Context Label Learning: Improving Background Class Representations in Semantic Segmentation

Zeju Li, *Graduate Student Member, IEEE*, Konstantinos Kamnitsas, Cheng Ouyang, Chen Chen, and Ben Glocker

Abstract—Background samples provide key contextual information for segmenting regions of interest (ROIs). However, they always cover a diverse set of structures, causing difficulties for the segmentation model to learn good decision boundaries with high sensitivity and precision. The issue concerns the highly heterogeneous nature of the background class, resulting in multi-modal distributions. Empirically, we find that neural networks trained with heterogeneous background struggle to map the corresponding contextual samples to compact clusters in feature space. As a result, the distribution over background logit activations may shift across the decision boundary, leading to systematic over-segmentation across different datasets and tasks. In this study, we propose context label learning (CoLab) to improve the context representations by decomposing the background class into several subclasses. Specifically, we train an auxiliary network as a task generator, along with the primary segmentation model, to automatically generate context labels that positively affect the ROI segmentation accuracy. Extensive experiments are conducted on several challenging segmentation tasks and datasets. The results demonstrate that CoLab can guide the segmentation model to map the logits of background samples away from the decision boundary, resulting in significantly improved segmentation accuracy. Code is available at https://github.com/ZerojumpLine/CoLab.

Zeju Li, Cheng Ouyang, and Ben Glocker are with the BioMedIA Group, Department of Computing, Imperial College London, SW7 2AZ London, U.K. (e-mail: zeju.li18@imperial.ac.uk; c.ouyang@imperial.ac.uk; b.glocker@imperial.ac.uk).

Konstantinos Kamnitsas is with the BioMedIA Group, Department of Computing, Imperial College London, SW7 2AZ London, U.K., also with the School of Computer Science, University of Birmingham, B15 2TT Birmingham, U.K., and also with the Department of Engineering Science, University of Oxford, OX3 7DQ Oxford, U.K. (e-mail: konstantinos.kamnitsas12@imperial.ac.uk).

Chen Chen was with HeartFlow Inc., Mountain View, CA 94041 USA. She is now with the BioMedIA Group, Department of Computing, Imperial College London, SW7 2AZ London, U.K. (e-mail: chen.chen15@imperial.ac.uk).

Index Terms—Underfitting, multi-task learning, self-supervised learning, image segmentation.

I. INTRODUCTION

CONVOLUTIONAL neural networks (CNNs) are state-of-the-art approach for semantic image segmentation. Their large number of trainable parameters make them capable to fit to different kinds of tasks yielding high performance in terms segmentation accuracy [37]. Yet, in real world applications, CNNs seem sometimes unable to capture and generalize from complex, heterogeneous training sets when the amount of training data is limited [28]. Specifically, in the case of medical image segmentation, samples from the background class, which provide essential contextual information for regions of interest (ROIs), make up the majority of the training set while containing diverse sets of structures with heterogeneous characteristics, making it hard for the segmentation model to learned accurate decision boundaries.

When trained with datasets with a highly heterogeneous background class, the segmentation model is prone to under-fit these contextual samples and fail to separate the ones which share characteristics similar to the ROI samples. The model then produces false positives (FP), yielding systematic over-segmentation. In this study, we argue that underfitting of the contextual information is a main cause of degraded segmentation performance by affecting precision (calculated as $\frac{\text{TP}}{\text{TP+FP}}$). We observe that better context representation where the background class is decomposed into several subclasses, for example, using additional anatomy labels, significantly improves the ROI segmentation accuracy. We show some examples of human-defined context labels in Fig. 1(a). In the case of liver tumor segmentation, for example, it is beneficial to also have labels for the liver available in addition to the tumor class. Empirically, we find that when training with human-defined context labels, the segmentation model can yield better performance in terms of Dice similarity coefficient (DSC) and 95% Hausdorff distance (HD), as shown in Fig. 1(b). However, these human-defined context labels are not always available and difficult and time-consuming to obtain. In many applications, only the ROI labels, e.g., tumor class, are available. Here, we propose context label learning (CoLab),
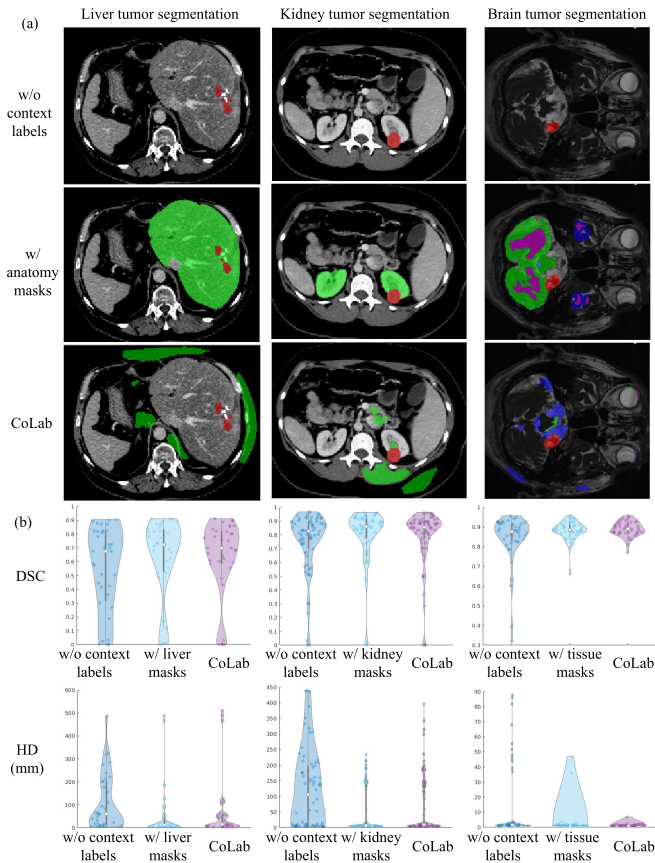
Fig. 1. (a) Examples of ROI classes (in red color) along with human-defined context labels and context labels generated automatically by the proposed context label learning (CoLab) in background. Here, human-defined context labels (e.g., liver, kidney, brain tissue) provide additional anatomical information to benefit representation learning. In a similar manner, CoLab automatically finds and labels task-related tissues/structures in background to benefit corresponding ROI segmentation, see Fig. 8 for more details. (b) Self-supervised context labels generated by CoLab can significantly improve the segmentation of different structures with improved Dice similarity coefficient (DSC) and reduced 95% Hausdorff distance (HD), compared to the one trained w/o context labels. The performance gains are comparable and sometimes even higher than the ones using human-defined context labels.

which automatically generates context labels to improve the learning of a context representation yielding better ROI segmentation accuracy. As demonstrated in Fig. 1(b), CoLab can bring similar improvements when compared with training with human-defined context labels, without the need for expert knowledge.

