Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation

Yuyin Zhou[1*†] Zhe Li[2†] Song Bai[3] Chong Wang[4*] Xinlei Chen[5*]
Mei Han[6*] Elliot Fishman[7] Alan L. Yuille[1]

[1]Johns Hopkins University [2]Google Research [3]University of Oxford [4]ByteDance Inc.
[5]Facebook [6]PAII Inc. [7]The Johns Hopkins Medical Institute

Abstract

Accurate multi-organ abdominal CT segmentation is essential to many clinical applications such as computer-aided intervention. As data annotation requires massive human labor from experienced radiologists, it is common that training data are partially labeled, e.g.*, pancreas datasets only have the pancreas labeled while leaving the rest marked as background. However, these background labels can be misleading in multi-organ segmentation since the "background" usually contains some other organs of interest. To address the background ambiguity in these partially-labeled datasets, we propose Prior-aware Neural Network (PaNN) via explicitly incorporating anatomical priors on abdominal organ sizes, guiding the training process with domain-specific knowledge. More specifically, PaNN assumes that the average organ size distributions in the abdomen should approximate their empirical distributions, prior statistics obtained from the fully-labeled dataset. As our training objective is difficult to be directly optimized using stochastic gradient descent, we propose to reformulate it in a min-max form and optimize it via the stochastic primal-dual gradient algorithm. PaNN achieves state-of-the-art performance on the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault", a competition on organ segmentation in the abdomen. We report an average Dice score of* ***84.97****%, surpassing the prior art by a large margin of* ***3.27****%.*

1. Introduction

This work focuses on multi-organ segmentation in abdominal regions which contain multiple organs such as liver, pancreas and kidneys. The segmentation of internal structures on medical images, *e.g.*, CT scans, is an essential prerequisite for many clinical applications such as computer-aided diagnosis, computer-aided intervention and radiation therapy. Compared with other internal structures such as heart or brain, abdominal organs are much more difficult to segment due to the morphological and structural complexity, low contrast of soft tissues, *etc*.

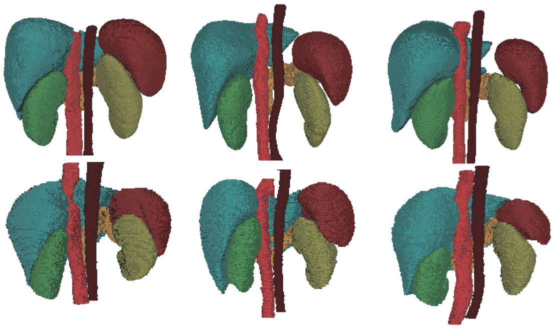

Figure 1. 3D Visualization of several abdominal organs (liver, spleen, left kidney, right kidney, aorta, inferior vena cava) to show the similarity of patient-wise abdominal organ size distributions.

With the development of deep convolutional neural networks (CNNs), many medical image segmentation problems have achieved satisfactory results only when full-supervision is available [33, 32, 45, 41, 30, 4]. Despite the recent progress, the annotation of medical radiology images is extremely expensive, as it must be handled by experienced radiologists and carefully checked by additional experts. This results in the lack of high-quality labeled training data. More critically, how to efficiently incorporate domain-specific expertise (*e.g.*, anatomical priors) with segmentation models [10, 25], such as organ shape, size, remains an open issue.

Our key observation is that, in medical image analysis domain, instead of scribbles [17, 36, 37] , points [3] and image-level tags [26, 27, 40], there exists a considerable number of datasets in the form of abdominal CT scans [31, 33, 34]. To meet different research goals or practical usages, these datasets are annotated to target different organs (a subset of abdominal organs), *e.g.*, pancreas datasets [31] only have the pancreas labeled while leaving the rest marked as background.

[*] This work was partly done when Yuyin Zhou, Chong Wang, Xinlei Chen and Mei Han were at Google.

[†] Equal Contribution.

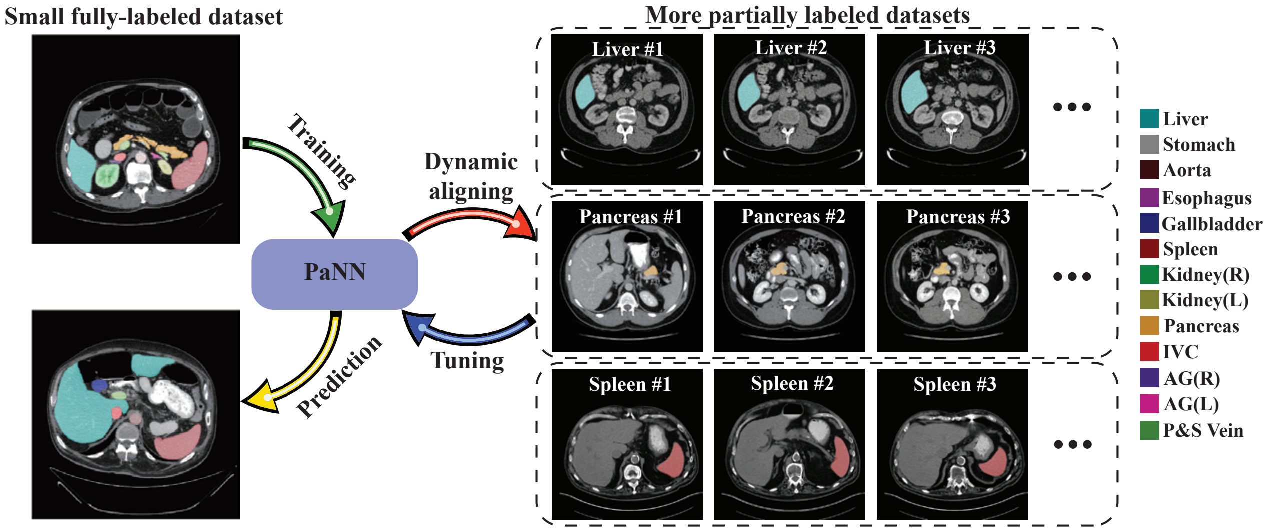

Figure 2. Overview of the proposed PaNN for partially-supervised multi-organ segmentation. It is trained with a small set of fully-labeled dataset and several partially-labeled datasets. The PaNN regularizes that the organ size distributions of the network output should approximate their prior statistics in the abdominal region obtained from the fully-labeled dataset.

