Medical Image Analysis 70 (2021) 101979

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Marginal loss and exclusion loss for partially supervised multi-organ segmentation

Gonglei Shi[a,b], Li Xiao[a,*], Yang Chen[b], S. Kevin Zhou[a,c,*]

[a] Medical Imaging, Robotics, Analytic Computing Laboratory & Engineering (MIRACLE), Key Lab of Intelligent Information Processing of Chinese Academy of Sciences, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[b] School of Computer Science and Engineering, Southeast University, Nanjing, 210000, China
[c] School of Biomedical Engineering & Suzhou Institute For Advanced Research, University of Science and Technology, Suzhou, 215123, China

ARTICLE INFO

Article history:
Received 5 July 2020
Revised 2 December 2020
Accepted 20 January 2021
Available online 3 February 2021

Keywords:
Multi-organ segmentation
Partially labeled dataset
Marginal loss
Exclusion loss

ABSTRACT

Annotating multiple organs in medical images is both costly and time-consuming; therefore, existing multi-organ datasets with labels are often low in sample size and mostly partially labeled, that is, a dataset has a few organs labeled but not all organs. In this paper, we investigate how to learn a single multi-organ segmentation network from a union of such datasets. To this end, we propose two types of novel loss function, particularly designed for this scenario: (i) marginal loss and (ii) exclusion loss. Because the background label for a partially labeled image is, in fact, a 'merged' label of all unlabelled organs and 'true' background (in the sense of full labels), the probability of this 'merged' background label is a marginal probability, summing the relevant probabilities before merging. This marginal probability can be plugged into any existing loss function (such as cross entropy loss, Dice loss, etc.) to form a marginal loss. Leveraging the fact that the organs are non-overlapping, we propose the exclusion loss to gauge the dissimilarity between labeled organs and the estimated segmentation of unlabelled organs. Experiments on a union of five benchmark datasets in multi-organ segmentation of liver, spleen, left and right kidneys, and pancreas demonstrate that using our newly proposed loss functions brings a conspicuous performance improvement for state-of-the-art methods without introducing any extra computation.

1. Introduction

Multiple organ segmentation has been widely used in clinical practice, including diagnostic interventions, treatment planning, and treatment delivery (Ginneken et al., 2011; Sykes, 2014). It is a time-consuming task in radiotherapy treatment planning, with manual or semi-automated tools (Heimann and et al., 2009) frequently employed to delineate organs at risk. Therefore, to increase the efficiency of organ segmentation, auto-segmentation methods such as statistical models (Cerrolaza et al., 2015; Okada et al., 2015), multi-atlas label fusion (Xu et al., 2015; Tong et al., 2015; Suzuki et al., 2012), and registration-free methods (Saxena et al., 2016; Lombaert et al., 2014; He et al., 2015) have been developed. Unfortunately, these methods are likely affected by image deformation and inter-subject variability and their success in clinical applications is limited.

Deep learning based medical image segmentation methods have been widely used in the literature to perform the classification of each pixel/voxel for a given 2D/3D medical image and has significantly improved the performance of multi-organ auto-segmentation. One prominent model is U-Net (Ronneberger et al., 2015), along with its latest variant nnUNet (Isensee et al., 2018), which learns multiscale features with skip connections. Other frameworks for multi-organ segmentation include (Wang et al., 2019; Binder et al., 2019; Gibson et al., 2018). There is a rich body of subsequent works (Okada et al., 2012; Chu et al., 2013; Suzuki et al., 2012; Liu et al., 2020; Gibson et al., 2018), focusing on improving existing frameworks by finding and representing the interrelations based on canonical correlation analysis especially by constructing and utilizing the statistical atlas.

However, almost all current segmentation models rely on fully annotated data (Zhao et al., 2019; Chen et al., 2018; Yang et al., 2017) with strong supervision. To curate a large-scale fully annotated dataset is a challenging task, both costly and time-consuming. It is also a bottleneck in the multi-organ segmentation research area that current labeled data sets are often low in sample size and mostly partially labeled. That is, a data set has a few

* Corresponding author.
E-mail addresses: xiaoli@ict.ac.cn (L. Xiao), zhoushaohua@ict.ac.cn (S.K. Zhou).

https://doi.org/10.1016/j.media.2021.101979
1361-8415/

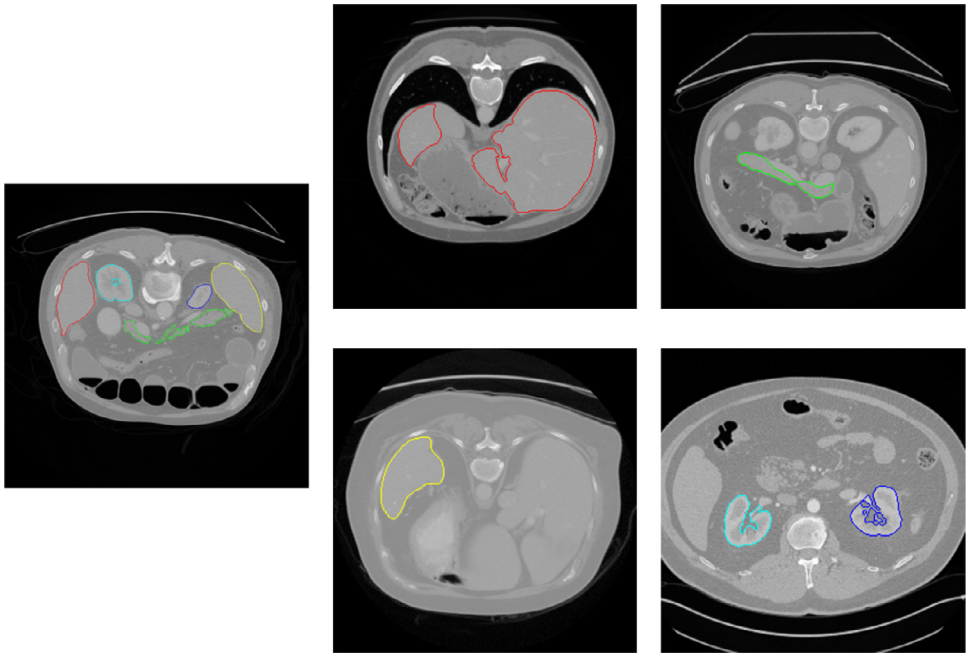

Fig. 1. Five typical annotated images from five different datasets, one image per dataset. The colored edges show the annotated organ boundaries (red for liver, yellow for spleen, green for pancreas, blue for left kidney, and cyan for right kidney). The left image shows the case of a fully annotated data set and the amount of such data set is usually small. The right four images are partially labeled. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

organs labeled but not all organs (as shown in Fig. 1). Such partially annotated datasets obviate the use of segmentation methods that require full supervision.

It becomes a research problem of practical need on how to make full use of these partially annotated data to improve the segmentation accuracy and robustness. In the case of sufficient network model capabilities, a larger amount of data typically means that it is more likely to represent the actual distribution of data in reality, hence leading to better overall performance. Motivated by this, in this paper we investigate how to learn a **single multi-organ segmentation network** from the union of such partially labeled data sets. Such learning does not introduce any extra computation.

To this end, we propose two types of loss functions particularly designed for this task: (i) **marginal loss** and (ii) **exclusion loss**. Firstly, because the background label for a partially labeled image is, in fact, a 'merged' label of all unlabeled organs and 'true' background (in the sense of full labels), the probability of this 'merged' background label is a marginal probability, summing the relevant probabilities before merging. This marginal probability can be plugged into any existing loss function such as cross entropy (CE) loss, Dice loss, etc. to form a **marginal loss**. In this paper, we propose to use marginal cross entropy loss and marginal Dice loss in the experiment. Secondly, in multi-organ segmentation, there is a one-to-one mapping between pixels and labels, different organs are mutually exclusive and not allowed to overlap. This leads us to propose the **exclusion loss**, which adds the exclusiveness as prior knowledge on each labeled image pixel. In this way, we make use of the explicit relationships of given ground truth in partially labeled data, while mitigating the impact of unlabeled categories on model learning. Using the state-of-the-art network model (e.g., nnUNet (Isensee et al., 2018)) as the backbone, we successfully learn a single multi-organ segmentation network that outputs the full set of organ labels (plus background) from a union of five benchmark organ segmentation datasets from different sources. Refer to Fig. 1 for image samples from these datasets.

In the following, after a brief survey of related literature in Section 2, we provide the derivation of marginal loss and exclusion loss in Section 3. The two types of loss function can be applied to pretty much any loss function that relies on posterior class probabilities. In Section 4, extensive experiments are then presented to demonstrate the effectiveness of the two loss functions. By successfully pooling together partially labeled datasets, our new method can achieve significant performance improvement, which is essentially **a free boost** as these auxiliary datasets are existent and already labeled. Our method outperforms two state-of-the-art models (Zhou et al., 2019; Fang and Yan, 2020) for partially annotated data learning. We conclude the paper in Section 5.

2. Related work

2.1. Multi-organ segmentation models

Many pioneering works have been done on multi-organ segmentation, using traditional machine learning methods or deep learning methods. In (Okada et al., 2015; Xu et al., 2015; Tong et al., 2015; Suzuki et al., 2012; Shimizu et al., 2007; Wolz et al., 2013), a multi-altas based strategy is used for segmentation, which registers an unseen test image with multiple training images and use the registration map to propagate the labels in the training images to generate final segmentation. However, its performance is limited by image registration quality. In (Heimann and Meinzer, 2009; Cootes et al., 2001; Chen et al., 2012), prior knowledge of statistical models is employed to achieve multi-organ segmentation. There are also some methods that directly use deep learning semantic segmentation networks for multi-organ segmentation (Gibson et al., 2018; Wang et al., 2019; Kohlberger et al., 2011; Lay et al., 2013). Besides, there are prior approaches that combine the above-mentioned different methods (Chu et al., 2013; Lu et al., 2012) to achieve better multi-organ segmentation. However, all these methods rely on the availability of fully labelled images.