The contribution of this study can be summarized as follows: 1) With the observations of six datasets, we conclude that underfitting of the background class consistently degrades the segmentation performance by decreasing precision. 2) We find that better context representations with a decomposition of the background class can improve segmentation performance. 3) We propose CoLab, a flexible and generic method to automatically generate soft context labels. We validate CoLab with extensive experiments and find consistent improvements where the segmentation accuracy is en par and sometimes better compared to the case where human annotated context labels are available.

II. RELATED WORK

A. Class Imbalance

CoLab is related to the class imbalance problem as background commonly constitutes the majority class in image segmentation. However, methods to combat class imbalance mostly focus explicitly on improving the performance of the minority categories [15], [24], [36]. Most approaches ignore the characteristics of majority background class as it is not contributing much to the common evaluation metrics of sensitivity, precision, and DSC of the ROI classes. CoLab focuses specifically on the representation of the background class and is complementary to methods tackling class imbalance such as loss reweighting strategies [7].

Previous studies adopt coarse-to-fine strategies to reduce FP in segmentation tasks with class imbalance [32], [35], [41]. However, ROI samples missed in the coarse stage cannot be recovered with later stages. In contrast, CoLab is trained in an end-to-end manner and can reduce varied kinds of FP.

B. Multi-Task Learning

CoLab, which is formulated as multi-label classification, can be seen as a form of multi-task learning (MTL). Current MTL methods train the model with different predefined tasks together with the main task using a shared feature representation [6], [20], [31]. Previous works also attempted to incorporate spatial prior [11], [39] or task prior [38], [40] into model training with some predefined auxiliary tasks and optimization functions. In contrast, CoLab reformulates the main task by decomposing the background class with context labels and automatically generate the auxiliary task in a self-supervised manner. We argue that CoLab can have a direct impact on the main task by extending the label space.

The main methodology of the CoLab strategy is inspired by some recently proposed methods which aim to generate the weights for pre-defined auxiliary tasks or labels through a similar meta-learning framework [23], [27]. In this study, CoLab is specifically designed for semantic segmentation with heterogeneous background classes which is a common scenario in medical imaging.

III. CONTEXT LABELS IN IMAGE SEGMENTATION

A. Preliminaries

We consider CNNs for multi-class segmentation with a total of c classes. Given a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ with N samples, where $\boldsymbol{y}_i \in \mathbb{R}^c$ is the one-hot encoded label for the central pixel in the image sample $\boldsymbol{x}_i \in \mathbb{R}^d$, such that $\mathbf{1} \cdot \boldsymbol{y}_i = 1 \ \forall \ i$. A segmentation model $f_\phi(\cdot)$ learns class representations of the input sample $\boldsymbol{x}_i$, noted as $\boldsymbol{z}_i = f_\phi(\boldsymbol{x}_i) \in \mathbb{R}^c$. We obtain the predicted probability $\boldsymbol{p}_i$ that the real class of $\boldsymbol{x}_i$ is j via a softmax function with $p_{ij} = \mathrm{e}^{z_{ij}} / \sum_{j=1}^{c} \mathrm{e}^{z_{ij}}$. Typically the model is optimized by minimizing the empirical risk $R_{\mathcal{L}_{seg}}(f_\phi) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{seg}(f_\phi(\boldsymbol{x}_i), \boldsymbol{y}_i)$ computed on the training set. The segmentation loss $\mathcal{L}_{seg}$ can be defined as the

sum of losses $\mathcal{L}$ over c classes:

$$\mathcal{L}_{seg}(f_\phi(\boldsymbol{x}_i), \boldsymbol{y}_i) = \sum_{j=1}^{c} \mathcal{L}(p_{ij}, y_{ij})$$

$$= \underbrace{\sum_{j=1}^{c-1} \mathcal{L}(p_{ij}, y_{ij})}_{\text{ROI classes}} + \underbrace{\mathcal{L}(p_{ic}, y_{ic})}_{\text{background class}}, \quad (1)$$

where $\mathcal{L}$ is a criterion for a specific class, such as cross entropy (CE) or soft DSC [25]. Here, we further decompose $\mathcal{L}_{seg}$ into two terms, including an ROI loss (computed on $c - 1$ foreground classes) and a background loss (computed on the background class, only).

We aim to improve segmentation performance by augmenting the background class with auxiliary context classes. Specifically, we propose to utilize context labels assigned to different and decomposed background regions. In order to divide the background class into $t > 1$ classes, we create another model $\tilde{f}_\theta(\cdot)$ with $\tilde{z}_i = \tilde{f}_\theta(\boldsymbol{x}_i) \in \mathbb{R}^{c+t-1}$ and predicted probability $\tilde{\boldsymbol{p}}_i$. We also require an additional one-hot label $\tilde{\boldsymbol{y}}_i \in \mathbb{R}^{c+t-1}$, where we require $\tilde{y}_{ij} = y_{ij} \, \forall \, i, \, j \in \{1, \cdots, c-1\}$ and $\sum_{j=c}^{c+t-1} \tilde{y}_{ij} = y_{ic} \, \forall \, i$. With this notion, the segmentation loss can be written as:

$$\mathcal{L}_{seg}(\tilde{f}_\theta(\boldsymbol{x}_i), \tilde{\boldsymbol{y}}_i) = \sum_{j=1}^{c+t-1} \mathcal{L}(\tilde{p}_{ij}, \tilde{y}_{ij})$$

$$= \underbrace{\sum_{j=1}^{c-1} \mathcal{L}(\tilde{p}_{ij}, \tilde{y}_{ij})}_{\text{ROI classes}} + \underbrace{\sum_{j=c}^{c+t-1} \mathcal{L}(\tilde{p}_{ij}, \tilde{y}_{ij})}_{\text{background classes}}, \quad (2)$$

In the following method section, we always consider a simplified but common case where we only have one ROI class ($c = 2$) for simplicity. In other words, we only consider binary ROI segmentation with $f_\phi(\boldsymbol{x}_i) \in \mathbb{R}^2$, although the method can be naturally extended to multi-class ROI segmentation. With this assumption, Eq. 2 can be simplified as:

$$\mathcal{L}_{seg}(\tilde{f}_\theta(\boldsymbol{x}_i), \tilde{\boldsymbol{y}}_i) = \underbrace{\mathcal{L}(\tilde{p}_{i1}, \tilde{y}_{i1})}_{\text{ROI class}} + \underbrace{\sum_{j=2}^{t+1} \mathcal{L}(\tilde{p}_{ij}, \tilde{y}_{ij})}_{\text{background classes}} \quad (3)$$

B. False Positives on Background Class

To better understand the effect of the background class on the model learning, we train multiple CNNs on segmentation datasets which contain heterogeneous background. We conduct experiments on challenging tasks including liver tumor segmentation in computed tomography (CT) images [5], kidney tumor segmentation in CT images [12], colon tumor segmentation in CT images [34], vestibular schwannoma (VS) segmentation in T2-weighted magnetic resonance (MR) images [33], brain stroke lesion segmentation in T1-weighted MR images [21] and pancreatic tumor segmentation in CT images [34]. We adopt a well-configured 3D U-Net [13] as the segmentation model for all the experiments, which has been demonstrated to yield competitive results across different

medical image segmentation tasks. The detailed data and network configurations are summarized in Section V.