The aim of this work is to fully leverage these existing partially-annotated datasets to assist multi-organ segmentation, which we refer to as *partial supervision*. To address the challenge of partial supervision, an intuitive solution is to simply train a segmentation model directly on both the labeled data and the partially-labeled data in the semi-supervised manner [29, 2, 26]. However, it 1) fails to take advantages of the fact that medical images are naturally more constrained compared with natural images [24]; 2) is intuitively misleading as it treats the unlabeled pixels/voxels as background. To overcome these issues, we propose Prior-aware Neural Network (PaNN) to handle such background ambiguity via incorporating prior knowledge on organ size distributions. We achieve this via a prior-aware loss, which acts as an auxiliary and soft constraint to regularize that the average output size distributions of different organs should approximate their prior proportions. Based on the anatomical similarities (Fig. 1) across different patient scans [10, 25, 15], the prior proportions are estimated by statistics from the fully-labeled data. The overall pipeline is illustrated in Fig. 2. It is important to note that the training objective is hard to be directly optimized using stochastic gradient descent. To address this issue, we propose to formulate our objective in a min-max form, which can be well optimized via the stochastic primal-dual gradient algorithm [20]. To summarize, our contributions are three-fold:

1) We propose Prior-aware Neural Network, which incorporates domain-specific knowledge from medical images, to facilitate multi-organ segmentation via using partially-annotated datasets.

2) As the training objective is difficult to be directly optimized using stochastic gradient descent, it is essential to reformulate it in a min-max form and optimize via stochastic primal-dual gradient [20].

3) PaNN significantly outperforms previous state-of-the-arts even using fewer annotations. It achieves 84.97% on the MICCAI2015 challenge "Multi-Atlas Labeling Beyond the Cranial Vault" in the free competition for organ segmentation in the abdomen.

2. Related Work

Currently, the most successful deep learning techniques for semantic segmentation stem from a common forerunner, *i.e.*, Fully Convolutional Network (FCN) [21]. Based on FCN, many recent advanced techniques have been proposed, such as DeepLab [5, 6, 7], SegNet [1], PSPNet [43], RefineNet [18], *etc*. Most of these methods are based on supervised learning, hence requiring a sufficient number of labeled training data to train. To cope with scenarios where supervision is limited, researchers begin to investigate the weakly-supervised setting [26, 27, 9], *e.g.*, only bounding-boxes or image-level labels are available, and the semi-supervised setting [26, 35], *i.e.*, unlabeled data are used to enlarge the training set. Papandreou *et al*. [26] propose EM-Adapt where the pseudo-labels of the unknown pixels are estimated in the expectation step and standard SGD is performed in the maximization step. Souly *et al*. [35] demonstrate the usefulness of generative adversarial networks for semi-supervised segmentation.

In the medical imaging domain, it becomes more intractable to acquire sufficient labeled data due to the difficulty of annotation, as the annotation has to be done by experts. Although fully-supervised methods (*e.g.*, UNet [30], VoxResNet [4], DeepMedic [14], 3D-DSN [11], HNN [32]) have achieved remarkable performance improvement in tasks such as brain MR segmentation, abdominal single-organ segmentation and multi-organ segmentation, semi- or weakly-supervised learning is still a far more realistic solution. For example, Bai *et al*. [2] proposed an EM-based iterative method, where a CNN is alternately trained on labeled and post-processed unlabeled sets. In [42], supervised and unsupervised adversarial costs are involved to address semi-supervised gland segmentation. DeepCut [29] shows that weak annotations such as bounding-boxes in medical image segmentation can also be utilized by performing an iterative optimization scheme like [26].

However, these methods fail to capture the anatomical priors [19]. Inclusion of priors in medical imaging could potentially have much more impact compared with their usage in natural images since anatomical objects in medical images are naturally more constrained in terms of shape, location, size, *etc*. Some recent works [10, 25] demonstrate that these priors can be learned by a generative model. But these methods will induce heavy computational overhead. Kervadec *et al*. [15] proposed that directly imposing inequality constraints on sizes is also an effective way of incorporating anatomical priors. Unlike these methods, we propose to learn from partial annotations by embedding the abdominal region statistics in the training objective, which requires no additional training budget.

3. Prior-aware Neural Network

Our work aims to address the multi-organ segmentation problem with the help of multiple existing partially-labeled datasets. Given a CT scan where each element indicates the Housefield Unit (HU) of a voxel, the goal is to find the predicted labelmap of each pixel/voxel.

3.1. Partial Supervision

We consider a new supervision paradigm, *i.e*., partial supervision, for multi-organ segmentation. This is motivated by the fact that there exists a considerable number of datasets with only one or a few organs labeled in the form of abdominal CT scans [31, 33, 34] in medical image analysis, which can serve as partial supervision for multi-organ segmentation (see the list in the supplementary material). Based on domain knowledge, our approach assumes the following characteristics of the datasets which are common in medical image analysis. First, the scanning protocols of medical images are well standardized, *e.g*., brain, head and neck, chest, abdomen, and pelvis in CT scans, which means that the internal structures are consistent in a limited range according to the scanning protocol (see Fig. 1). Second, internal organs have anatomical and spatial relationships such as gastrointestinal track, *i.e*., stomach, duodenum, small intestine, and colon are connected in a fixed order.