2.2. Multi-organ segmentation with partially annotated data learning

Very limited works have been done on medical image segmentation with partially-supervised learning. Zhou et al. (Zhou et al., 2019) learns a segmentation model in the case of partial labeling by adding a prior-aware loss in the learning objective to match the distribution between the unlabeled and labeled datasets. However, it use the predicted value of the network as pseudo-label on the missing partial label data, and hence involves extra memory and time consumption. Instead, our work trains a single multi-class network. Since only two loss terms are added, it needs nearly no additional training time and memory cost. Dmitriev et al. (Dmitriev and Kaufman, 2019) propose a unified, highly efficient segmentation framework for robust simultaneous learning of multi-class datasets with missing labels. But the network can only learn from datasets with single-class labels. Fang et al. (Fang and Yan, 2020) hierarchically incorporate multi-scale features at various depths for image segmentation, further develop a unified segmentation strategy to train three separate datasets together, and finally achieve multi-organ segmentation by learning from the union of partially labeled and fully labeled datasets. Though this paper also uses a loss function that amounts to our marginal cross entropy, its main focus is on proposing the hierarchical network architecture. In contrast, we concentrate on studying the impact of the marginal loss including both marginal cross entropy and marginal Dice loss. Furthermore, it is worth mentioning that none of the above works considers the mutual exclusiveness, a well-known attribute between different organs. We propose a novel exclusion loss term, exploiting the fact that organs are mutually exclusive and adding the exclusiveness as prior knowledge on each image pixel.

2.3. Partially annotated data learning in other tasks

A few existing methods have been developed on classification and object detection tasks using partially annotated data. Yu et al. (Yu et al., 2014) propose an empirical risk minimization framework to solve multi-label classification problem with missing labels; Wu et al.(Wu et al., 2015) train a classifier with multi-label learning with missing labels to improve object detection problem. Cour et al. (Cour et al., 2011) propose a convex learning formulation based on the minimization of a loss function appropriate for the partially labeled setting. Besides, as far as semi-supervised learning is concerned, a number of researches have been developed to solve (He et al., 2019; Zhu et al., 2018; Xiao et al., 2019) classification problems or detection problems in the absence of annotations.

3. Method

The goal of our work is to train a single multi-class segmentation network Ψ by using a large number of partially annotated data in addition to a few fully labeled data for baseline training. Learning under such a setup is enabled by the novel losses we propose below.

Segmentation is achieved by grouping pixels (or voxels) of the same label. A labeled pixel has two attributes: (i) pixel and (ii) label. Therefore, it is possible to improve the segmentation performances by exploiting the pixel or label information. To be more specific, we leverage some prior knowledge on each image pixel, such as its anatomical location or its relation with other pixels, to facilitate the network for better segmentation; we also merge or split labels to help the network focus more on specific task requirements. In this work, we apply the two ideas on multi-organ segmentation tasks as follows. Firstly, due to a large amount of partially labeled images, we merge all unlabeled organ pixels with the background label, which forms a **marginal loss**. Secondly, regarding a well known prior knowledge that organs are mutually exclusive, we design an **exclusion loss**, which adds exclusion information on each image pixel, to further reduce the segmentation errors.

3.1. Regular cross-entropy loss and regular dice loss

The loss function is generally proposed for a specific problem. A common idea for loss functions are based on classification tasks which optimize the intra-class difference and reduce the intra-class variation, for example contrastive loss (Hadsell et al., 2006), triplet Loss (Schroff et al., 2015), center loss (Wen et al., 2016), large margin softmax loss (Liu et al., 2016), angular softmax (Li et al., 2018) and cosine embeding loss (Wang et al., 2018). The cross entropy loss (Long et al., 2015) is the most representative loss function, which is commonly used in deep learning. There are also some loss functions designed to optimize the global performance for semantic segmentation, such as Dice loss (Long et al., 2015), Tversky loss (Salehi et al., 2017), combo loss (Taghanaki et al., 2019), Lovasz-Softmax loss (Berman et al., 2018). Besides, some losses are proposed specifically to improve a given loss function, for example, the focal loss (Lin et al., 2017) is developed based on cross-entropy loss (Long et al., 2015) to better solve class imbalance problem. Here we focus on the cross-entropy loss and regular Dice loss that are most commonly used in multi-organ segmentation.

Suppose that, for a multi-class classification task with N labels with its label index set as $\Omega_N = \{C_1, C_2, \ldots, C_N\}$, its data sample x (i.e., an image pixel in image segmentation) belongs to one of N classes, say class C_n, which is encoded as an N-dimensional one-hot vector $\hat{y}_n = [y_1, y_2, \ldots, y_N]$ with $y_n = 1$ and all others 0. A multi-class classifier consists of a set of response functions $\{a_n(x);\ n \in \Omega_N\}$, which constitutes the outputs of the segmentation network Ψ. From these response functions, the posterior classification probabilities are computed by a softmax function,

$$p_n = \frac{\exp(a_n)}{\sum_{k\in\Omega_N}\exp(a_k)}, \quad n \in \Omega_N. \tag{1}$$

To learn the classifier, the regular cross-entropy loss is often used, which is defined as follows:

$$L_{rCE} = -\sum_{n\in\Omega_N} y_n \log(p_n). \tag{2}$$

Besides, the Dice score coefficient (DSC) is often used, which measures the overlap between the segmentation map and ground truth. The dice loss is defined as $1 - DSC$:

$$L_{rDice} = \sum_{n\in\Omega_N}(1 - 2\cdot\frac{y_n\ p_n}{y_n + p_n}) \tag{3}$$

3.2. Marginal loss

For an image with incomplete segmentation label, it is possible that the pixels for some given classes are not 'properly' provided. To deal with such situations, we assume that there are a reduced number of $M < N$ classes in a partially-labeled dataset with its corresponding label index set as $\Omega'_M = \{C'_1, C'_2, \ldots, C'_M\}$. For each merged class label $m \in \Omega'_M$, there is a corresponding subset $\Phi_m \subset \Omega_N$, which is comprised of all the label indexes in Ω_N that can be merged into the same class m. Because the labels are exclusive in multi-organ segmentation, we have $\Omega_N = \cup_{m\in\Omega'_M}\Phi_m$.

Fig. 2 illustrates the process of label merging, using an example of four organ classes $C_i, i = 1, 2, 3, 4$. After the merging, there are two classes C'_1 and C'_2, with C_1 and C_2 are combined together to form a new merged label C'_1 and C_3 and C_4 to form a new label C'_2.

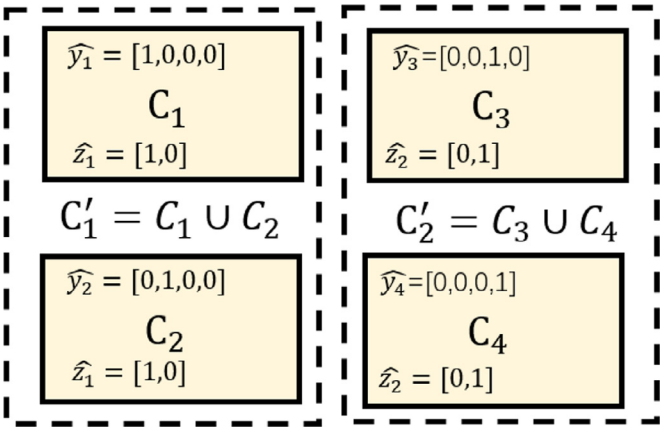

Fig. 2. Venn Diagram to illustrate the marginal loss. The dataset contains three classes C_1, C_2, C_3, C_4, the partially labeled dataset only contains two class labels, with C_1 and C_2 merged together as C_1' and C_3 and C_4 merged together as C_2'.

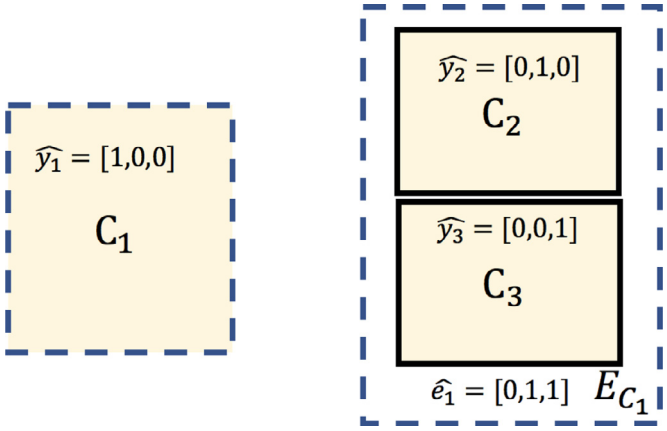

Fig. 3. Venn Diagram to illustrate the exclusion loss. There are three mutually exclusive classes C_1, C_2, and C_3. The exclusion set for C_1 is $E_{C_1} = C_2 \cup C_3$.

The classification probability for the merged class m is a marginal probability

$$q_m = \sum_{n \in \Phi_m} p_n. \tag{4}$$

Also, the one-hot vector for a class $m \in \Omega_M'$ is denoted as $\hat{z}_m = [z_1, z_2, \ldots, z_M]$, which is M-dimensional with $z_m = 1$ and all others 0.

Consequently, we define marginal cross-entropy loss and marginal Dice loss as follows:

$$L_{mCE} = -\sum_{m \in \Omega_M'} z_m \log(q_m). \tag{5}$$

$$L_{mDice} = \sum_{m \in \Omega_M'} \left(1 - 2 \cdot \frac{z_m\, q_m}{z_m + q_m}\right). \tag{6}$$

We use marginal cross entropy as an example to perform the gradient calculation. Firstly, referring to Eqs. (1) and (4), the gradient of the output m of a softmax node to the network node a_j is:

$$\frac{\partial q_m}{\partial a_j} = \sum_{n \in \Phi_m} \frac{\partial p_n}{\partial a_j} = p_j[\pi_j(\Phi_m) - q_m], \tag{7}$$

where $\pi_j(\Phi_m)$ is a boolean indicator function that tells if j is in Φ_m. p and q are the classification probabilities of regular and marginal softmax functions.The derivative gradient of L_{mCE} to the network node a_j is

$$\begin{aligned} \frac{\partial L_{mCE}}{\partial a_j} &= -\sum_{m \in \Omega_M'} \frac{z_m}{q_m} \frac{\partial q_m}{\partial a_j} \\ &= -\sum_{m \in \Omega_M'} \frac{z_m}{q_m} p_j[\pi_j(\Phi_m) - q_m] = \left[1 - \frac{z_{\bar{m}}}{q_{\bar{m}}}\right] p_j, \end{aligned} \tag{8}$$

where $\bar{m}$ is the only class index that makes $\pi_j(\Phi_m) = 0$.