Visualizations of the segmentation results are shown in Fig. 2(a). When trained with binary segmentation tasks and without context labels, the models are prone to over-segment the ROIs with many FP. Specifically, the model trained on the binary liver tumor task predicts other organs outside liver as liver tumor; the model trained on the binary kidney tumor task predicts parts of the healthy kidney regions as kidney tumor; while the model trained on the binary brain tumor task predicts the surrounding brain tissue as brain tumor; the model trained on the binary brain lesion task predicts unrelated healthy brain regions as brain lesion.

C. Underfitting of Background Samples

To study the model behavior when trained with heterogeneous background classes, we can monitor the logit distribution of samples from ROI and background classes for the test data. Our observations for liver tumor, kidney tumor, and brain tumor segmentation are summarized in Fig. 3(a, d, g).

We find that the CNN models map the ROI samples to a compact cluster in the logit space while background samples form a more dispersed distribution. This indicates that the model cannot easily map all background samples to a single cluster representing the background class. Although the model seems to separate ROI from background samples in the feature space, it builds complex background representations and unable to capture an accurate decision boundary between ROI and immediate context. A possible reason is that the CNN uses most of its capacity to extracted the common features among background regions with different characteristics. Unfortunately, these shared features are not very discriminate. Specifically, we observe that the logit distribution of background samples overlaps with the learned decision boundary. This is the reason why the model predicts many FP leading to over-segmentation of ROI structures when the background class is heterogeneous. We hypothesise that the width of the distribution serves as an indicator of the heterogeneity of a specific class and is a sign of the difficulty during learning. As a result of sample heterogeneity, a CNN may struggle to reduce intra-class variation of the background class and underfit the background samples, failing to recognize the background samples that share similar characteristics with ROI samples.

It should be noted that we do not observe significant difference between the logit distribution of training background samples and test background samples, as also shown in previous studies [18], [19], indicating that the logit shift of background samples is indeed due to underfitting instead of overfitting.

D. The Effect of Context Labels

The availability of context labels greatly helps with the ROI segmentation, which provide additional signals to CNN to fit the training data of heterogeneous background samples. We confirm this empirically by including human-defined context labels in the model training. Specifically, we adopt
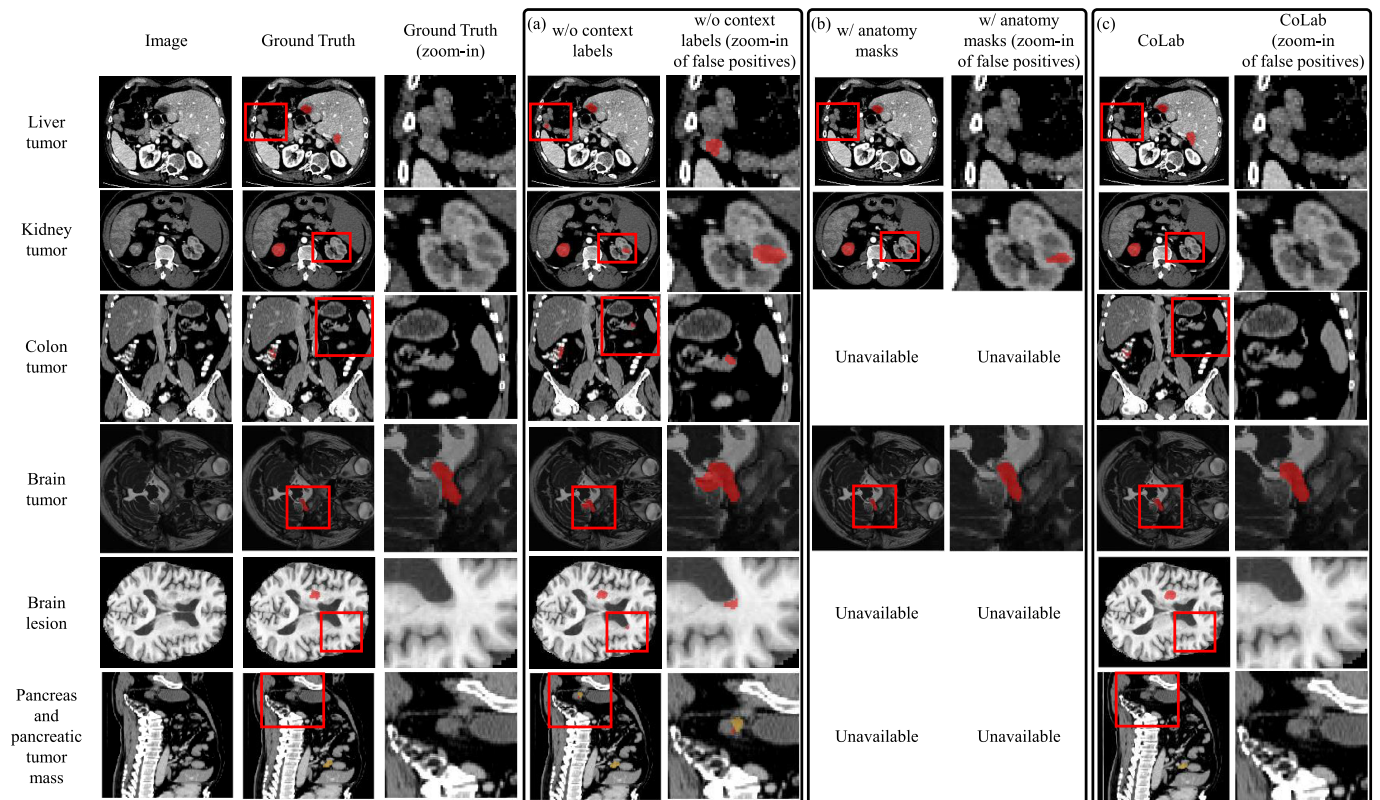
Fig. 2.   Visualization of different datasets and segmentation results on test data when training w/o and w/ different types of context labels. (a) Models trained without any context labels are prone to over-segmentation of the ROI (red/red + orange) with different kinds of FP (marked with red boxes). (b) Empirically, we observe that adding human-defined context labels such as anatomy masks for supervision helps reduce FP and therefore obtains higher overall segmentation accuracy. (c) The proposed CoLab achieves a similar benefit without requiring additional human annotation by producing context labels automatically.

liver masks ($t = 2$) for liver tumor segmentation, kidney masks ($t = 2$) for kidney tumor segmentation, brain tissue masks including ventricles, deep grey matter, cortical grey matter, white matter and other tissues ($t = 6$) for brain tumor segmentation. The kidney and liver masks are manually annotated, while the brain tissue masks are generated automatically using the paired T1-weighted MR images and a multi-atlas label propagation with expectation–maximisation based refinement (MALP-EM) [17]. In order to make a fair comparison, we make sure that all the experiments share the same training schedule except the label space. We sample the training patches only considering ROIs. Specifically, we make sure 50% training patches to contain ROIs and sample the other half of the training patches uniformly.