The partially-supervised setting can be formally defined as below. Given a fully-labeled dataset $\mathbf{S}_{\mathrm{L}} = \{\mathbf{I}_{\mathrm{L}}, \mathbf{Y}_{\mathrm{L}}\}$ with the annotation $\mathbf{Y}_{\mathrm{L}}$ known and T partially-labeled datasets $\mathbf{S}_{\mathrm{P}} = \{\mathbf{S}_{\mathrm{P}_1}, \mathbf{S}_{\mathrm{P}_2}, ...\mathbf{S}_{\mathrm{P}_T}\}$ with the t-th dataset defined as $\mathbf{S}_{\mathrm{P}_t} = \{\mathbf{I}_{\mathrm{P}_t}, \mathbf{Y}_{\mathrm{P}_t}\}$. $\mathrm{L} = \{1, 2, ..., n_{\mathrm{L}}\}$ and $\mathrm{P}_t = \{1, 2, ..., n_{\mathrm{P}_t}\}$ denote the image indices for $\mathbf{S}_{\mathrm{L}}$ and $\mathbf{S}_{\mathrm{P}_t}$, respectively. For each element $y_{ij} \in \mathbf{Y}_{\mathrm{L}}$, y_{ij} denotes the annotation of the j-th pixel in the i-th image $\mathbf{I}_i \in \mathbf{I}_{\mathrm{P}_t}$ and is selected from $\mathcal{L}$, where $\mathcal{L}$ denotes the abdominal organ space, *i.e*., $\mathcal{L} = \{\mathrm{spleen}, \mathrm{pancreas}, \mathrm{liver}, ...\}$. For the t-th partially-labeled dataset $\mathbf{S}_{\mathrm{P}_t}$, $y_{ij} \in \mathbf{Y}_{\mathrm{P}_t}$ is selected from $\mathcal{L}_{\mathrm{P}_t} \subseteq \mathcal{L}$. In 2D-based segmentation models, the i-th input $\mathbf{I}_i$ is a sliced 2D image from either Axial, Coronal or Saggital view of the whole CT scan [45, 32, 44, 39]. In 3D-based segmentation models, $\mathbf{I}_i$ is a cropped 3D patch from the whole CT volume [8, 22]. Note that semi-supervision and fully-supervision are two extreme cases of partial supervision, when the set of partial labels is an empty set ($\mathcal{L}_{\mathrm{P}_t} = \oslash$) and is equal to the complete set ($\mathcal{L}_{\mathrm{P}_t} = \mathcal{L}$), respectively.

A naive solution is to simply train a segmentation network from both the fully-labeled data and the partially-labeled data and alternately update the network parameters and the segmentations (pseudo-labels) for the partially-labeled data [44, 2]. While these EM-like approaches have achieved significant improvement compared with fully-supervised methods, they require high-quality pseudo-labels and fail to explicitly incorporate anatomical priors on shape or size.

To address this issue, we propose a Prior-aware Neural Network (PaNN), aiming at explicitly embedding anatomical priors without incurring any additional budget. More specifically, the anatomical priors are enforced by introducing an additional penalty which acts as a soft constraint to regularize that the average output distributions of organ sizes should mimic their empirical proportions. This prior is obtained by calculating the organ size statistics of the fully-labeled dataset. An overview of the overall framework is shown in Fig. 2, and the detailed training procedures will be introduced in the following sections.

3.2. Prior-aware Loss

Consider a segmentation network parameterized by $\mathbf{\Theta}$, which outputs probabilities $\mathbf{p}$. Let $\mathbf{q} \in \mathbb{R}^{(|\mathcal{L}|+1)\times 1}$ be the label distribution in the fully-labeled dataset, with q^l describing the proportion of the l-th label (organ). Then, we estimate the average predicted distribution of the pixels in the partially-labeled datasets as

$$\bar{\mathbf{p}} = \frac{1}{N} \sum_{t=1}^{T} \sum_{i \in \mathrm{P}_t} \sum_{j} \mathbf{p}_{ij}, \tag{1}$$

where $\mathbf{p}_{ij} = [p_{ij}^0, p_{ij}^1, ..., p_{ij}^{|\mathcal{L}|}]$ denotes the probability vector of the j-th pixel in the i-th input slice $\mathbf{I}_i$, and N is the total number of pixels/voxels. Recall that T is the total number of partially-labeled datasets.

To embed the prior knowledge, the prior-aware loss is defined as

$$
\begin{aligned}
&\mathrm{KL}_{\mathrm{marginal}}(\mathbf{q}|\bar{\mathbf{p}}) \triangleq \textstyle\sum_l \mathrm{KL}(q^l|\bar{p}^l) \\
&= -\textstyle\sum_l \left(q^l \log \bar{p}^l + (1-q^l)\log(1-\bar{p}^l)\right) + \mathit{const} \\
&= -\{\mathbf{q}\log\bar{\mathbf{p}} + (1-\mathbf{q})\log(1-\bar{\mathbf{p}})\} + \mathit{const},
\end{aligned}
\tag{2}
$$

which measures the matching probability of the two distributions $\mathbf{q}$ and $\bar{\mathbf{p}}$ via Kullback-Leibler divergence. Note that each class is treated as one vs. rest when calculating the matching probabilities. Therein, the rationale of Eq. (2) is that the output distributions $\bar{\mathbf{p}}$ of different organ sizes should approximate their empirical marginal proportions $\mathbf{q}$, which generally reflects the domain-specific knowledge.

Note that $\mathbf{q}$ is a global estimation of label distribution of the fully-labeled training data, which remains unchanged. Consequently, $\mathrm{H}(\mathbf{q})$ is constant which can be omitted during the network training. Nevertheless, we observe that it is still problematic to directly apply stochastic gradient descent, as we will detail in Sec. 3.3.

Specifically in our case, our final training objective is

$$
\min_{\mathbf{\Theta},\mathbf{Y}_{\mathrm{P}}} \mathcal{J}_{\mathrm{L}}(\mathbf{\Theta}) + \lambda_1 \mathcal{J}_{\mathrm{P}}(\mathbf{\Theta},\mathbf{Y}_{\mathrm{P}}) + \lambda_2 \mathcal{J}_{\mathrm{C}}(\mathbf{\Theta}),
\tag{3}
$$

where $\mathcal{J}_{\mathrm{L}}(\mathbf{\Theta})$ and $\mathcal{J}_{\mathrm{P}}(\mathbf{\Theta},\mathbf{Y}_{\mathrm{P}})$ are the cross entropy loss on the fully-labeled data and the partially-labeled data, respectively. And $\mathbf{Y}_{\mathrm{P}}$ denotes the computed pseudo-labels as well as existing partial labels from the partially-labeled dataset(s). Note that the prior-aware loss $\mathcal{J}_{\mathrm{C}}$ is used as a soft global constraint to stablize the training process. Concretely, $\mathcal{J}_{\mathrm{L}}(\mathbf{\Theta})$ is defined as

$$
\mathcal{J}_{\mathrm{L}} = -\frac{1}{N}\sum_{i\in\mathrm{L}}\sum_{j}\sum_{l=0}^{|\mathcal{L}|}\mathbb{1}(y_{ij}=l)\log p_{ij}^{l},
\tag{4}
$$

where p_{ij}^l denotes the softmax probability of the j-th pixel in the i-th image to the l-th category. $\mathcal{J}_{\mathrm{P}}(\mathbf{\Theta},\mathbf{Y}_{\mathrm{P}})$ is given by

$$
\begin{aligned}
\mathcal{J}_{\mathrm{P}} = -\frac{1}{N}\sum_{t=1}^{T}\sum_{i\in\mathrm{P}_t}\sum_{j}\sum_{l=0}^{|\mathcal{L}|}\{&\mathbb{1}(y_{ij}=l)\log p_{ij}^{l} \\
&+\mathbb{1}(y'_{ij}=l)\log p_{ij}^{l}\},
\end{aligned}
\tag{5}
$$

where the first term corresponds to the pixels with their labels $\mathbf{Y}_{\mathrm{P}}$ given, *i.e.*, $y_{ij}\in\mathcal{L}_{\mathrm{P}_t}$. The second term corresponds to unlabeled background pixels, and $\mathbf{Y}_{\mathrm{P}}$ needs to be estimated during the model training as a kind of pseudo-supervision, *i.e.*, $y'_{ij}\in\mathcal{L}-\mathcal{L}_{\mathrm{P}_t}$.

