep reinforcement learning for weakly-supervised lymph node segmentation in ct images. *IEEE Journal of Biomedical and Health Informatics*, 25(3): 774–783, 2020.

Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2606–2616, 2022.

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, 2022a.

Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80: 102517, 2022b.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

Michael E Mortenson. *Mathematics for computer graphics applications*. Industrial Press Inc., 1999.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Isabella Nogues, Le Lu, Xiaosong Wang, Holger Roth, Gedas Bertasius, Nathan Lay, Jianbo Shi, Yohannes Tsehay, and Ronald M Summers. Automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 388–397. Springer, 2016.

Hirohisa Oda, Holger R Roth, Kanwal K Bhatia, Masahiro Oda, Takayuki Kitasaka, Shingo Iwano, Hirotoshi Homma, Hirotsugu Takabatake, Masaki Mori, Hiroshi Natori, et al. Dense volumetric detection and segmentation of mediastinal lymph nodes in chest ct images. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 1057502. SPIE, 2018.

Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6913–6922, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.

Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

Ren and Malik. Learning a classification model for segmentation. In *Proceedings ninth IEEE international conference on computer vision*, pages 10–17. IEEE, 2003.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

Burr Settles. Active learning literature survey. 2009.

Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Q Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Liza J Stapleford, Joshua D Lawson, Charles Perkins, Scott Edelman, Lawrence Davis, Mark W McDonald, Anthony Waller, Eduard Schreibmann, and Tim Fox. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. *International Journal of Radiation Oncology\* Biology\* Physics*, 77(3):959–966, 2010.

Guotai Wang, Xinglong Liu, Chaoping Li, Zhiyong Xu, Jiugen Ruan, Haifeng Zhu, Tao Meng, Kang Li, Ning Huang, and Shaoting Zhang. A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2653–2663, 2020a.

Lu Wang, Dong Guo, Guotai Wang, and Shaoting Zhang. Annotation-efficient learning for medical image segmentation based on noisy pseudo labels and adversarial learning. *IEEE Transactions on Medical Imaging*, 40(10):2795–2807, 2020b.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019.

Guoping Xu, Hanqiang Cao, Jayaram K Udupa, Yubing Tong, and Drew A Torigian. Disegnet: A deep dilated convolutional encoder-decoder architecture for lymph node segmentation on pet/ct images. *Computerized Medical Imaging and Graphics*, 88:101851, 2021.

Shuwei Zhai, Guotai Wang, Xiangde Luo, Qiang Yue, Kang Li, and Shaoting Zhang. Paseg: learning from point annotations for 3d medical image segmentation using contextual regularization and cross knowledge distillation. *IEEE Transactions on Medical Imaging*, 2023.

Xingyu Zhao, Peiyi Xie, Mengmeng Wang, Wenru Li, Perry J Pickhardt, Wei Xia, Fei Xiong, Rui Zhang, Yao Xie, Junming Jian, et al. Deep learning–based fully automated detection and segmentation of lymph nodes on multiparametric-mri for rectal cancer: A multicentre study. *EBioMedicine*, 56, 2020.

Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.