3.3. Exclusion loss

It happens in multi-organ segmentation tasks that some classes are mutually exclusive to each other. The exclusion loss is designed to add the exclusiveness as an additional prior knowledge on each image pixel. We define an exclusion subset for a class n as E_n, which comprises all (or a part of) the label indexes that are mutually exclusive with class n. The exclusion label information is encoded as an N-dimensional vector $\hat{e}_n = [e_1, e_2, \ldots, e_N]$, which is obtained as:

$$\hat{e}_n = \sum_{k \in E_n} \hat{y}_k. \tag{9}$$

Note that $\hat{e}_n$ is still an N-dimensional vector, but it is not an one-hot vector any more. Fig. 3 shows the procedure of applying exclusion loss. Assuming that organ classes C_1, C_2 and C_3 are mutually exclusive, the labels of C_2 and C_3 form the exclusion subset E_{C_1}.

We expect that the intersection between the segmentation prediction p_n from the network and e_n is as small as possible. Following the Dice coefficient, the formula for the exclusion Dice loss is given as:

$$L_{eDice} = \sum_{n \in \Omega_N} 2 \cdot \frac{e_n \cdot p_n}{e_n + p_n}. \tag{10}$$

The exclusion cross-entropy loss is defined accordingly:

$$L_{eCE} = \sum_{n \in \Omega_N} e_n \log(p_n + \epsilon), \tag{11}$$

where ϵ is introduced to avoid the trap of $-\infty$. We set $\epsilon = 1$.

4. Experiments and results

4.1. Problem setting and benchmark dataset

We consider a partially-supervised multi-organ segmentation task that is common in practice (such as Fig. 1). For each partially annotated image, we restrict it with only one label. For clarity of description, we assume that F denotes the fully-labeled segmentation dataset and $P_i; i \in \{1, 2, \ldots, C\}$ denotes a dataset of partially-annotated images that contain only a partial list of organ label(s). The datasets $P_{1:C}$ do not overlap in terms of their organ labels. For an image in P_i, there is a 'merged' background, which is the union of real background and missing organ labels. We jointly learn a single segmentation network Ψ using $F \cup P_1 \cup \ldots \cup P_C$, assisted by the proposed loss functions.

For our experiments, we choose liver, spleen, pancreas, left kidney and right kidney as the segmentation targets and use the following benchmark datasets.

- Dataset F. We use Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge (Landman et al., 2017) as fully annotated base dataset F. It is composed of 30 CT images with segmentation labels of 13 organs, including liver, spleen, right kidney, left kidney, pancreas, and other organs (gallbladder, esophagus, stomach, aorta, inferior vena cava, portal vein and splenic vein, right adrenal gland, and left adrenal gland) we hereby ignore.
- Dataset P_1. We refer to the task03 liver dataset from the Decathlon-10 (Simpson et al., 2019) challenge as P_1. It is composed of 131 CT's with annotations for liver and liver cancer re-

Table 1
Usage of experimental dataset.

Network	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$
Ψ_F^{mc}: multiclass (F)	√		√		√		√	√		
$\Psi_{F+P_1}^{b}$: binary liver ($F+P_1$)	√	√								
$\Psi_{F+P_2}^{b}$: binary spleen ($F+P_2$)			√	√						
$\Psi_{F+P_3}^{b}$: binary pancreas ($F+P_3$)					√	√				
$\Psi_{F+P_4}^{b}$: binary kidney ($F+P_4$)							√	√	√	√
$\Psi_{P_1}^{b}$: binary liver (P_1)		√								
$\Psi_{P_2}^{b}$: binary spleen (P_2)				√						
$\Psi_{P_3}^{b}$: binary pancreas (P_3)						√				
$\Psi_{P_4}^{t}$: ternary kidney (P_4)									√	√
Ψ_{All}^{mc}: multiclass ($F+P_{1:4}$)	√	√	√	√	√	√	√	√	√	√
total # of training CT	24	105	24	33	24	225	24	24	168	168
total # of testing CT	6	26	6	8	6	56	6	6	42	42

Table 2
A summary description of the datasets.

Dataset	Modiality	Num of labeled samples	Annotated organs	axis	image voxel range	spacing range
MALBCVWC	CT	30	liver / right kidney / left kidney / /pancreas /spleen / other structures	z y x	85 ~ 198 512 512	2.50 ~ 5.00 0.59 ~ 0.98 0.59 ~ 0.98
Decathlon-Liver	CT	126	liver	z y x	74 ~ 984 512 512	0.70 ~ 5.00 0.56 ~ 1.00 0.56 ~ 1.00
Decathlon-Spleen	CT	41	spleen	z y x	31 ~ 168 512 512	1.50 ~ 8.00 0.61 ~ 0.98 0.61 ~ 0.98
Decathlon-Pancreas	CT	281	pancreas	z y x	31 ~ 751 512 512	0.70 ~ 7.50 0.61 ~ 0.98 0.61 ~ 0.98
KiTS	CT	210	left kidney and right kidney	z y x	29 ~ 1059 512 512	0.50 ~ 5.00 0.44 ~ 1.04 0.44 ~ 1.04

gion. We merge the cancer label into the liver label and obtain a binary-class (liver vs background) dataset.
- Dataset P_2. We refer to the task09 spleen dataset from the Decathlon-10 challenge as P_2. It includes 41 CT's with spleen segmentation label.
- Dataset P_3. We refer to the task07 pancreas dataset from the Decathlon-10 challenge as P_3. It includes 281 CT's with pancreas and its cancer segmentation label. The cancer label is merged into the pancreas label to obtain a binary-class (pancreas vs background) dataset.
- Dataset P_4. We refer to KiTS (Heller et al., 2019) challenge dataset as P_4. Since the offered 210 CT segmentation makes no distinction between left and right kidneys, we manually divide it into left and right kidneys according to the connected component. Cancer label is merged into the according kidney label.

The spatial resolution of all these datasets are resampled to $(1.5 \times 1.5 \times 3)mm^3$. We split the datesets into training and testing. we randomly choose 6 samples from F, 26 samples from P_1 and 8 samples from P_2, 56 samples from P_3 and 42 samples from P_4 as testing. The others are used for training. Table 2 also provides a summary description of the datasets.

4.2. Segmentation networks

We set up the training of 10 deep segmentation networks for comparison as in Table 1.

- Ψ_F^{mc}: a multiclass segmentation network based on F.
- $\Psi_{P_1}^{b}$: a binary segmentation network for liver only based on P_1.
- $\Psi_{P_2}^{b}$: a binary segmentation network for spleen only based on P_2.
- $\Psi_{P_3}^{b}$: a binary segmentation network for pancreas only based on P_3.
- $\Psi_{P_4}^{t}$: a ternary segmentation network for left kidney and right kidney only based on P_4.
- $\Psi_{F+P_1}^{b}$: a binary segmentation network for liver only based on F and P_1. Note that the spleen, pancreas, left kidney and right kidney labels in F are merged into background.
- $\Psi_{F+P_2}^{b}$: a binary segmentation network for spleen only based on F and P_2. Note that the liver, pancreas, left kidney and right kidney labels in F are merged into background.
- $\Psi_{F+P_3}^{b}$: a binary segmentation network for pancreas only based on F and P_3. Note that the liver, spleen, left kidney and right kidney labels in F are merged into background.
- $\Psi_{F+P_4}^{t}$: a ternary segmentation network for left kidney and right kidney only based on F and P_4. Note that the liver, spleen, pancreas labels in F are merged into background.
- Ψ_{All}^{mc}: a multi-class segmentation network based on F, P_1, P_2, P_3 and P_4.

4.3. Implementation details on partially segmentation labels

When training partially labeled dataset, the remaining classes are merged as background. Taking a sample only contains P_1 labels as an example, the current background region contains unlabeled spleen, pancreas, left or right kidney and the true background. In Fig. 2 of the marginal loss,the missing-labeled organs and back-

Table 3
The Dice coefficients obtained by deep segmentation networks under different loss combinations and on different datasets.

Ψ_F^{mc}: Multiclass (F)											
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
rCE	.945 ± 0.013	.819 ± 0.027	.855 ± 0.019	.917 ± 0.005	.768 ± 0.080	.679 ± 0.043	.873 ± 0.007	.866 ± 0.008	.865 ± 0.018	.873 ± 0.019	0.846
rDC	.945 ± 0.014	.837 ± 0.031	.857 ± 0.018	.914 ± 0.007	.768 ± 0.012	.673 ± 0.047	.720 ± 0.005	.821 ± 0.007	.812 ± 0.016	.917 ± 0.012	0.826
rCE+rDC	.960 ± 0.004	.850 ± 0.022	.859 ± 0.022	.918 ± 0.005	.802 ± 0.007	.695 ± 0.042	.929 ± 0.013	.939 ± 0.012	.889 ± 0.010	.903 ± 0.008	0.874
	$\Psi_{P_1}^b$		$\Psi_{P_2}^b$		$\Psi_{P_3}^b$		$\Psi_{P_4}^t$				
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
rCE	.917 ± 0.013	.872 ± 0.007	.768 ± 0.030	.938 ± 0.014	.673 ± 0.020	.720 ± 0.041	.821 ± 0.008	.812 ± 0.014	.917 ± 0.006	.913 ± 0.018	0.835
rDC	.931 ± 0.027	.883 ± 0.008	.817 ± 0.027	.940 ± 0.015	.670 ± 0.019	.715 ± 0.041	.817 ± 0.009	.807 ± 0.012	.908 ± 0.007	.900 ± 0.017	0.839
rCE+rDC	.938 ± 0.027	.904 ± 0.007	.830 ± 0.025	.954 ± 0.011	.687 ± 0.020	.728 ± 0.042	.815 ± 0.013	.813 ± 0.013	.924 ± 0.005	.917 ± 0.012	0.851
	$\Psi_{F+P_1}^b$		$\Psi_{F+P_2}^b$		$\Psi_{F+P_3}^b$		$\Psi_{F+P_4}^t$				
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
rCE	.950 ± 0.009	.875 ± 0.010	.817 ± 0.019	.943 ± 0.011	.789 ± 0.012	.734 ± 0.005	.883 ± 0.006	.917 ± 0.004	.937 ± 0.011	.920 ± 0.013	0.877
rDC	.950 ± 0.006	.890 ± 0.011	.863 ± 0.014	.941 ± 0.011	.778 ± 0.009	.700 ± 0.005	.867 ± 0.005	.933 ± 0.012	.925 ± 0.015	.938 ± 0.014	0.879
rCE+rDC	.960 ± 0.012	.899 ± 0.008	.869 ± 0.014	.945 ± 0.011	.823 ± 0.007	.753 ± 0.006	.917 ± 0.005	.940 ± 0.012	.947 ± 0.007	.950 ± 0.011	0.900
Ψ_{All}^{mc}: Multiclass ($F + P_1 + P_2 + P_3 + P_4$)											
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
mCE	.920 ± 0.013	.877 ± 0.018	.857 ± 0.018	.941 ± 0.013	.772 ± 0.007	.748 ± 0.038	.900 ± 0.006	.867 ± 0.007	.918 ± 0.009	.925 ± 0.017	0.873
mDC	.949 ± 0.008	.901 ± 0.013	.860 ± 0.011	.948 ± 0.009	.778 ± 0.006	.725 ± 0.050	.878 ± 0.007	.869 ± 0.007	.923 ± 0.010	.925 ± 0.012	0.876
mCE+mDC	.965 ± 0.012	.954 ± 0.012	.891 ± 0.015	.966 ± 0.010	.807 ± 0.007	.791 ± 0.057	.942 ± 0.012	.948 ± 0.013	.974 ± 0.012	**.974** ± 0.019	.921
mCE+mDC +eCE+eDC	**.969** ± 0.012	**.957** ± 0.009	**.924** ± 0.009	**.970** ± 0.008	**.836** ± 0.006	**.808** ± 0.041	**.946** ± 0.012	**.952** ± 0.013	**.978** ± 0.013	.972 ± 0.004	**.931**