3.3. Derivation

By substituting Eq. (1) into Eq. (2) and expanding $\mathbf{q}, \bar{\mathbf{p}}$ into scalars, we rewrite Eq. (2) as

$$
\begin{aligned}
\mathcal{J}_{\mathrm{C}} = -\sum_{l=0}^{|\mathcal{L}|}\{&q^l \log \frac{1}{N}\sum_{t=1}^{T}\sum_{i\in\mathrm{P}_t}\sum_{j} p_{ij}^{l} + \\
&(1-q^l)\log(1-\frac{1}{N}\sum_{t=1}^{T}\sum_{i\in\mathrm{P}_t}\sum_{j} p_{ij}^{l})\} + \mathit{const}.
\end{aligned}
\tag{6}
$$

From Eq. (2) and Eq. (6) we can see that the average distribution $\bar{\mathbf{p}}$ of organ sizes is inside the logarithmic loss, which is very different from standard machine learning loss such as Eq. (4) and Eq. (5) where the average is outside logarithmic loss. And directly minimizing by stochastic gradient descent is very difficult as the true gradient induced by Eq. (2) is not a summation of independent terms, the stochastic gradients would be intrinsically biased [20].

To remedy this, we propose to optimize the KL divergence term using stochastic primal-dual gradient [20]. Our goal here is to transform the prior-aware loss into an equivalent min-max problem by taking the sample average out of the logarithmic loss. We introduce two auxiliary variables to assist the optimization, *i.e.*, the primal variable α and the dual variable β. First, the following identity holds

$$
-\log\alpha = \max_{\beta}\left(\alpha\beta + 1 + \log(-\beta)\right)
\tag{7}
$$

due to the property of the log function. Based on Eq. (7), we define $\boldsymbol{\nu}\in\mathbb{R}^{|\mathcal{L}|\times 1}$ as the dual variable associated to the primal variable $\bar{\mathbf{p}}$, and define $\boldsymbol{\mu}\in\mathbb{R}^{|\mathcal{L}|\times 1}$ as the dual variable associated to the primal variable $(1-\bar{\mathbf{p}})$. Then, we have

$$
\begin{aligned}
-\log\bar{p}^l &= \max_{\nu^l}\left(\bar{p}^l\nu^l + 1 + \log(-\nu^l)\right) \\
-\log(1-\bar{p}^l) &= \max_{\mu^l}\left((1-\bar{p}^l)\mu^l + 1 + \log(-\mu^l)\right),
\end{aligned}
\tag{8}
$$

where ν^l (or μ^l) denotes the l-th element of $\boldsymbol{\nu}$ (or $\boldsymbol{\mu}$). Substituting them into Eq. (2)/Eq. (6), maximizing the KL divergence is equivalent to the following min-max optimization problem:

$$
\begin{aligned}
&\min_{\mathbf{\Theta}}\max_{\boldsymbol{\nu},\boldsymbol{\mu}} \sum_l q^l\left(\bar{p}^l\nu^l + 1 + \log(-\nu^l)\right) \\
&\qquad + \sum_l (1-q^l)\left((1-\bar{p}^l)\mu^l + 1 + \log(-\mu^l)\right) \\
\Leftrightarrow\ &\min_{\mathbf{\Theta}}\max_{\boldsymbol{\nu},\boldsymbol{\mu}} \sum_l \left(q^l\nu^l - (1-q^l)\mu^l\right)\bar{p}^l + q^l\log(-\nu^l) \\
&\qquad + \sum_l (1-q^l)\left(\mu^l + \log(-\mu^l)\right),
\end{aligned}
\tag{9}
$$

which brings the sample average out of the logarithmic loss. Note that we ignore the constant in the above formulas.

Algorithm 1: The training procedure of PaNN

```
Input:
Fully-labeled training data $\mathbf{S}_{\mathrm{L}}$;
Partially-labeled training data $\mathbf{S}_{\mathrm{P}}$;
Hyperparameters: $\lambda_1$, $\lambda_2$;
Output:
Segmentation model $\mathbf{\Theta}$;
begin
    Train the segmentation model $\mathbf{\Theta}$ on $\mathbf{S}_{\mathrm{L}}$;
    Compute the prior distribution $\mathbf{q}$ on $\mathbf{S}_{\mathrm{L}}$;
    Initialize $\boldsymbol{\nu} = -1/\mathbf{q}$ and $\boldsymbol{\mu} = 1/(1-\mathbf{q})$;
    repeat
        Estimate pesudo-labels $\mathbf{Y}_{\mathrm{P}}$ with $\mathbf{\Theta}$;
        Update $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ via stochastic gradient ascent;
        Update $\mathbf{\Theta}$ via stochastic gradient descent;
return $\mathbf{\Theta}$
```

3.4. Model Training

We consider training a fully convolutional network [21, 6, 30] for multi-organ segmentation, where the input images are either 2D slices [39, 32, 45] or 3D cropped patches [8, 22]. The training procedure can be divided into two stages.

In the first stage, we only train on the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$ by optimizing Eq. (4) via stochastic gradient descent (also means $\lambda_1 = 0$ and $\lambda_2 = 0$ in Eq. (3)). The goal of this stage is to find a proper initialization $\mathbf{\Theta_0}$ for the network weights, which can stabilize the training procedure in the second stage.