The observations on test and training set are summarized in Fig. 4. We find that the models yield overall better performance when trained with anatomy masks, as indicated by improved DSC (defined as $DSC = 2\frac{\text{sensitivity}\cdot\text{precision}}{\text{sensitivity}+\text{precision}}$). Furthermore, we observe that the model trained with context labels yields higher precision while preserving similar sensitivity. The observation is consistent with the visualization results in Fig. 2(b), where we find the models trained with human-defined context labels reduce FP.

Similarly, we visualize the corresponding network behaviour in Fig. 3(b, e, h). As we only consider the segmentation performance of the ROI (class 1), we visualize the logits in the plane of (z_1, max($z_2, \ldots, z_{t+1}$)). We observe that the model trained with anatomy masks maps the background samples

to a narrower distribution and reduces the background logits shift across the decision boundary. This indicates that the models fit the training data better with the help of context labels. Instead of building generic filters for all the background samples, the CNNs can dedicate specific filters to model a more homogeneous subparts of the background samples that share common characteristics. The models are faced with a simplified segmentation task with homogeneous background subclasses yielding better overall performance with the same model capacity.

Although anatomy masks are found to be effective context labels, they are not always available in real-world applications. Specifically acquiring manually annotated context labels is time-consuming and would require significant efforts from human experts to generate annotations at large scale. Therefore, we propose CoLab which can automatically discover specific soft context labels using a meta-learning strategy. CoLab benefits the segmentation model training by making it to fit the background samples better, achieving comparable or even better performance when compared with models trained on manually defined anatomy masks.

IV. CoLab

A. Overview

Now we consider CNNs for binary semantic segmentation. Typically, we are given a baseline segmentation model $f_\phi(\cdot)$ that maps the input image x_i to the label space $f_\phi(x_i) \in \mathbb{R}^2$.
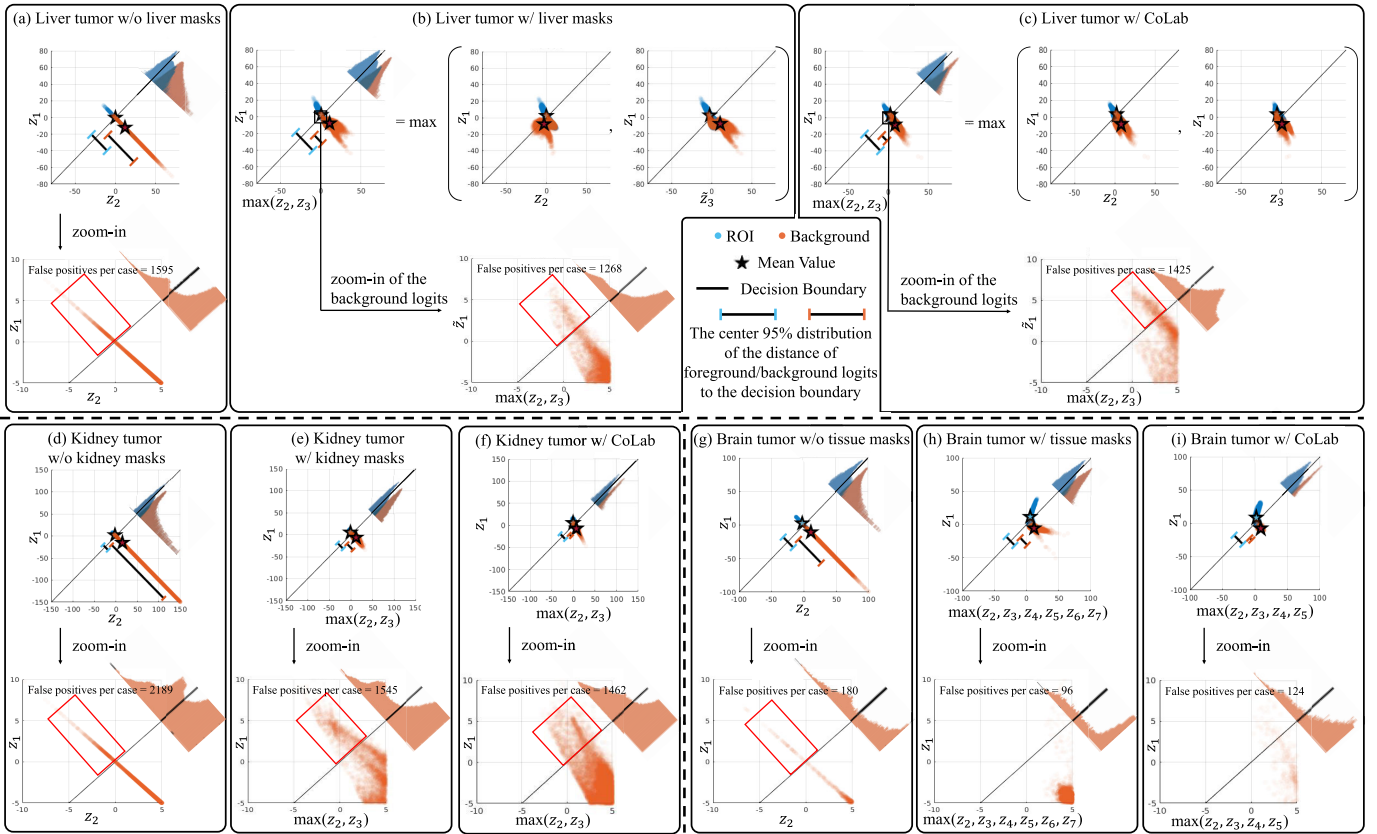
Fig. 3. The distribution of activations (logits) after the classification layer using ROI (blue) or background samples (orange) of test data as network input. Here, z_1 denotes logits for ROI classes, while z_i ($i > 1$) denotes logits for background classes. We also report the number of pixel-level predictions that are FP. Without context labels (a, d, g), CNNs fail to map the background samples to compact clusters. This is because CNNs are forced to learn generic filters for the background samples, yielding similar activations regardless of different background structures with diverse patterns/appearance. Consequently, the logit distribution of background samples would spread across the decision boundary, leading to increased FP. Extending the label space with anatomy masks (b, e, h) or CoLab (c, f, i) can alleviate this issue.