ground represent C_3 and C_4 in Fig. 2,which are merged as the new background C_2'. The liver given ground truth labels P_1 represents C_1'.

When forming the exclusion subset, as shown in Fig. 2, the area where the liver is located is represented as C_1 and the other organ labels(spleen, pancreas, left or right kidney) are formed as the exclusion subset E_{C_1} for the liver.

4.4. Training procedure

For training the above networks except Ψ_{All}^{mc}, we use the regular CE loss, regular Dice loss, and their combination. For training the network Ψ_{All}^{mc}, when involves partial labels we need to invoke the marginal CE loss, marginal Dice loss, and their combination. Further, for Ψ_{All}^{mc} we experiment the use of exclusion Dice loss and exclusion CE loss.

Considering the impact of the varying axial resolutions of different data sets in the original CT image on the training process, we resample the 3D CT image to $(1.5 \times 1.5 \times 3)mm^3$ and then extract the patch with the shape [190,190,48] as input to illustrate the merit of our loss functions. For comparison, we use the same parameter settings in all networks; therefore there is no inference time difference among them.During training, we use 250 batches per epoch and 2 patches per batch. In order to ensure the stability of model training, We set at least one patch to contain foreground voxels in each batch during training. The initial learning rate of the network is 1e-1. Whenever the loss reduction is less than 1e-3 in consecutive 10 epochs, the learning rate decays by 20%.

We train 3D nnUNet (Isensee et al., 2018) for all segmentation networks. We choose the 3D nnUNet because it is known to be a state-of-the-art segmentation network. While there are other network architectures (Fang and Yan, 2020) that might achieve comparable performance, we expect similar empirical observations from our ablation studies even based on the other networks.

For the network Ψ_{All}^{mc}, we train it in two stages in order to prevent the instability caused by large loss value at the beginning of the training. In the first stage, we only use the fully annotated dataset F. The goal is to minimize the regular loss function using the Adam optimizer. The purpose of the first phase is to give the network an initial weight on multi-class segmentation in order to prevent the large loss value when applying the marginal loss functions. In the second stage, each epoch is trained jointly using the union of five datasets. In each epoch, we randomly select 500 patches from each training dataset with a batch size of 2. Depending on the source of the slice, we use either the regular loss, if from F, or the marginal loss and the exclusion loss, if from $P_i (i \in \{1, 2, 3, 4\})$. In actual experiment, the first stage consists of 60 epochs and the second stage 140 epochs. Online evaluation is done during training and samples are randomly selected from the testing set for evaluation. The checkpoint with the best online evaluation result is selected as the final model.

4.5. Ablation studies

We use two standard metrics for gauging the performance of a segmentation method: Dice coefficient and Hausdorff distance (HD). A higher Dice coefficient or a lower HD means a better segmentation result. Table 3 shows the mean and standard deviation of Dice coefficients of the results obtained by the deep segmentation networks under different loss combinations and with different dataset usages, from which we make the following observations.

The effect of pooling together more data. The experimental results obtained by the models jointly trained from combinations of the datasets F and $P_i (i \in \{1, 2, 3, 4\})$ are generally better than those by the models trained from a single labeled dataset alone. As shown in Table 3 and Table 4, when comparing the performance of $\Psi_{F+P_i}^b$ vs $\Psi_{P_i}^b (i \in \{1, 2, 3, 4\})$, the former generally outperforms the latter. For example, when using rCE+rDC as the loss, the mean Dice coefficient is boosted from 0.851 to 0.900 (the according HD is reduced by 37.5%). When comparing the performance of $\Psi_{F+P_i}^{mc} (i \in \{1, 2, 3, 4\})$ vs Ψ_F^{mc}, again the former is better than the latter, the mean dice coefficient is increased from 0.874 to 0.900 (the according HD is reduced by 28.7%).

The importance of CE and Dice losses. When comparing the importance of CE and Dice losses, in general, it is inconclusive which one is better, depending on the setup. For example, the Dice loss works better on liver segmentation while the CE loss signifi-

Table 4
The Hausdorff distances obtained by deep segmentation networks under different loss combinations and on different datasets.

Ψ_F^{mc}: Multiclass (F)											
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
rCE	**2.14** $\pm$ 1.69	23.31 $\pm$ 7.25	16.76 $\pm$ 7.00	8.81 $\pm$ 6.12	3.68 $\pm$ 1.46	23.15 $\pm$ 3.92	2.31 $\pm$ 0.59	3.63 $\pm$ 0.35	9.12 $\pm$ 11.58	15.32 $\pm$ 20.84	10.82
rDC	2.44 $\pm$ 2.19	23.61 $\pm$ 4.96	19.32 $\pm$ 8.86	8.71 $\pm$ 6.64	3.67 $\pm$ 2.04	23.75 $\pm$ 4.31	2.14 $\pm$ 0.30	3.65 $\pm$ 0.20	8.76 $\pm$ 7.26	7.37 $\pm$ 7.55	10.34
rCE+rDC	3.21 $\pm$ 1.72	17.36 $\pm$ 3.64	**16.11** $\pm$ 6.98	8.71 $\pm$ 6.40	6.31 $\pm$ 1.29	21.37 $\pm$ 4.75	2.17 $\pm$ 0.14	3.31 $\pm$ 0.07	8.50 $\pm$ 6.88	6.25 $\pm$ 6.81	9.33
	$\Psi_{P_1}^b$		$\Psi_{P_2}^b$		$\Psi_{P_3}^b$		$\Psi_{P_4}^t$				
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
rCE	17.32 $\pm$ 3.90	6.31 $\pm$ 3.94	28.32 $\pm$ 9.56	3.76 $\pm$ 0.35	19.36 $\pm$ 3.40	6.55 $\pm$ 4.38	15.38 $\pm$ 5.25	16.47 $\pm$ 5.41	5.07 $\pm$ 7.62	6.32 $\pm$ 21.42	12.49
rDC	12.85 $\pm$ 4.37	7.04 $\pm$ 3.44	22.15 $\pm$ 7.00	1.59 $\pm$ 0.47	17.55 $\pm$ 4.54	6.98 $\pm$ 3.05	23.65 $\pm$ 3.52	19.13 $\pm$ 5.26	6.14 $\pm$ 0.31	6.70 $\pm$ 0.37	12.38
rCE+rDC	18.76 $\pm$ 3.42	4.00 $\pm$ 3.06	25.67 $\pm$ 7.31	1.13 $\pm$ 0.20	18.36 $\pm$ 4.17	5.46 $\pm$ 3.79	13.66 $\pm$ 5.37	17.33 $\pm$ 7.02	**1.02** $\pm$ 0.20	1.89 $\pm$ 0.22	10.73
	$\Psi_{F+P_1}^b$		$\Psi_{F+P_2}^b$		$\Psi_{F+P_3}^b$		$\Psi_{F+P_4}^t$				
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
rCE	6.25 $\pm$ 1.69	8.22 $\pm$ 3.29	30.19 $\pm$ 7.60	2.17 $\pm$ 0.38	13.72 $\pm$ 1.37	9.21 $\pm$ 3.46	7.13 $\pm$ 5.52	8.23 $\pm$ 0.93	7.13 $\pm$ 7.92	6.33 $\pm$ 20.87	9.86
rDC	6.49 $\pm$ 1.14	11.25 $\pm$ 3.50	16.61 $\pm$ 7.27	2.24 $\pm$ 0.20	15.17 $\pm$ 1.17	21.34 $\pm$ 4.42	3.21 $\pm$ 0.34	6.12 $\pm$ 0.63	6.23 $\pm$ 1.14	7.21 $\pm$ 0.63	9.59
rCE+rDC	2.63 $\pm$ 0.94	7.49 $\pm$ 3.05	16.85 $\pm$ 7.27	1.65 $\pm$ 0.17	8.16 $\pm$ 0.89	8.56 $\pm$ 3.64	3.46 $\pm$ 0.30	10.70 $\pm$ 2.06	2.24 $\pm$ 0.34	4.66 $\pm$ 0.97	6.64
Ψ_{All}^{mc}: Multiclass ($F + P_1 + P_2 + P_3 + P_4$)											
Loss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
mCE	8.32 $\pm$ 3.86	15.16 $\pm$ 4.88	17.84 $\pm$ 7.12	2.24 $\pm$ 0.58	12.17 $\pm$ 0.81	8.19 $\pm$ 3.59	4.97 $\pm$ 0.73	15.55 $\pm$ 5.19	6.18 $\pm$ 7.52	7.52 $\pm$ 7.27	9.81
mDC	3.72 $\pm$ 3.42	12.71 $\pm$ 3.46	23.62 $\pm$ 6.92	2.44 $\pm$ 0.10	12.36 $\pm$ 0.91	7.18 $\pm$ 3.98	8.19 $\pm$ 0.54	8.85 $\pm$ 6.02	9.16 $\pm$ 7.18	6.55 $\pm$ 7.69	9.48
mCE+mDC	2.71 $\pm$ 1.16	**2.94** $\pm$ 2.90	21.67 $\pm$ 7.56	1.05 $\pm$ 0.09	4.49 $\pm$ 0.93	4.92 $\pm$ 3.48	1.68 $\pm$ 0.29	1.52 $\pm$ 0.18	1.77 $\pm$ 0.74	**1.58** $\pm$ 0.34	4.43
mCE+mDC +eCE+eDC	2.84 $\pm$ 1.53	4.04 $\pm$ 2.64	17.58 $\pm$ 7.27	**1.00** $\pm$ 0.09	**3.24** $\pm$ 0.69	**3.96** $\pm$ 3.27	**1.43** $\pm$ 0.14	**1.28** $\pm$ 0.07	3.13 $\pm$ 0.58	1.68 $\pm$ 0.68	**4.02**

cantly outperforms the Dice loss on left kidney segmentation. Also fusing CE and Dice losses is in general beneficial in terms of our results as it usually brings a gain in segmentation performance. For example, when using Ψ_F^b, the average dice loss reaches 0.874 for rCE+rDC, while that for rCE and rDC is 0.846 and 0.826, respectively.