In the second stage, we train the model on the union of the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$ and partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$ via Eq. (3). As can be drawn, we have two groups of variables, *i.e.*, the network weights $\mathbf{\Theta}$ and the three auxiliary variables $\{\boldsymbol{\nu}, \boldsymbol{\mu}, \mathbf{Y}_{\mathrm{P}}\}$. We adopt an alternating optimization, which can be decomposed into two subproblems:

- **Fixing $\mathbf{\Theta}$, Updating $\{\boldsymbol{\nu}, \boldsymbol{\mu}, \mathbf{Y}_{\mathrm{P}}\}$.** With the network weights $\mathbf{\Theta}$ given, we can first estimate the pesudo-labels $\mathbf{Y}_{\mathrm{P}}$ of background pixels in the partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$. Meanwhile, the optimization of $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ is a maximization problem. Hence, we do stochastic gradient *ascent* to learn $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$. As for the initialization, we set $\boldsymbol{\nu}$ to $-1/\mathbf{q}$ and set $\boldsymbol{\mu}$ to $-1/(1-\mathbf{q})$, respectively.

- **Fixing $\{\boldsymbol{\nu}, \boldsymbol{\mu}, \mathbf{Y}_{\mathrm{P}}\}$, Updating $\mathbf{\Theta}$.** By fixing the three auxiliary variables, we can then update the network weights $\mathbf{\Theta}$ via the standard stochastic gradient *descent*.

As can be seen, our algorithm is formulated as a min-max optimization. We summarize the detailed procedure of optimization in Algorithm 1.

4. Experiments

4.1. Experiment Setup

Datasets and Evaluation Metric. We use the training set released in the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge as the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$, which contains 30 abdominal CT scans with 3779 axial contrast-enhanced abdominal clinical CT images in total. For each case, 13 anatomical structures are annotated, including spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava (IVC), portal vein & splenic vein, pancreas, left adrenal gland, right adrenal gland. Each CT volume consists of $85 \sim 198$ slices of 512×512 pixels, with a voxel spatial resolution of $([0.54 \sim 0.54] \times [0.98 \sim 0.98] \times [2.5 \sim 5.0])\mathrm{mm}^3$.

As for the partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$, we use a spleen segmentation dataset[1] (referred as **A**), a pancreas segmentation dataset[2] (referred as **B**) and a liver segmentation dataset[1] (referred as **C**). To make these partially-labeled datasets balanced, 40 cases are evenly selected from each dataset to constitute the partial supervision.

Following the standard cross-validation evaluation [33, 32, 23, 45, 39], we randomly partition the fully-labeled dataset $\mathbf{S}_L$ into 5 complementary folds, each of which contains 6 cases, then apply the standard 5-fold cross-validation. For each fold, we use 4 folds (*i.e.*, 24 cases) as full supervision and test on the remaining fold.

The evaluation metric we use is the Dice-Sørensen Coefficient (DSC), which measures the similarity between the prediction voxel set $\mathcal{Z}$ and the ground-truth set $\mathcal{Y}$. Its mathematical definition is $\mathrm{DSC}(\mathcal{Z}, \mathcal{Y}) = \frac{2 \times |\mathcal{Z} \cap \mathcal{Y}|}{|\mathcal{Z}| + |\mathcal{Y}|}$. We report an average DSC of all the testing cases over the 13 labeled anatomical structures for performance evaluation.

Implementation Details. Similar to [45, 32, 33, 39], we use the soft tissue CT window range of $[-125, 275]$ HU. The intensities of each slice are then rescaled to $[0.0, 255.0]$. Random rotation of $[0, 15]$ is used as an online data augmentation. Our implementations are based on the current state-of-the-art 2D[3] [7, 6] and 3D models[4] [30, 28]. We provide an extensive study about how partially-labeled datasets facilitate multi-organ segmentation task and list thorough comparisons under different settings.

As described in Sec. 3.4, the whole training procedure is divided into two stages. The first stage is the same as fully-supervised training, *i.e.*, we train exclusively on the fully-labeled dataset $\mathcal{S}_{\mathrm{L}}$ for a certain number of iterations M1.

In the second stage, we switch to the min-max optimization on the union of the fully-labeled dataset and partially-labeled datasets for M2 iterations. In each mini-batch, the sampling rate of labeled data and partially-labeled data is $3 : 1$. It has been suggested [2] that it is less necessary

[1] Available at http://medicaldecathlon.com

[2] Available at https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

[3] https://github.com/tensorflow/models/tree/master/research/deeplab

[4] https://github.com/DLTK/DLTK

Model	**Supervision**	**Partially-labeled dataset**			**Average Dice**
		A	B	C	
ResNet50 [12]	Full				0.7535
	Semi [2]	✓			0.7593
			✓		0.7632
				✓	0.7596
		✓	✓	✓	**0.7669**
	Partial **(ours)**	✓			0.7650
			✓		0.7662
				✓	0.7631
		✓	✓	✓	**0.7705**
	PaNN **(ours)**	✓			0.7716
			✓		0.7712
				✓	0.7705
		✓	✓	✓	**0.7833**
ResNet101 [12]	Full				0.7614
	Semi [2]	✓			0.7637
			✓		0.7649
				✓	0.7647
		✓	✓	✓	**0.7719**
	Partial **(ours)**	✓			0.7714
			✓		0.7695
				✓	0.7684
		✓	✓	✓	**0.7735**
	PaNN **(ours)**	✓			0.7770
			✓		0.7819
				✓	0.7748
		✓	✓	✓	**0.7904**
3D-UNet [8]	3D-UNet-fully-sup				0.7066
	Semi [2]	✓	✓	✓	0.7193
	Partial **(ours)**	✓	✓	✓	0.7163
	PaNN **(ours)**	✓	✓	✓	**0.7208**

Table 1. Performance comparison (DSC) with fully-supervised and semi-supervised methods. **Bold underline** denotes the best results, **bold** denotes the second best results.

to update the pseudo-label $\mathbf{Y}_{\mathrm{P}}$ per iteration. Hence, $\mathbf{Y}_{\mathrm{P}}$ is updated every 10K iterations in practice. In addition, the hyperparameters λ_1 and λ_2 are set to be 1.0 and 0.1, respectively. The same decay policy of learning rate is utilized as that used in the first stage. In the second stage, the initial learning rate for the minimization step and the maximization step are set as 10^{-5} and 2×10^{-5}, respectively.