In order to fit the label space with context labels, we first extend the classification layer of $f_\phi(\cdot)$ with additional $t-1$ output neurons and obtain $\tilde{f}_\theta(\cdot)$ which map x_i to $\tilde{f}_\theta(x_i) \in \mathbb{R}^{t+1}$. We employ another model $g_\omega(\cdot)$ as the task generator parameterized by ω to produce context labels, with the network output $o_i = g_\omega(x_i) \in \mathbb{R}^t$. We have no requirements for the backbone of $g_\omega(\cdot)$ and empirically keep it the same with that of $f_\phi(\cdot)$.

We illustrate the training process of the proposed CoLab in Fig. 5. At the beginning of each iteration, we obtain the context labels termed as distance constrained label $\hat{\tilde{y}}_i$ by taking use of o_i and the ground truth y. This process is marked with ① and will be illustrated in Section IV-B and IV-C. Then, we calculate a new θ^* with one step of gradient descent with ② to access the impact of $\hat{\tilde{y}}_i$ on model training of ROI segmentation. Next, we optimize ω based on the second-order derivatives through a meta learning scheme with ③, which would be demonstrated in Section IV-D. Finally, we optimize the segmentation model $\tilde{f}_\theta(\cdot)$ based on the updated context labels with ④.

B. Label Aggregation

We calculate the context probability $(q_{ij})_{j=1}^t$ based on o_i via the softmax function as:

$$q_{ij} = \frac{e^{o_{ij}}}{\sum_{j=1}^t e^{o_{ij}}}. \tag{4}$$

The extended label $\tilde{y}_i \in \mathbb{R}^{t+1}$ is calculated by aggregating the original label $y_i = (y_{i1}, y_{i2})$ and the context probability q_i with:

$$\tilde{y}_{ij} = \begin{cases} y_{i1} & \text{if } j = 1, \\ q_{ij} & \text{if } j > 1 \text{ and } y_{i1} = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The label aggregation process is also illustrated in Fig. 6. By doing so, we can decompose the background class y_{i2} into t subclasses while ensure that the $\tilde{y}_i$ contains sufficient information about the ROI segmentation. For multi-class segmentation with totally c classes, $\tilde{y}_i$ can be calculated as:

$$\tilde{y}_{ij} = \begin{cases} y_{ij} & \text{if } j < c, \\ q_{ij} & \text{if } j \geq c \text{ and } y_{ij} = 0 \ \forall j < c, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

C. Context Constraints

Compared with background samples which are further away, the ones closer to the ROIs share similar characteristics with ROI samples and are more likely to be misclassified. In order to make the segmentation model focus more on those hard background samples that are close to ROIs, we make an assumption that all samples which are distant from the ROI need less attention and should be safely assigned the same
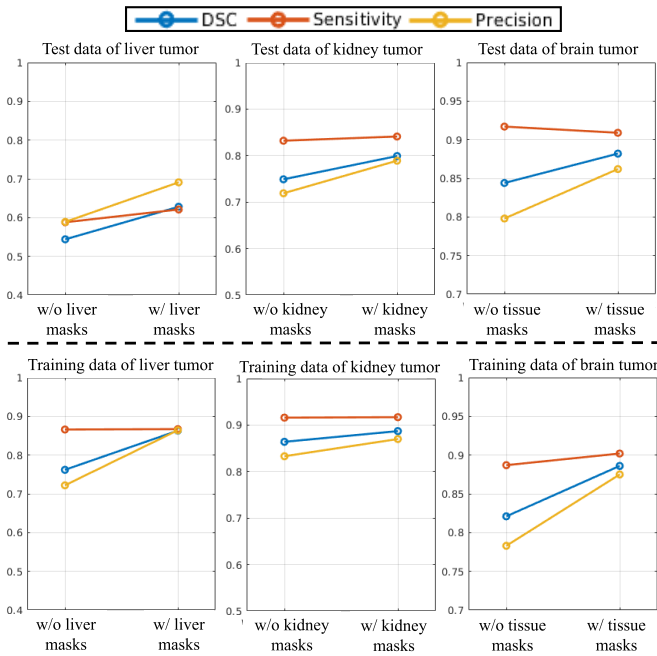
Fig. 4. Performance on liver tumor, kidney tumor and brain tumor segmentation on training data (lower part) and testing data (upper part) when networks were trained w/o and w/ context labels. Extending the label space with anatomy masks can help the model improve the performance by increasing the precision, while the sensitivity is largely retained.

background label. Specifically, we create a hard label $\boldsymbol{b}_i \in \mathbb{R}^{t+1}$ to represent the the background samples that are far from the ROI:

$$b_{ij} = \begin{cases} 1 & \text{if } j = 2, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

By utilizing the ROI label $\boldsymbol{y}_i$, we calculate the corresponding distance map d_i which is the Euclidean distance of a pixel to the closest boundary point of the ROI of any class. We set d_i to be zero for pixels inside the ROI.

We then calculate a soft dilated mask $\mathcal{M}_i$ based on d_i:

$$\mathcal{M}_i = \begin{cases} 1 & \text{if } d_i < m, \\ e^{\frac{-d_i + m}{\tau}} & \text{otherwise,} \end{cases} \tag{8}$$

where m is the margin controlling the model's focus on the pixels which are neighbouring to ROI and τ is the temperature to control the probabilities of the dilated regions. Empirically, m is set as 30 and τ is set as 20 for all our experiments. $\mathcal{M}_i$ would be 1 for pixels around ROIs and decrease close to 0 for pixels far from ROIs. We visualize an example of $\mathcal{M}_i$ with liver tumor in Fig. 7. The distance constrained label then is calculated as:

$$\hat{\tilde{\boldsymbol{y}}}_i = \mathcal{M}_i \tilde{\boldsymbol{y}}_i + (1 - \mathcal{M}_i)\boldsymbol{b}_i. \tag{9}$$

In this way, only the regions neighbouring to the tumor are considered to be classified as the contextual background class. Specifically, both $\tilde{f}_\theta(\cdot)$ and $g_\omega(\cdot)$ would be trained to focus on the regions which are close to ROIs.