The combined effect of data pooling and using marginal loss. It is evident that the segmentation network Ψ_{All}^{mc} exhibits a significant performance gain, enabled by joint training on the five datasets. It brings a 4.7% increases (.921 vs 0.874) in average dice coefficient for test images when compared with Ψ_F^{mc}, which is trained on F alone when using the dice loss and CE. Specifically, it brings an average 5.45% improvement on liver segmentation (.965 vs 0.960 on F test images and 0.954 vs 0.850 on P_1 test images), an average 4.0% improvement on spleen segmentation (.891 vs 0.859 on F test images and 0.966 vs 0.918 on P_2 test images), an average 5.05% improvement on pancreas segmentation (.807 vs 0.802 on F test images and 0.791 vs 0.695 on P_3 test images), and an average 4.45% improvement on kidney segmentation (.945 vs 0.934 on F test images and 0.974 vs 0.896 on P_4 test images).

It is interesting to notice that adding only a partially labeled dataset on the binary segmentation task does not always improve performance. For example, training a binary segmentation network (rCE+rDC loss) using F+P1 results in better accuracy for the livers. Such improvement only happens for the test images from F but not for the test images from P1. The same thing happens to F+P2 for the spleen. This may due to the feature differences between different datasets, and adding an extra dataset for training may not always improve the predictions on the original dataset. However, it is also worth mentioning that using Ψ_{All}^{mc} for training can improve the performance on all the four partially labeled datasets, this can be attributed to the benefit of multi-task learning and the exclusion loss.

The effect of exclusion loss. In addition, the exclusion loss brings significant performance boosting. The final results have been effectively improved by an average of 1.0% increases of Dice coefficient compared to the results obtained without the exclusion loss. This confirms that our proposed exclusion loss can promote the proper learning of the mutual exclusion between two labels. But it should be noted that exclusion loss is more like an auxiliary loss for partial label learning.

In sum, with the help of our newly proposed marginal loss and exclusion loss which enable the joint training of both fully labelled and partially labelled dataset, it brings a 3.1% increase (.931 vs 0.900) in dice coefficient. Such a performance improvement is essentially **a free boost** because these datasets are existent and already labeled.

Hausdorff distance. Table 4 shows the mean Hausdorff distance of the testing results, from which similar observations are made. Notably, jointly training from the five datasets, enabled by the marginal loss, can effectively increase the performances, especially it reduces the average distance from 9.33 to 4.43 (a 52.5% reduction) when using the Dice loss. Adding exclusion dice can further improve the performances (4.43 to 4.02, another 9.3% reduction). The main reason for the big HD values for say spleen $\in F$ is that sometime a small part of predicted spleen segmentation appears in non-spleen region. This does not affect the Dice coefficient but creates an outlier HD value.

The impact of loss weight. In order to further explore the impact of marginal loss and exclusion loss on the performance, we set up the training of a series of models to understand the influence of the weight ratio of marginal and exclusion losses. All the models are trained on the union of F and all the partially-annotated datasets. We experiment with ten different weight ratios: 4:1, 3:1, 2:1, 1:1, 1:2, 1:3, 1:4, 1:0, and 0:1. The dice coefficients and Hausdorff distances are reported in Table 5. Results demonstrate that a weight ratio of 1:2 achieves the best results on almost all the metrics. It is interesting to observe that, when only using exclusion loss (experiment with a weight of 0:1), there is nearly no performance improvement on pancreas and kidney comparing with Ψ_F^{mc}, which uses only F for training (as in Tables 3 and

Table 5
The Dice coefficients and Hausdorff distances obtained by the segmentation network Ψ_{All}^{mc} using different loss weight combinations.

mLoss:eLoss	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
4:1	.962 ± 0.011	.931 ± 0.029	.884 ± 0.019	.954 ± 0.013	.775 ± 0.007	.795 ± 0.058	.935 ± 0.013	.939 ± 0.012	.960 ± 0.014	.962 ± 0.013	0.910
3:1	.964 ± 0.007	.952 ± 0.017	.890 ± 0.015	.968 ± 0.010	.792 ± 0.008	.789 ± 0.055	.936 ± 0.007	.938 ± 0.013	.967 ± 0.013	.964 ± 0.008	0.916
2:1	**.970** ± 0.004	**.957** ± 0.009	.894 ± 0.018	**.970** ± 0.007	.833 ± 0.005	**.808** ± 0.038	.934 ± 0.010	.948 ± 0.014	.974 ± 0.013	.969 ± 0.013	0.926
1:1	.965 ± 0.006	.954 ± 0.015	.893 ± 0.016	.966 ± 0.009	**.844** ± 0.018	.792 ± 0.059	**.953** ± 0.009	**.959** ± 0.004	.977 ± 0.020	**.972** ± 0.007	.928
1:2	.969 ± 0.012	**.957** ± 0.009	**.924** ± 0.009	**.970** ± 0.008	.836 ± 0.006	**.808** ± 0.041	.946 ± 0.012	.952 ± 0.013	**.978** ± 0.013	**.972** ± 0.004	**.931**
1:3	.968 ± 0.009	.954 ± 0.013	.910 ± 0.017	.966 ± 0.008	.783 ± 0.011	.790 ± 0.056	.945 ± 0.011	.950 ± 0.012	.970 ± 0.014	.965 ± 0.015	0.920
1:4	.966 ± 0.008	.953 ± 0.016	.887 ± 0.016	.965 ± 0.010	.767 ± 0.022	.782 ± 0.059	.944 ± 0.011	.949 ± 0.016	.954 ± 0.014	.957 ± 0.005	0.913
1:0	.965 ± 0.012	.954 ± 0.012	.891 ± 0.015	.966 ± 0.010	.807 ± 0.007	.791 ± 0.057	.942 ± 0.012	.948 ± 0.013	.974 ± 0.012	.974 ± 0.019	0.921
0:1	.967 ± 0.012	.930 ± 0.035	.904 ± 0.020	.958 ± 0.011	.785 ± 0.015	.678 ± 0.057	.926 ± 0.008	.934 ± 0.006	.950 ± 0.019	.941 ± 0.018	0.897
4:1	2.89 ± 0.69	4.39 ± 1.92	21.43 ± 7.82	1.41 ± 0.40	6.76 ± 2.10	8.42 ± 3.90	1.85 ± 0.10	2.01 ± 0.25	8.12 ± 8.32	4.39 ± 2.40	6.17
3:1	2.51 ± 0.40	4.17 ± 4.42	19.47 ± 7.79	**1.00** ± 0.00	5.92 ± 2.29	5.11 ± 3.50	1.90 ± 0.09	2.01 ± 0.25	4.18 ± 1.95	3.75 ± 0.43	5.00
2:1	1.81 ± 0.20	4.05 ± 4.91	22.89 ± 7.86	**1.00** ± 0.00	3.44 ± 0.60	**3.96** ± 3.18	2.81 ± 1.80	1.50 ± 0.12	**1.25** ± 0.60	1.60 ± 0.95	4.43
1:1	1.98 ± 0.21	**2.93** ± 3.09	21.63 ± 8.35	1.05 ± 0.09	8.72 ± 3.87	5.17 ± 3.34	1.58 ± 0.17	8.04 ± 5.61	1.79 ± 1.92	1.66 ± 0.30	5.46
1:2	2.83 ± 1.53	4.04 ± 2.64	17.58 ± 7.27	**1.00** ± 0.09	**3.24** ± 0.69	**3.96** ± 3.27	1.43 ± 0.14	1.28 ± 0.07	3.13 ± 0.08	1.68 ± 0.68	**4.02**
1:3	**1.41** ± 0.41	3.03 ± 2.82	21.50 ± 9.73	**1.00** ± 0.00	8.02 ± 3.34	5.28 ± 3.46	**1.41** ± 0.14	**1.00** ± 0.13	6.76 ± 0.62	3.13 ± 0.79	5.25
1:4	2.19 ± 0.51	3.14 ± 3.17	21.88 ± 7.94	1.05 ± 0.09	8.42 ± 3.82	5.38 ± 3.44	12.18 ± 8.95	1.43 ± 0.14	8.76 ± 0.62	4.14 ± 0.79	6.86
1:0	2.71 ± 1.16	2.94 ± 2.90	21.67 ± 7.56	1.05 ± 0.09	4.49 ± 0.93	4.92 ± 3.48	1.68 ± 0.29	1.52 ± 0.18	1.77 ± 0.74	**1.58** ± 0.34	4.43
0:1	2.86 ± 1.56	6.08 ± 7.66	**12.95** ± 9.59	1.21 ± 0.05	5.25 ± 0.70	8.58 ± 4.34	2.77 ± 1.55	8.52 ± 3.75	12.79 ± 16.34	8.77 ± 2.21	6.98

Table 6
Data sensitivity: 5 sets of experiments with different number of fully labeled and single labeled data.