For 2D implementations, the initial learning rate of the first stage is 2×10^{-5} and a *poly* learning rate policy is employed. M1 and M2 are set as 40K and 30K, respectively. Following [33, 7, 14], we apply multi-scale inputs (scale factors are $\{0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$) in both training and testing phase. For 3D implementations, the initial learning rate of the first stage is $5e^{-4}$ and a fixed learning rate policy is employed. M1 and M2 are set as 80K and 100K, respectively.

4.2. Experimental Comparison

We compare the proposed PaNN with a series of state-of-the-art algorithms, including 1) the fully-supervised approach (denoted as "-fully-sup"), where we train exclusively only on the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$, 2) the semi-supervised approach (denoted as "-semi-sup"), where we train the network on both the fully-labeled dataset $\mathbf{S}_{\mathrm{L}}$ and the partially-labeled dataset(s) $\mathbf{S}_{\mathrm{P}}$ while treating $\mathbf{S}_{\mathrm{P}}$ as unlabeled following the representative method [2], and 3) the naive partially-supervised approach (denoted as "-partial-sup"), where we also train the network on both $\mathbf{S}_{\mathrm{L}}$ and $\mathbf{S}_{\mathrm{P}}$ while treating the partial labels as they are. Different from PaNN, we set $\lambda_2 = 0$ in Eq. (3) to verify the efficacy of the prior-aware loss.

Benefit of Partial Supervision. As shown from Table 1, among three kinds of supervisions, partial supervision obtains the best performance followed by the semi-supervision and full supervision. It is no surprise to observe such a phenomenon for two reasons. First, compared with full supervision, semi-supervision has more training data, though part of them is not annotated. Second, compared with semi-supervision, partial supervision involves more annotated pixels in the organ of interest.

Effect of PaNN. From Table 1, PaNN generally achieves better performance than the naive partially-supervised methods, which demonstrates the effectiveness of our proposed PaNN. For example, when setting the partial dataset as the union of **A**, **B** and **C**, PaNN achieves the best result either using 2D models or 3D models. 2D models generally observe a better performance in each setting compared with 3D models. This is probably due to the fact that current 3D models only act on local patches (*e.g.*, $64 \times 64 \times 64$), which results in lacking holistic information [38]. A detailed discussion of 2D and 3D models is listed in [16]. More specifically, PaNN outperforms the naive partially-supervised method by 1.28% with ResNet-50 and by 1.69% with ResNet-101 as the backbone model, respectively. Additionally, we also observe a convincing performance gain of 0.45% using 3D UNet [8, 30] as the backbone model.

Meanwhile, by increasing the number of partially-labeled datasets (from using only **A**, **B** or **C** to the union of three), the performance improvements of different methods are also different. For example, with the ResNet-101 as the backbone, the largest improvement obtained under semi-supervision is 0.82% (from 76.37% to 77.19%), and that of partial supervision is 0.51% (from 76.84% to 77.35%). By contrast, PaNN obtains a much more remarkable improvement of 1.56% (from 77.48% to 79.04%). Such an observation suggests that PaNN is capable of handling more partially-labeled training data and is less susceptible to the background ambiguity.

Organ-by-organ Analysis. To reveal the detailed effect of PaNN, we present an organ-by-organ analysis in Fig. 3. We use ResNet-50 as the backbone model (ResNet-101 has a similar trend) and the partially-labeled dataset **C** (indicates that the liver is the target organ).

In Fig. 3, we observe clear statistical improvements over the fully-supervised method for almost every organ (p-values $p < 0.001$ hold for 11/13 of all abdominal organs). Great improvements are also observed for those difficult or-

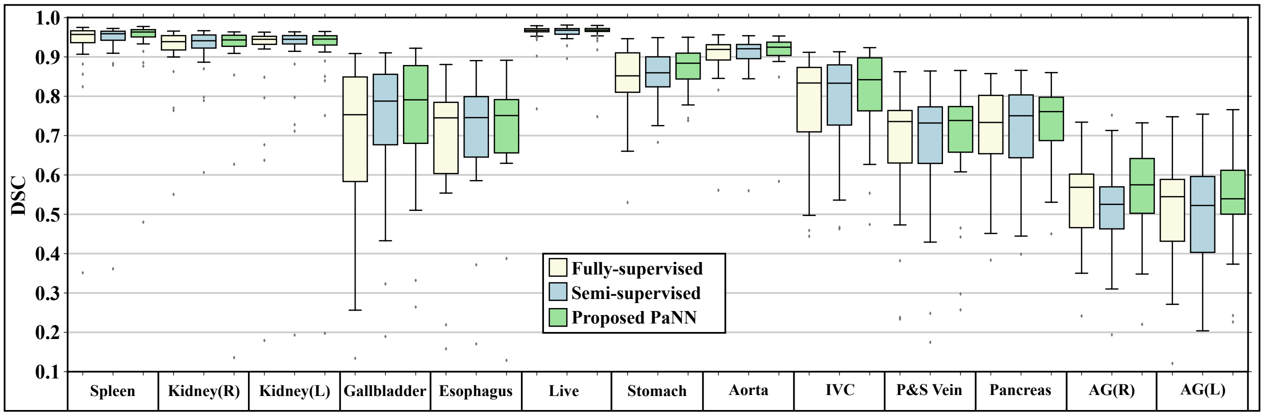

Figure 3. Performance comparison (DSC) in box plots of 13 abdominal structures, where the partially-labeled dataset **C** is used with ResNet-50 as the backbone model. Our proposed PaNN improves the overall mean DSC and also reduces the standard deviation. Kidney/AG (R), Kidney/AG (L) stand for the right and left kidney/adrenal gland, respectively.