D. Task Generator Optimization With Meta-Gradients

We formulate the optimization of CoLab as a bi-level problem:

$$\min_\omega \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{ROI}(\tilde{f}_{\theta*}(\boldsymbol{x}_i), \boldsymbol{y}_i) \tag{10}$$

$$s.t. \quad \theta^* = \underset{\theta}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{seg}(\tilde{f}_\theta(\boldsymbol{x}_i), \hat{\tilde{\boldsymbol{y}}}_i)), \tag{11}$$

where $\mathcal{L}_{ROI}(\tilde{f}_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i) = \mathcal{L}_c(\tilde{p}_{i1}, y_{i1})$ is computed on the ROI class. Different from $\mathcal{L}$, we choose $\mathcal{L}_c$ to be a criteria that can be calculated in the one-versus-all manner, such as binary cross entropy (BCE) and soft DSC loss, to represent the binary segmentation performance of the ROI class with more than two output logits.

We train the model using a batch of training samples with batch size n. For simplicity, we shorten $\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{ROI}(\tilde{f}_{\theta*}(\boldsymbol{x}_i), \boldsymbol{y}_i)$ as $\mathcal{L}_{ROI}(\theta^*)$ and $\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{seg}(\tilde{f}_\theta(\boldsymbol{x}_i), \hat{\tilde{\boldsymbol{y}}}_i)$ as $\mathcal{L}_{seg}(\theta, \omega)$ in the following descriptions. The bi-level optimization problem defined in Eq. 10 and 11 can be solved with gradient descent [10], [30]. Specifically, the derivative of $\mathcal{L}_{ROI}(\theta^*)$ w.r.t. ω can be calculated by applying the chain rule:

$$\nabla_\omega \mathcal{L}_{ROI}(\theta^*) = (\frac{\partial \theta^*}{\partial \omega})^\mathsf{T} \nabla_\theta \mathcal{L}_{ROI}(\theta^*), \tag{12}$$

where $\nabla_\omega = (\frac{\partial}{\partial \omega})^\mathsf{T}$ and $\nabla_\theta = (\frac{\partial}{\partial \theta})^\mathsf{T}$. One can compute $\frac{\partial \theta^*}{\partial \omega}$ based on implicit function theorem [4]. However, the derived result would contain a Hessian which is computational expensive and not always possible to access for deep neural networks. Among the many heuristics for the gradient approximation [10], [22], [30], we follow the solution described in [9], [22] to approximate θ^* by a single optimization step. Specifically, we sample a batch of training data and approximate the optimal inner variable θ^* with a step of gradient decent:

$$\theta^* \approx \theta - \alpha \nabla_\theta \mathcal{L}_{seg}(\theta, \omega), \tag{13}$$

where α is step size, which is kept the same with the learning rate of θ. We differentiate this equation w.r.t. ω from both sides yielding:

$$\frac{\partial \theta^*}{\partial \omega} = -\alpha \nabla^2_{\theta, \omega} \mathcal{L}_{seg}(\theta^*, \omega), \tag{14}$$

where $\nabla^2_{\theta, \omega} = \frac{\partial \nabla_\theta}{\partial \omega}$. By substituting Eq. 14 into Eq. 12, we can obtain the gradient on ω and update it with:

$$\begin{aligned} \omega^{t+1} &= \omega^t - \beta \nabla_{\omega^t} \mathcal{L}_{ROI}(\theta^*) \\ &= \omega^t + \alpha\beta \nabla^2_{\omega^t, \theta} \mathcal{L}_{seg}(\theta^*, \omega^t) \nabla_\theta \mathcal{L}_{ROI}(\theta^*), \end{aligned} \tag{15}$$

where β is the learning rate to update ω. In this way, the task generator g_ω is explicitly trained to produce effective context labels with the second-order gradients. Although Eq. 15 contains an expensive vector-matrix product, it is feasible to calculate with prevailing machine learning frameworks such as PyTorch [29]. We find CoLab can be efficient as it costs as little as 30% additional training time. We summarize the implementation details in supplementary material. The full algorithm is summarized in Algorithm 1.
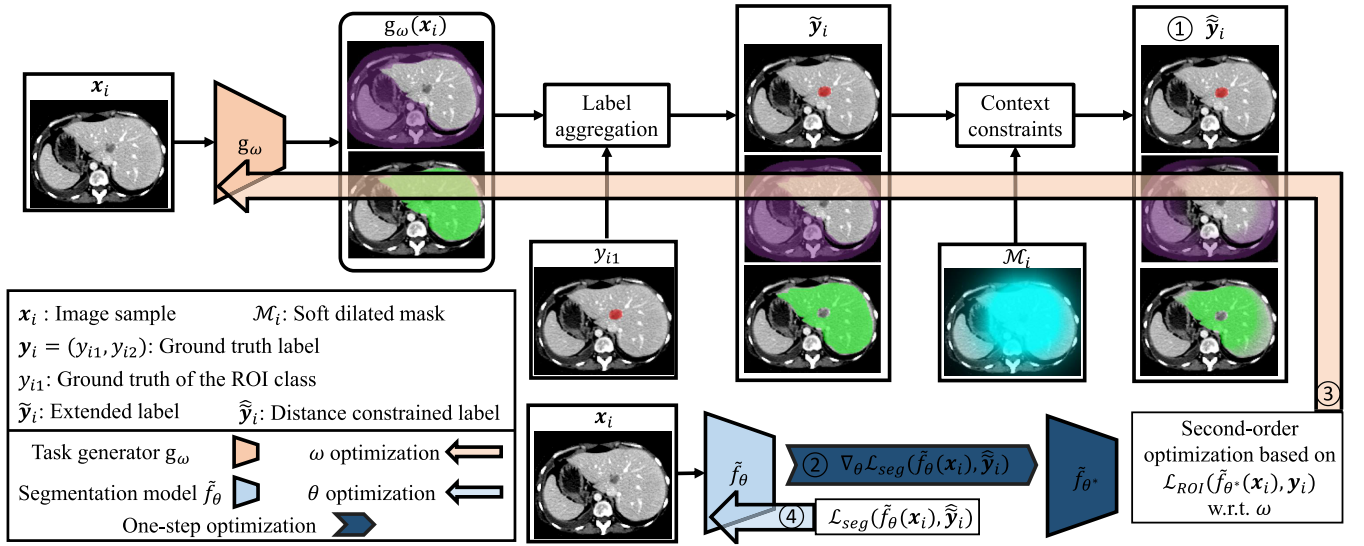
Fig. 5. Illustration of the training process of CoLab. $g_\omega(\cdot)$ is optimized to generate context labels such that it can help reduce the training loss of the ROI in the most effective way, see Section IV-A for more details.