full : partial	Total # of annotated organs	Liver	Spleen	Pancreas	L Kidney	R Kidney	All
24/00	120	**.960** ± 0.004	**.859** ± 0.022	**.802** ± 0.007	**.929** ± 0.013	**.939** ± 0.012	**.874**
19/05	100	.938 ± 0.012	.852 ± 0.017	.784 ± 0.058	.879 ± 0.015	.843 ± 0.015	.859
14/10	80	.930 ± 0.013	.843 ± 0.020	.602 ± 0.045	.876 ± 0.015	.840 ± 0.009	0.818
09/15	60	.902 ± 0.017	.812 ± 0.021	.605 ± 0.047	.851 ± 0.013	.821 ± 0.004	0.798
04/20	40	.888 ± 0.014	.732 ± 0.017	.595 ± 0.048	.851 ± 0.013	.803 ± 0.005	0.774
24/00	120	**3.21** ± 1.72	**16.11** ± 6.98	**6.31** ± 1.29	**2.17** ± 0.14	**3.31** ± 0.07	**9.33**
19/05	100	8.35 ± 0.62	24.58 ± 7.53	8.72 ± 0.94	5.72 ± 0.53	12.66 ± 5.03	12.00
14/10	80	8.75 ± 0.69	26.14 ± 7.64	23.75 ± 3.36	8.15 ± 0.92	11.75 ± 6.43	15.71
09/15	60	9.01 ± 1.18	21.18 ± 7.99	21.97 ± 3.93	7.32 ± 0.29	12.39 ± 7.62	14.37
04/20	40	8.99 ± 1.17	27.25 ± 6.78	23.76 ± 3.68	7.32 ± 0.94	13.75 ± 4.71	16.21

Table 7
Segmentation performance comparison in terms of Dice coefficients and Hausdorff distances between our proposed method and state-of-the-art methods.

Methods	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
PaNN(Zhou et al., 2019)	**.972** ± 0.010	.950 ± 0.006	.915 ± 0.008	.968 ± 0.005	.780 ± 0.011	.754 ± 0.036	.941 ± 0.006	.943 ± 0.004	.937 ± 0.013	.942 ± 0.005	0.906
PIPO(Fang and Yan, 2020)	.931 ± 0.004	.949 ± 0.013	.893 ± 0.007	.945 ± 0.004	.776 ± 0.008	.767 ± 0.042	.937 ± 0.015	.943 ± 0.015	.959 ± 0.004	.965 ± 0.013	.907
our work Ψ_{All}^{mc}	.969 ± 0.012	**.957** ± 0.009	**.924** ± 0.009	**.970** ± 0.008	**.836** ± 0.006	**.808** ± 0.041	**.946** ± 0.012	**.952** ± 0.013	**.978** ± 0.013	**.972** ± 0.004	**.931**
PaNN(Zhou et al., 2019)	**1.90** ± 0.95	4.07 ± 2.84	21.37 ± 5.96	1.05 ± 0.09	8.64 ± 1.11	5.44 ± 2.54	3.31 ± 0.58	1.30 ± 0.07	4.20 ± 0.80	1.55 ± 0.14	5.28
PIPO(Fang and Yan, 2020)	6.40 ± 0.79	13.87 ± 6.36	20.66 ± 6.12	2.41 ± 0.35	6.18 ± 1.04	5.98 ± 3.62	2.32 ± 0.33	1.31 ± 0.08	6.79 ± 1.53	**1.02** ± 0.05	6.69
our work Ψ_{All}^{mc}	2.84 ± 1.53	**4.04** ± 2.64	**17.58** ± 7.27	**1.00** ± 0.09	**3.24** ± 0.69	**3.96** ± 3.27	**1.43** ± 0.14	**1.28** ± 0.07	**3.13** ± 0.58	1.68 ± 0.68	**4.02**
p-value for PaNN vs our work Ψ_{All}^{mc}	5.35$E-01$	1.48$E-02$	6.70$E-02$	1.41$E-04$	3.53$E-02$	5.34$E-03$	5.61$E-01$	9.09$E-02$	9.86$E-05$	4.37$E-04$	
p-value for PIPO vs our work Ψ_{All}^{mc}	1.73$E-04$	5.64$E-01$	6.10$E-01$	6.82$E-02$	7.98$E-02$	1.02$E-01$	6.75$E-03$	7.04$E-02$	4.53$E-01$	5.71$E-01$	

Table 8
The impact of missing part of the partial label dataset on the training result.

Missing Dataset	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
P_1	.960 ± 0.024*	.907 ± .184*	.947 ± 0.037	.964 ± .005	.797 ± 0.056	.807 ± .110	.948 ± 0.006	.951 ± 0.010	.978 ± .014	.973 ± .022	.922
P_2	.968 ± 0.006	.949 ± .024	.912 ± 0.085*	.938 ± .006*	.834 ± 0.058	.798 ± 0.172	.947 ± .008*	.950 ± .007*	.978 ± 0.013	.972 ± .012	.916
P_3	.972 ± .036	.933 ± .051	.925 ± 0.051	.955 ± .010	.812 ± 0.065*	.726 ± .148*	.956 ± .008	.959 ± .010	.980 ± .013	.972 ± .025	.919
P_4	.971 ± .002	.952 ± .037	.928 ± 0.082	.964 ± .005	.832 ± .043	.813 ± .103	.954 ± .008	.959 ± .008	.882 ± .160*	.839 ± .215*	0.909
None	.969 ± 0.012	.957 ± 0.009	.924 ± 0.009	.970 ± 0.008	.836 ± 0.006	.808 ± 0.041	.946 ± 0.012	.952 ± 0.013	.978 ± 0.013	.972 ± 0.004	.931
P_1	11.63 ± 22.14*	7.49 ± 13.33	1.34 ± 0.44	1.05 ± 0.14	6.80 ± 4.56*	4.08 ± 3.67	1.38 ± 0.30	1.30 ± 0.36	2.41 ± 1.11	2.28 ± 0.92	3.98
P_2	2.65 ± 2.35	5.26 ± 6.47	12.69 ± 25.51	12.25 ± 6.27*	4.12 ± 1.53	8.13 ± 7.34*	1.37 ± 0.34	1.50 ± 0.31*	3.37 ± 0.45	2.56 ± 0.78	4.29
P_3	3.33 ± 2.53	8.04 ± 13.60*	21.17 ± 29.52*	1.36 ± 0.38	4.50 ± 1.81	7.20 ± 6.59	1.37 ± 0.34	1.43 ± 0.36	2.26 ± 0.72	3.14 ± 4.82	5.21
P_4	1.86 ± 0.46	3.15 ± 3.06	13.22 ± 26.14	1.05 ± 0.14	3.79 ± 1.81	3.95 ± 3.17	1.47 ± 0.41*	1.44 ± 0.29	20.53 ± 33.54*	8.70 ± 16.39*	5.90
None	2.84 ± 1.53	4.04 ± 2.64	17.58 ± 7.27	1.00 ± 0.09	3.24 ± 0.69	3.96 ± 3.27	1.43 ± 0.14	1.28 ± 0.07	3.13 ± 0.58	1.68 ± 0.68	4.02

Table 9
Split of experimental dataset on pathological organs.

Dataset part	Task07 Pancreas					
	PancreasF		*PancreasP*$_1$		*PancreasP*$_2$	
train/test	train	test	train	test	train	test
#	81	20	72	18	72	18

Table 10
The impact on tissue and tumor segmentation on pathological organs.

	Dice Coefficients					Hausdorff Distances				
Data Part	Normal Tissues$\in$ *Pancreas*$_F$	Tumor$\in$ *Pancreas*$_F$	Pancreas$\in$ *Pancreas*$_{P_1}$	Tumor$\in$ *Pancreas*$_{P_2}$	*Mean*	Normal Tissues$\in$ *Pancreas*$_F$	Tumor$\in$ *Pancreas*$_F$	Pancreas$\in$ *Pancreas*$_{P_1}$	Tumor$\in$ *Pancreas*$_{P_2}$	*Mean*
rCE+rDC	.767 ± .004	.415 ± .120	.779 ± .026	.305 ± .092	.567	6.74 ± 10.73	25.26 ± 41.25	12.67 ± 22.90	16.21 ± 24.79	15.24
our work	.769 ± .004	.524 ± .117	.770 ± .028	.391 ± .117	.614	7.62 ± .7.78	19.65 ± 35.95	13.27 ± 24.84	19.56 ± 30.47	15.03

Table 11
The Dice coefficients and Hausdorff distances change when adjusting val/tr split in *F*.

experiment	Liver$\in F$	Liver$\in P_1$	Spleen$\in F$	Spleen$\in P_2$	Pancreas$\in F$	Pancreas$\in P_3$	L Kidney$\in F$	R Kidney$\in F$	L Kidney$\in P_4$	R Kidney$\in P_4$	All
DC before adjust val/tr split	.969 ± 0.012	.957 ± 0.009	**.924** ± 0.009	.970 ± 0.008	.836 ± 0.006	.808 ± 0.041	.946 ± 0.012	.952 ± 0.013	.978 ± 0.013	.972 ± 0.004	.931
DC after adjust val/tr split	.970 ± 0.002	.956 ± 0.009	**.953** ± 0.018	.968 ± 0.003	.838 ± 0.038	.811 ± 0.105	.954 ± 0.012	.949 ± 0.007	.974 ± 0.007	.972 ± 0.003	.935
HD before adjust val/tr split	2.84 ± 1.53	4.04 ± 2.64	**17.58** ± 7.27	1.00 ± 0.09	3.24 ± 0.69	3.96 ± 3.27	1.43 ± 0.14	1.28 ± 0.07	3.13 ± 0.58	1.68 ± 0.68	4.02
HD after adjust val/tr split	1.94 ± 0.29	2.96 ± 2.73	**8.31** ± 6.58	1.00 ± 0.03	3.72 ± 0.43	4.27 ± 4.18	1.28 ± 0.20	1.43 ± 0.36	1.92 ± 0.75	1.56 ± 0.24	2.84

4). This indicates that exclusion loss is more suitable as an auxiliary loss to be used with marginal loss together.