Name	Spleen	Kidney(R)	Kidney(L)	Gallbladder	Esophagus	Liver	Aorta	IVC	Average Dice	Mean Surface Distance	Hausdorff Distance
AutoContext3DFCN [33]	0.926	0.866	0.897	0.629	0.727	0.948	0.852	0.791	0.782	1.936	26.095
deedsJointCL [13]	0.920	0.894	0.915	0.604	0.692	0.948	0.857	0.828	0.790	2.262	25.504
dltk0.1_unet_sub2 [28]	0.939	0.895	0.915	**0.711**	0.743	0.962	0.891	0.826	0.815	1.861	62.872
results_13organs_p0.7	0.890	0.898	0.883	0.685	0.754	0.936	0.870	0.819	0.817	4.559	38.661
PaNN* (**ours**)	**0.961**	**0.901**	**0.943**	0.704	**0.783**	**0.972**	**0.913**	**0.835**	**0.832**	**1.641**	**25.176**
PaNN (**ours**)	**0.968**	**0.920**	**0.953**	**0.729**	**0.790**	**0.974**	**0.925**	**0.847**	**0.850**	**1.450**	**18.468**

Table 2. Performance comparison on the 2015 MICCAI Multi-Atlas Abdomen Labeling challenge leaderboard. Our method achieves the largest Dice score and the smallest average surface distances and Hausdorff distances. PaNN* only uses 80% of the training data as the fully-supervised dataset and use the rest 20% data as partially-labeled data (by randomly removing labels of 8/13 organs), without using extra data. In this table, we only show 8/13 organs' average Dice scores due to the space limit.

gans, *i.e.*, organs either in small sizes or with complex geometric characteristics such as gallbladder (from 67.26% to 72.26%), esophagus (from 69.35% to 71.21%), stomach (from 84.09% to 87.21%), IVC (from 77.34% to 80.70%), portal vein & splenic vein (from 66.74% to 68.75%), pancreas (from 71.45% to 73.62%), right adrenal gland (from 53.65% to 55.56%) and left adrenal gland (from 49.51% to 53.63%). This promising result indicates that our method distills a reasonable amount of knowledge from additional partially-labeled data and the regularization loss can help facilitate the network to enhance the discriminative information to a certain degree.

Meanwhile, we also observe a distinct performance improvement for organs other than the partially-labeled structures (*i.e.*, the liver). For instance, the performance of gallbladder, stomach, IVC, pancreas are boosted from 68.97%, 85.57%, 78.59%, 71.94% to 72.26%, 87.21%, 80.70%, 73.62%, respectively. This suggests that the superiority of PaNN not only originates from more training data, but also from the fact that PaNN can effectively incorporate anatomical priors on organ sizes in abdominal regions, which is helpful for multi-organ segmentation.

Qualitative Evaluation. We also show a set of qualitative examples, *i.e.*, 5 slices from 3 cases, in Fig. 4, where we zoom in to visualize the finer details of the improved region.

In these samples, we observe that PaNN is the only method that successfully detects the pancreatic tail in Fig. 4(a). In Fig. 4(b), all other methods fail to detect the portal vein and splenic vein while PaNN demonstrates an almost perfect detection of these veins. For Fig. 4(c) to Fig. 4(e), apart from the evident improvements of the pancreas, left adrenal gland, one of the smallest abdominal organs, is also clearly segmented by PaNN.

4.3. MICCAI 2015 Multi-Atlas Labeling Challenge

We test our model in the 2015 MICCAI Multi-Atlas Abdomen Labeling challenge. The top model (denoted as "PaNN" in Table 2) we submit is based on ResNet-101, and trained on all 30 cases of the fully-labeled dataset $\mathbf{S}_L$ and the union of three partially-labeled datasets **A**, **B** and **C**. The evaluation metric employed in this challenge includes the Dice scores, average surface distances [32] and Hausdorff distances [22]. We compare PaNN with the other top submissions of the challenge leaderboard in Table 2. As it shows, the proposed PaNN achieves the best performance under all the three evaluation metrics, easily surpassing prior best result by a large margin. **Without using any additional data and even randomly removing partial labels** from the challenge data, our method (denoted as "PaNN*" in Table 2) still obtains the state-of-the-art result of 83.17%, outperforming the previous best result of

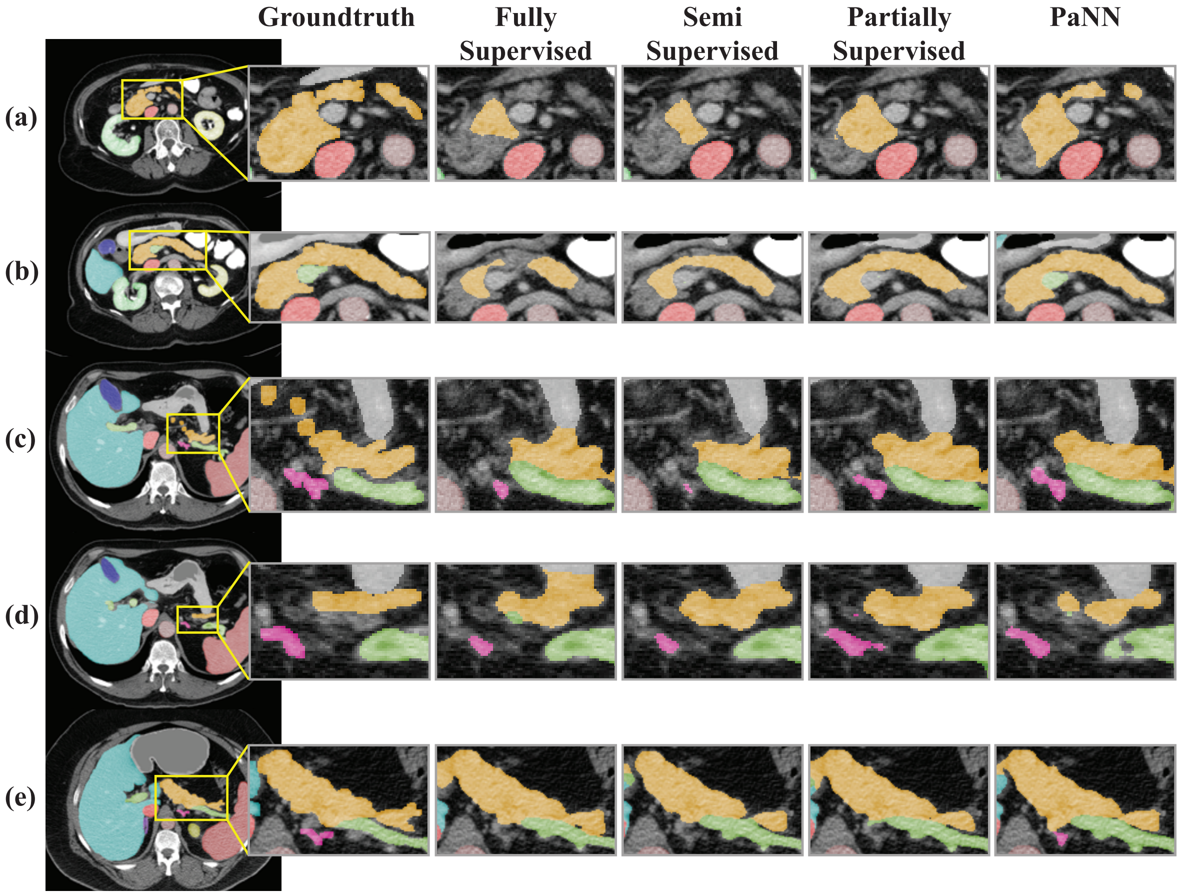

Figure 4. Qualitative comparison of different methods, where the partially-labeled dataset **C** is used as partial supervision with ResNet-101 as the backbone model. We exhibit 3 cases (5 slices) as examples. Improved segmentation regions are zoomed in from the axial view to demonstrate finer details.