Algorithm 1 Context Label Learning (CoLab)

Require:

$\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$: training data; t: the number of context classes, $f_\phi(\cdot)$: the segmentation model which produces $f_\phi(\boldsymbol{x}_i) \in \mathbb{R}^2$; $g_\omega(\cdot)$: the task generator which produce $g_\omega(\boldsymbol{x}_i) \in \mathbb{R}^t$.

α, β: learning rates to update θ and ω.

1: Extend the classification layer of $f_\phi(\cdot)$ to get $\tilde{f}_\theta(\cdot)$ which produce $\tilde{f}_\theta(\boldsymbol{x}_i) \in \mathbb{R}^{t+1}$.
2: **for** each iteration **do**
3:      Sample a batch of data $\mathcal{B} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ from $\mathcal{D}$.
4:      **for** a number of steps **do** ▷ *Note: One step is sufficient in our experiments.*
5:          Generate the context probability $\boldsymbol{q}_i$ based on the output of task generator $g_\omega(\boldsymbol{x}_i)$ via Eq. 4.
6:          Generate the extended label $\tilde{\boldsymbol{y}}_i$ by aggregating the context probability $\boldsymbol{q}_i$ with $\boldsymbol{y}_i$ via Eq. 5.
7:          Calculate a loss mask $\mathcal{M}_i$ via Eq. 8 and generate the distance constrained label $\hat{\boldsymbol{y}}_i$ via Eq. 9.
8:          Calculate a new θ^* with the gradient descent optimization algorithm via Eq. 13.
9:          Optimize ω based on meta-gradients via Eq. 15. ▷ *Training the task generator.*
10:      **end for**
11:      Update θ with gradient decent based on updated ω. ▷ *Training the segmentation model.*
12: **end for**

V. Experiments

A. Experimental Setup

1) Network Configurations: We use a state-of-the-art 3D U-Net [13] as the network backbone for both the segmentation model $f_\phi/\tilde{f}_\theta$ and the task generator g_ω. We normalize all datasets with the built-in pipeline following [13]. Specifically, we adopt case-wise Z-score normalization within brain masks
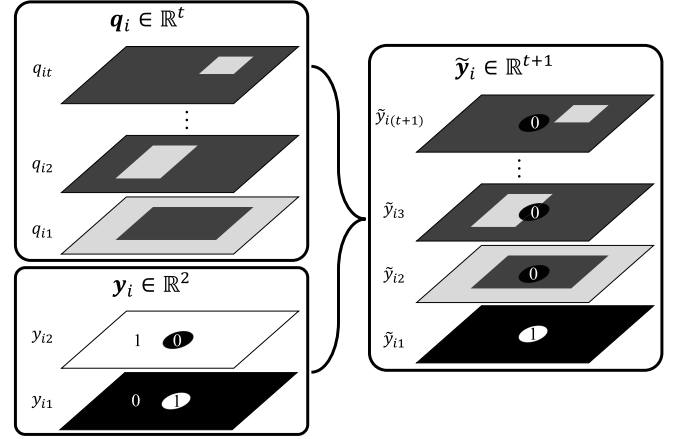

Fig. 6. Illustration of the label aggregation process. We generate the extended label $\tilde{y}_i$ by aggregating the original label y_i with the context probability q_i.
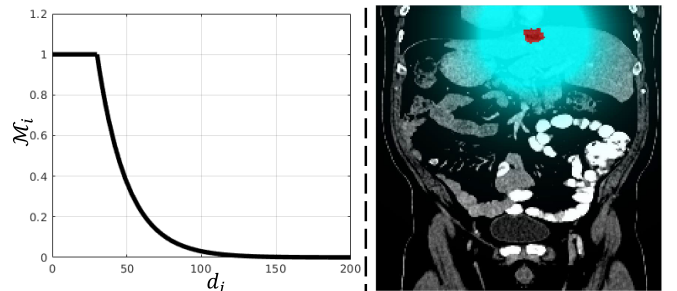

Fig. 7. The task generator focuses on the context which is near to the ROI. To do so, we train the segmentation model to classify the pixels which are outside the soft dilated ROI as the same class (pixels marked without color in the right figure). (Left part) The value of a soft dilated mask M_i with $m = 30$ and $\tau = 20$. (Right part) The visualization of the dilated mask M_i (blue) with a liver tumor (red) of a training case.

for MR images, while we employ dataset-wise Z-score normalization based on ROI samples for CT images after clipping the Hounsfield units (HU) from 0.5% to 99.5%. We utilize the default data augmentation policies for all experiments. We adopt a combination of CE and sample-wise soft DSC

loss with equal weight for $\mathcal{L}$ while using BCE for $\mathcal{L}_c$. We use a batch size of 2 and patch size of $80 \times 80 \times 80$. We train the networks for 1000 epochs for brain tumor and lesion segmentation and 2000 epochs for liver, kidney, colon and pancreas tumor segmentation as we observed that the network needed more iterations to converge latter tasks. We summarize the hyper-parameters of all experiments in supplementary material. All reported results are the average of two runs with different random seeds.

2) Liver Tumor Segmentation: We evaluate CoLab for liver tumor segmentation with the training dataset from the Liver Tumor Segmentation Challenge (LiTS) which contains 131 CT images. We exclude the samples which do not contain any liver tumor leaving 118 cases. We resample all CT images to a common voxel spacing of $1.9 \times 1.9 \times 2.5$ mm following [13]. We train models with 83 cases and test on 35 cases.

3) Kidney Tumor Segmentation: We further conduct experiments using the training dataset of the Kidney Tumor Segmentation Challenge (KiTS) which contains 210 CT images. We resample all CT images to a common voxel spacing of $1.6 \times 1.6 \times 3.2$ mm following [13]. We tested on 70 cases and used the other 140 cases as the training data.

4) Colon Tumor Segmentation: We evaluate CoLab for the case of colon tumor segmentation from CT images. We collect 126 colon cancer CT images from the training dataset of the Medical Segmentation Decathlon challenge [34]. We resample all CT images to the voxel spacing of $1.6 \times 1.6 \times 3.1$ mm following [13]. We train models with 88 cases and test on the other 38 cases.

5) Brain Tumor Segmentation: We also conduct experiments for brain tumor segmentation using the VS dataset [33] which contains 243 paired T1-weighted and T2-weighted MR images. We only used T2-weighted MR images for evaluating CoLab but used T1-weighted MR images to generate the brain structure masks with MALP-EM. We did not use brain masks for the histogram normalization of this task because VS can appear at the brain boundary and may be excluded when using common brain extraction algorithms. All MR images have the same isotropic voxel spacing of 1.0 mm^3. We train the models with 176 cases and test on 46 cases following [33].