The effect when missing parts of partially labeled dataset. Considering in practice, some of the partially labeled dataset may be missed, we perform another group of experiments to show whether the newly proposed loss can still effectively improve the performance. We perform a series of ablation studies, in which one of the partially labeled datasets, $P_i (i \in \{1, 2, 3, 4\})$, is not used during the training, the results are shown in 8. It can be observed that when missing $P_i (i \in \{1, 2, 3\}$ respectively, the segmentation performances has a certain drop on the corresponding organ (The average Dice of liver, spleen and pancreas on F decreases by 0.015). However, it is still much better when compares to training the fully annotated dataset F alone (as shown in Table 3). However, the lack of P_4 seems to have a large negative effect on the segmentation results of the left and right kidneys, which leads to a big drop on P_4 (Average Dice 0.861 vs 0.975 of left and right kidney), which may be caused by the large difference between the datasets.

The effect on pathological organs. In order to verify the performance of our method on pathological targets, we use the dataset task07 pancreas from the Decathlon-10. it has two foreground labels: normal tissues and tumors. We divide the data set into three data set as shown in Table 9. The first data set *OrganF* contains both labels of normal tissues and tumors, the second data set *Organ*P_1 merges the normal tissue and tumor labels into one label as organ, and the third data set *Organ*P_2 only keep the tumor labels. All the other experimental settings are the same as the previous ones.

The results are shown in Table 10. We can see a certain performance improvement on both pancreas segmentation and tumor segmentation.

The effect of the number of annotations. Finally, we perform a group of tests to measure the sensitivity of performance with the number of data annotation increases. We randomly split the fully annotated dataset F into a training set with 24 samples and a testing set with six samples and leave the testing set untouched. In the five sets of experiments reported in Table 6, we alter the training set by replacing some fully labeled data with single labeled data, while keeping the total number of the training data unchanged. For example, for a '14/10' split, we have 14 fully labels images with 5 organs, and the rest of 10 images are further randomly divided into 5 single-label groups of 2 images. For the 1st group, we can use its liver annotation. Similarly we use only the spleen, pancreas, left kidney, and right kidney labels for the 2nd to the 5th groups, respectively. As a result, we have a total of 14*5+2*5=80 annotated organs. Results in Table 6 confirm that the dice coefficient consistently decreases as the amount of annotation decreases, which is as expected.

4.6. Comparison with state-of-the-art

Our model is also compared with the other partially-supervised segmentation networks. The results are shown in Table 7. The Prior-aware Neural Network (PaNN) refers to the work by Zhou et al. (Zhou et al., 2019) which adds a prior-aware loss to learn partially labeled data. The pyramid input and pyramid output (PIPO)

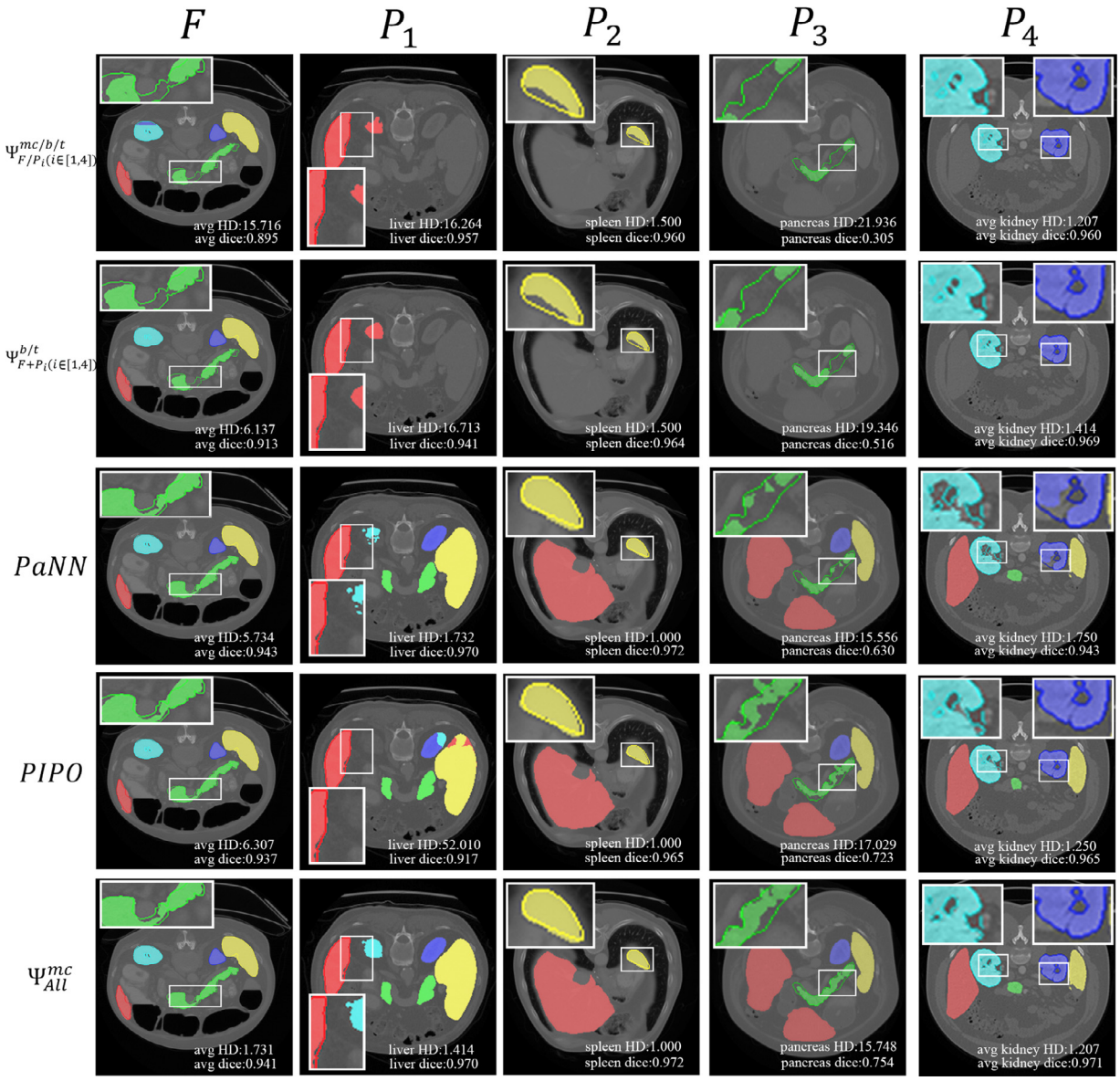

Fig. 4. The comparison of results obtained by different segmentation networks. The red area represents the liver, the yellow area represents the spleen, the green area represents the pancreas, the cyan and blue areas represent the right and left kidneys, respectively. The edge with deeper color means the ground truth given by the dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

refers to the work by Fang et al. (Fang and Yan, 2020) which develops a multi-scale structure as well as target adaptive loss to enable learning partially labeled data. Our work achieves a significantly better performance than these two methods. The average Dice reaches 0.931 for our model, while that for PaNN and PIPO is 0.906 and 0.907, respectively. Our method also greatly reduces the mean Hausdorff distance by 24.0% comparing with PaNN and 40.0% comparing with PIPO. Specifically, our method achieves slight better (except for Liver$\in F$) performance for large organs such as liver and spleen, but it brings a significant performance boost on small organs such as pancreas, left and right kidneys. Our work performs consistently better than the PIPO method on all the organs regardless the datasets, the improvement may be due to the use of 3D model as well as the exclusion loss.

Fig. 4 presents visualization of sample results of different methods. With the assistance of auxiliary datasets, the performances are significantly improved. Especially, there are situations occurring on all the other methods that the predicted organ region enters a different organ, which results a large HD value. The exclusion loss used in our method can effectively reduce such an error and greatly improve the HD performance. Besides, our method can achieve more meticulous segmentation results on some small organs such as pancreas and kidney, especially when there are small holes around the organ center.

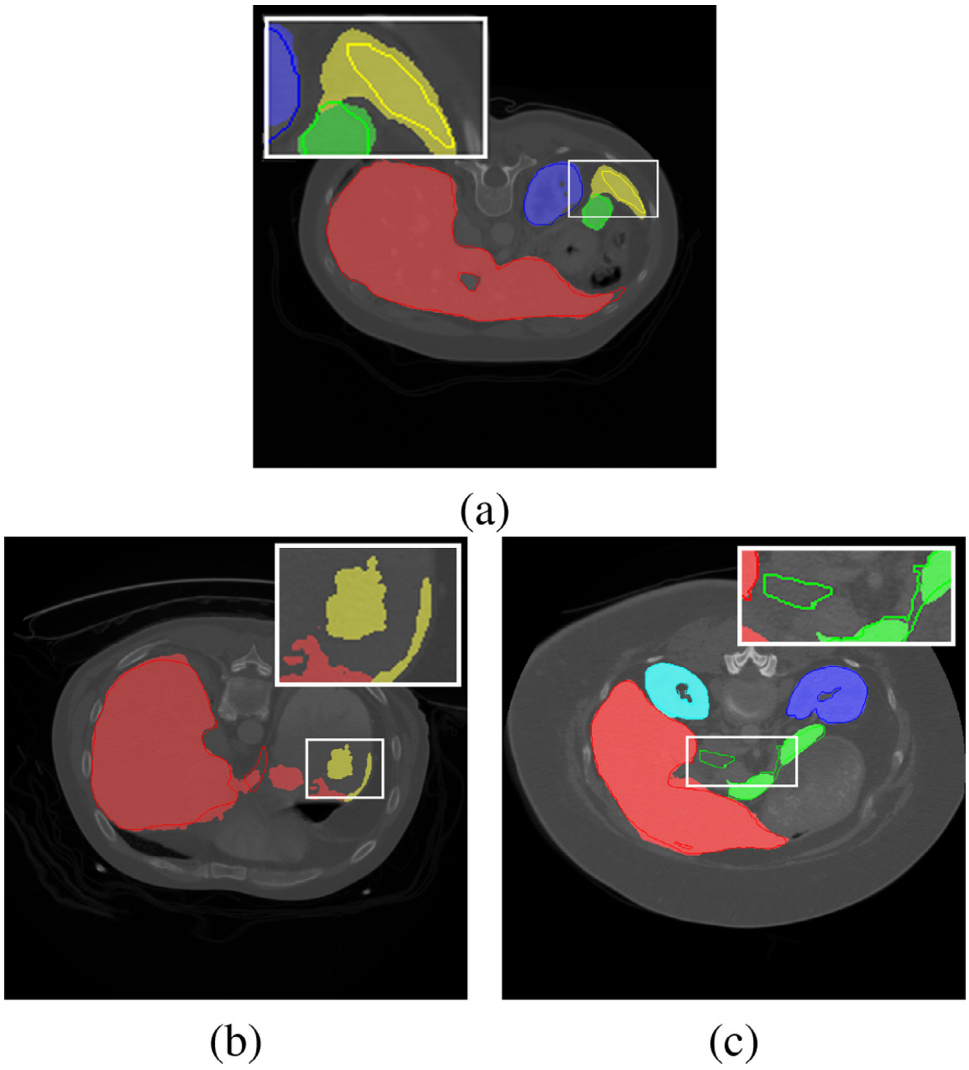

Fig. 5. Failure cases. The figure shows that there are still some regional predictions that have made big mistakes especially in spleen and pancreas.