Organ	Fully Supervised	Semi Supervised	Partially Supervised (**ours**)	PaNN (**ours**)
Gallbladder	0.8225	0.8399	0.8465	**0.8467**
Aorta	0.9110	0.9096	0.9121	**0.9133**
IVC	0.8083	0.8175	0.7995	**0.8266**
Pancreas	0.7831	0.7994	0.8079	**0.8193**
avg. Dice	0.9008	0.9060	0.9063	**0.9103**

Table 3. Performance comparison on a newly collected dataset. Full results are included in the supplementary material.

DLTK UNet [28] by 2% in average Dice. It is noteworthy that our method is far from its potential maximum performance as we only use 2D single view algorithms. It is suggested [45, 38, 44] that using multi-view algorithms or model ensemble can boost the performance further.

4.4. Generalization to Other Datasets

We also apply our algorithm to a different set of abdominal clinical CT images, where 20 cases are used for training and 15 cases are used for testing. A total of 9 structures (spleen, right kidney, left kidney, gallbladder, liver, stomach, aorta, IVC, pancreas) are manually labeled. Each case was segmented by four experienced radiologists, and confirmed by an independent senior expert. Each CT volume consists of $319 \sim 1051$ slices of 512×512 pixels, and has voxel spatial resolution of $([0.523 \sim 0.977] \times [0.523 \sim 0.977] \times 0.5) mm^3$. We use the union of all 3 datasets **A**, **B**, and **C** as the partial supervision. The results are summarized in Table 3, where the proposed PaNN also achieves better results compared with existing methods.

5. Conclusion

In this work, we have presented PaNN, for multi-organ segmentation, as a way to better utilize existing partially-labeled datasets. In several applications such as radiation therapy or computer-aided surgery, physicians and surgeons have been doing segmentation of target structures. Meanwhile, to handle the background ambiguity brought by the partially-labeled data, the proposed PaNN exploits the anatomical priors by regularizing the organ size distributions of the network output should approximate their prior statistics in the abdominal region. Our proposed PaNN shows promising results using state-of-the-art models.

Acknowledgements. This work was partly supported by the Lustgarten Foundation for Pancreatic Cancer Research.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[2] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *MICCAI*, 2017. 2, 3, 5, 6

[3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. Whats the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565, 2016. 1

[4] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 2017. 1, 2

[5] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, 2016. 2

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018. 2, 5

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. 2018. 2, 5, 6

[8] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 2016. 3, 5, 6

[9] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015. 2

[10] A. V. Dalca, J. Guttag, and M. R. Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018. 1, 2, 3

[11] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng. 3d deeply supervised network for automatic liver segmentation from ct volumes. In *MICCAI*, 2016. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[13] M. P. Heinrich. Multi-organ segmentation using deeds, self-similarity context and joint fusion. 2015. 7

[14] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017. 2, 6

[15] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical Image Analysis*, 2019. 2, 3

[16] M. Lai. Deep learning for medical image segmentation. *arXiv preprint arXiv:1505.02000*, 2015. 6

[17] D. Lin, J. Dai, J. Jia, K. He, and J. 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. 1

[18] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 2

[19] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 3

[20] Y. Liu, J. Chen, and L. Deng. An unsupervised learning method exploiting sequential output statistics. In *NIPS*, 2017. 2, 4

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 5

[22] F. Milletari, N. Navab, and S. A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, 2016. 3, 5, 7

[23] I. Nogues, L. Lu, X. Wang, H. Roth, G. Bertasius, N. Lay, J. Shi, Y. Tsehay, and R. M. Summers. Automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in ct images. In *MICCAI*, 2016. 5

[24] M. S. Nosrati and G. Hamarneh. Incorporating prior knowledge in medical image segmentation: a survey. *arXiv preprint arXiv:1607.01092*, 2016. 2

[25] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. ORegan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2018. 1, 2, 3

[26] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015. 1, 2, 3

[27] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. 2015. 1, 2

[28] N. Pawlowski, S. I. Ktena, M. C. Lee, B. Kainz, D. Rueckert, B. Glocker, and M. Rajchl. Dltk: State of the art reference implementations for deep learning on medical images. *arXiv preprint arXiv:1711.06853*, 2017. 5, 7, 8

[29] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2):674–683, 2017. 2, 3

[30] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1, 2, 5, 6

[31] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *MICCAI*, 2015. 1, 3

[32] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Medical image analysis*, 45:94–107, 2018. 1, 2, 3, 5, 7

[33] H. R. Roth, C. Shen, H. Oda, T. Sugino, M. Oda, Y. Hayashi, K. Misawa, and K. Mori. A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation. In *MICCAI*, 2018. 1, 3, 5, 6, 7

[34] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 1, 3

[35] N. Souly, C. Spampinato, and M. Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017. 2

[36] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *CVPR*, pages 1818–1827, 2018. 1

[37] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov. On regularized losses for weakly-supervised cnn segmentation. In *ECCV*, pages 507–522, 2018. 1

[38] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *arXiv preprint arXiv:1804.08414*, 2018. 6, 8

[39] Y. Wang, Y. Zhou, P. Tang, W. Shen, E. K. Fishman, and A. L. Yuille. Training multi-organ segmentation networks with sample selection by relaxed upper confident bound. In *MICCAI*, 2018. 3, 5

[40] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and i will show you where it is. In *CVPR*, pages 3190–3197, 2014. 1

[41] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *CVPR*, pages 8280–8289. 1

[42] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI*, 2017. 3

[43] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2

[44] Y. Zhou, Y. Wang, P. Tang, S. Bai, W. Shen, E. Fishman, and A. Yuille. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140. IEEE, 2019. 3, 8

[45] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille. A fixed-point model for pancreas segmentation in abdominal ct scans. In *MICCAI*, 2017. 1, 3, 5, 8