5. Discussions and conclusions

In this paper, we propose two new types of loss function that can be used for learning a multi-class segmentation network based on multiple datasets with partial organ labels. The marginal loss enables the learning due to the presence of 'merged' labels, while the exclusion loss promotes the learning by adding the mutual exclusiveness as prior knowledge on each labeled image pixel. Our extensive experiments on five benchmark datasets clearly confirm that a significant performance boost is achieved by using marginal loss and exclusion loss. Our method also greatly outperforms existing frameworks for partially annotated data learning.

However, our proposed method is far from perfect. Fig. 5 shows three typical failure cases. In case (a),the size of spleen is smaller compares to the others, which leads a prediction with high false positive rate, the dice coefficients of this sample between ground truth and prediction is thus dropped greatly. This case also causes a drop of average dice value of spleen on the whole dataset F. To demonstrate this, we performed a new test by simply replacing this sample with a sample in the training set. As shown in Table 11, the average prediction accuracy of spleen in F is greatly improved (0.924 vs 0.953 in DC and 17.57 vs 8.31 in HD) by this simple adjustment. In case (b), the background has similar features to liver so the liver prediction on the right side is wrong. In case (c), our method also has some misjudgment on spleen and pancreas. We will generalize the current method for improved segmentation performances by incorporating more knowledge about the organs, such as using shape adversarial prior (Yang et al., 2017). Furthermore, in future we will extend the marginal loss and exclusion loss on other tasks for partially labeled annotated learning and explore the use of other loss functions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Gonglei Shi: Conceptualization, Methodology, Data curation, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Li Xiao:** Resources, Supervision, Writing - review & editing. **Yang Chen:** Writing - review & editing. **S. Kevin Zhou:** Conceptualization, Methodology, Resources, Supervision, Writing - review & editing.

Acknowledgement

This work was supported by the State's Key Project of Research and Development Plan (No. 2017YFA0104302, 2017YFC0109202 and 2017YFC0107900); the National Natural Science Foundation (No. 61871117); and the Science and Technology Program of Guangdong (No. 2018B030333001). Li Xiao thanks CCF-Tencent Open Fund for support.

References

Berman, M., Rannen Triki, A., Blaschko, M.B., 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4413–4421.

Binder, T., Tantaoui, E.M., Pati, P., Catena, R., Set-Aghayan, A., Gabrani, M., 2019. Multi-organ gland segmentation using deep learning. Front. Med. (Lausanne) 6, 173.

Cerrolaza, J.J., Reyes, M., Summers, R.M., González-Ballester, M.Á., Linguraru, M.G., 2015. Automatic multi-resolution shape modeling of multi-organ structures. Med. Image Anal. 25 (1), 11–21.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.-A., 2018. Voxresnet: deep voxelwise residual networks for brain segmentation from 3D MR images. Neuroimage 170, 446–455.

Chen, X., Udupa, J.K., Bagci, U., Zhuge, Y., Yao, J., 2012. Medical image segmentation by combining graph cuts and oriented active appearance models. IEEE Trans. Image Process. 21 (4), 2035–2046.

Chu, C., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Hayashi, Y., Nimura, Y., Rueckert, D., Mori, K., 2013. Multi-organ segmentation based on spatially-divided probabilistic atlas from 3D abdominal CT images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 165–172.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23 (6), 681–685.

Cour, T., Sapp, B., Taskar, B., 2011. Learning from partial labels. Journal of Machine Learning Research 12 (42), 1501–1536.

Dmitriev, K., Kaufman, A.E., 2019. Learning multi-class segmentations from single–class datasets. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Fang, X., Yan, P., 2020. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging doi:10.1109/TMI.2020.3001036. 1–1

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE Trans. Med. Imaging 37 (8), 1822–1834.

Ginneken, B.V., Schaefer-Prokop, C.M., Prokop, M., 2011. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261 (3), 719–732.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2. IEEE, pp. 1735–1742.

He, B., Huang, C., Jia, F., 2015. Fully automatic multi-organ segmentation based on multi-boost learning and statistical shape model search. CEUR Workshop Proc 1390, 18–21.

He, Z.-F., Yang, M., Gao, Y., Liu, H.-D., Yin, Y., 2019. Joint multi-label classification and label correlations with missing labels and feature selection. Knowl. Based Syst. 163, 145–158.

Heimann, T., et al., 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans. Med. Imaging 28 (8), 1251–1265.

Heimann, T., Meinzer, H.-P., 2009. Statistical shape models for 3D medical image segmentation: a review. Med. Image Anal. 13 (4), 543–563.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. NnU-net: self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486.

Kohlberger, T., Sofka, M., Zhang, J., Birkbeck, N., Wetzl, J., Kaftan, J., Declerck, J., Zhou, S.K., 2011. Automatic multi-organ segmentation using learning-based segmentation and level set optimization. In: Fichtinger, G., Martel, A., Peters, T. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 338–345.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2017. Multi-atlas labeling beyond the cranial vault-workshop and challenge.

Lay, N., Birkbeck, N., Zhang, J., Zhou, S.K., 2013. Rapid multi-organ segmentation using context integration and discriminative models. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (Eds.), Information Processing in Medical Imaging. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 450–462.

Li, Y., Gao, F., Ou, Z., Sun, J., 2018. Angular softmax loss for end-to-end speaker verification. In: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, pp. 190–194.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Liu, W., Wen, Y., Yu, Z., Yang, M., 2016. Large-margin softmax loss for convolutional neural networks. In: International Conference on Machine Learning, 2, p. 7.

Liu, Y., Gargesha, M., Qutaish, M., Zhou, Z., Scott, B., Yousefi, H., Lu, Z., Wilson, D.L., 2020. Deep learning based multi-organ segmentation and metastases segmentation in whole mouse body and the cryo-imaging cancer imaging and therapy analysis platform (citap). In: Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, 11317. International Society for Optics and Photonics, p. 113170V.

Lombaert, H., Zikic, D., Criminisi, A., Ayache, N., 2014. Laplacian forests: Semantic image segmentation by guided bagging. In: International Conference on Medical Image Computing & Computer-assisted Intervention.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Lu, C., Zheng, Y., Birkbeck, N., Zhang, J., Kohlberger, T., Tietjen, C., Boettger, T., Duncan, J.S., Zhou, S.K., 2012. Precise segmentation of multiple organs in ct volumes using learning-based approach and information theory. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 462–469.

Okada, T., Linguraru, M.G., Hori, M., Summers, R.M., Tomiyama, N., Sato, Y., 2015. Abdominal multi-organ segmentation from CT images using conditional shape–location and unsupervised intensity priors. Med. Image Anal. 26 (1), 1–18.

Okada, T., Linguraru, M.G., Hori, M., Suzuki, Y., Summers, R.M., Tomiyama, N., Sato, Y., 2012. Multi-organ segmentation in abdominal ct images. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 3986–3989.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 379–387.

Saxena, S., Sharma, N., Sharma, S., Singh, S., Verma, A., 2016. An automated system for atlas based multiple organ segmentation of abdominal CT images. British Journal of Mathematics & Computer Science 12, 1–14. doi:10.9734/BJMCS/2016/20812.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. FaceNet: A unified embedding for face recognition and clustering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823.

Shimizu, A., Ohno, R., Ikegami, T., Kobatake, H., Nawano, S., Smutek, D., 2007. Segmentation of multiple organs in non-contrast 3d abdominal ct images. Int. J. Comput. Assist. Radiol. Surg. 2 (3–4), 135–142.

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

Suzuki, M., Linguraru, M.G., Okada, K., 2012. Multi-organ segmentation with missing organs in abdominal CT images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 418–425.

Sykes, J., 2014. Reflections on the current status of commercial automated segmentation systems in clinical practice. J. Med. Radiat. Sci. 61 (3), 131–134.

Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., Hamarneh, G., 2019. Combo loss: handling input and output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics 75, 24–33.

Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D., 2015. Discriminative dictionary learning for abdominal multi-organ segmentation. Med. Image Anal. 23 (1), 92–104.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018. Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274.

Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L., 2019. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Med. Image Anal. 55, 88–102.

Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision. Springer, pp. 499–515.

Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2013. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Trans. Med. Imaging 32 (9), 1723–1730.

Wu, B., Lyu, S., Hu, B.-G., Ji, Q., 2015. Multi-label learning with missing labels for image annotation and facial action unit recognition. Pattern Recognit. 48 (7), 2279–2289.

Xiao, L., Zhu, C., Liu, J., Luo, C., Liu, P., Zhao, Y., 2019. Learning from suspected target: Bootstrapping performance for breast cancer detection in mammography. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham, pp. 468–476.

Xu, Z., Burke, R.P., Lee, C.P., Baucom, R.B., Poulose, B.K., Abramson, R.G., Landman, B.A., 2015. Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning. Med. Image Anal. 24 (1), 18–27.

Yang, D., Xu, D., Zhou, S.K., Georgescu, B., Chen, M., Grbic, S., Metaxas, D., Comaniciu, D., 2017. Automatic liver segmentation using an adversarial image-to-image network. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L.,

Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention MICCAI 2017. Springer International Publishing, Cham, pp. 507–515.
Yu, H.-F., Jain, P., Kar, P., Dhillon, I., 2014. Large-scale multi-label learning with missing labels. In: International conference on machine learning, pp. 593–601.
Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019. Data augmentation using learned transformations for one-shot medical image segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 8543–8553.
Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L., 2019. Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10672–10681.
Zhu, P., Xu, Q., Hu, Q., Zhang, C., Zhao, H., 2018. Multi-label feature selection with missing labels. Pattern Recognit. 74, 488–502.