6) Brain Stroke Lesion Segmentation: Additionally, we evaluate CoLab with brain stroke lesion segmentation using the Anatomical Tracings of Lesions After Stroke (ATLAS) dataset [21] which contains 220 T1-weighted MR images. The MR images all have the same voxel spacing of 1.0 mm^3. We randomly selected 145 cases as training data and left the rest 75 cases for testing.

7) Pancreas and Pancreatic Tumor Mass Segmentation: We further evaluate Colab in the setting of multi-class segmentation. Specifically, we train the segmentation model to segment three classes including pancreas, pancreatic tumor mass and background. We aim to find the context labels which can benefit the segmentation of both foreground classes. We collect 281 CT images containing pancreas tumor from the training dataset of the Medical Segmentation Decathlon challenge [2], [3], [8], [34]. We resample all the CT images to the voxel spacing of $1.3 \times 1.3 \times 2.6$ mm following [13]. We randomly split the dataset into 197 cases for training and 84 cases for testing.

B. Compared Methods and Processing

1) Context Labels Based on k-Means: We compare CoLab with alternative approaches for context label generation, including a context label generation via clustering. Here, we take pixels inside the body masks or brain masks as samples and employ k-means [1] to construct t clusters. We kept t the same with the class number of human-defined anatomy masks. Specifically, we chose $t = 2$ for liver, kidney, colon and pancreas tumor segmentation and $t = 6$ for brain tumor and brain stroke lesion segmentation.

2) Context Labels Based on Dilated Masks: We also compare with a baseline using dilated masks of the ROI as the context labels. This idea is somewhat similar to label smoothing [14], [26] where models are trained with blurred ROI labels. Specifically, we take the soft dilated masks $\mathcal{M}_i$ defined in Eq. 8 and set the context probability as $\boldsymbol{p}_i = [1 - \mathcal{M}_i, \mathcal{M}_i]^\mathsf{T}$.

3) Context Labels Predicted With External Datasets: We further evaluate and compare to an approach which leverages prior knowledge from other datasets. Specifically, we trained a segmentation model with 20 CT images using data from [16] which contains labels of 14 abdominal organs including liver and kidney. We resample the 20 CT images to the voxel spacing of $1.6 \times 1.6 \times 3.2$ mm. We reduce the dependency of the model on longitudinal axis and trained the segmentation model with a patch size of $128 \times 128 \times 32$ as the slice numbers differ across datasets. We then apply this segmentation model to the resampled training split of LiTS as well as KiTS, and extract the liver and kidney masks as the automatically generated contextual anatomy masks.

4) Post-Processing: We also compare with a common strategy to suppress FP in segmentation based on component-based post-processing, which is widely adopted in many segmentation pipelines [13]. Specifically, we assume there is always one ROI and remove all but the largest region. For the cases when ROIs contain multiple classes (pancreas and pancreatic tumor segmentation), we take all the ROIs as a whole and only keep the largest component.

C. Quantitative Results

We summarize the quantitative results of the models trained with different types of context labels on five binary segmentation datasets in Table I and one multi-class segmentation dataset in Table II. We separate the results according to whether expert knowledge and manual labeling efforts are needed. Taking the manual segmentation results as the ground truth, we calculate DSC, sensitivity (SEN), precision (PRC) and HD. We show some segmentation results with CoLab in Fig. 2(c) and also summarize all the corresponding context labels in Fig. 8. We provide additional examples of context labels in supplementary material.

As discussed in Section III, the baseline segmentation model seems to underfit the heterogeneous background samples resulting in many FP and decreased precision. We find most evaluated context labels approaches can improve the models overall segmentation accuracy with increased precision. We observe that adding contextual labels consistently benefits model training and improves performance.

TABLE I

EVALUATION OF DIFFERENT BINARY SEGMENTATION TASKS WITH DIFFERENT TYPES OF CONTEXT LABEL GENERATION APPROACHES. WE DIVIDE THE RESULTS ACCORDING TO WHETHER EXPERT KNOWLEDGE AND LABELING EFFORTS ARE INCLUDED. THE BEST AND SECOND BEST RESULTS WITHOUT HUMAN EFFORTS ARE IN **BOLD** WITH THE BEST ALSO **UNDERLINED**

Task	Method	t	DSC	SEN	PRC	HD
Liver tumor [5]	w/o liver masks	1	54.4	58.8	58.9	111.1
	K-means [1]	2	**61.4**	**61.4**	67.0	71.9
	Dilated masks [26]	2	60.7	59.8	**68.0**	67.6
	CoLab	2	**62.5**	**62.8**	67.3	69.4
	CoLab	4	57.3	60.5	62.3	**56.3**
	CoLab	6	59.7	60.3	65.2	**43.6**
	w/ model-predicted liver masks [16]	2	62.4	61.6	70.6	44.1
	w/ liver masks [5]	2	62.8	62.1	69.1	53.5
Kidney tumor [12]	w/o kidney masks	1	74.9	83.2	71.9	120.4
	K-means [1]	2	**76.8**	83.5	74.3	87.1
	Dilated masks [26]	2	76.4	**83.9**	73.1	95.3
	CoLab	2	**78.5**	82.2	**77.7**	**75.7**
	CoLab	4	76.4	80.6	**76.5**	**63.7**
	CoLab	6	74.9	81.0	73.3	79.4
	w/ model-predicted kidney masks [16]	2	79.2	81.3	82.7	38.1
	w/ kidney masks [12]	2	79.9	84.1	78.9	54.7
Colon tumor [34]	w/o context labels	1	45.6	**53.9**	45.3	154.9
	K-means [1]	2	44.9	47.2	48.7	111.7
	Dilated masks [26]	2	**47.1**	49.8	**52.8**	**83.3**
	CoLab	2	**48.9**	**53.9**	50.3	86.0
	CoLab	4	45.3	48.1	**51.5**	**81.2**
	CoLab	6	46.1	**50.2**	49.7	88.9
Brain tumor [33]	w/o tissue masks	1	84.3	91.2	80.5	15.2
	K-means [1]	6	85.0	91.3	80.3	9.4
	Dilated masks [26]	2	84.8	91.6	81.1	8.8
	CoLab	2	85.2	90.4	82.5	7.8
	CoLab	4	**89.0**	**91.7**	**86.7**	**1.4**
	CoLab	6	87.9	**92.0**	84.9	**2.5**
	w/ tissue masks [17], [33]	6	88.2	90.9	86.2	3.1
Brain lesion [21]	w/o context labels	1	58.5	**68.2**	58.6	51.7
	K-means [1]	6	60.0	60.7	**70.5**	**19.5**
	Dilated masks [26]	2	61.9	**69.8**	63.3	35.3
	CoLab	2	**62.8**	**68.2**	67.4	28.1
	CoLab	4	**62.6**	68.1	65.4	25.1
	CoLab	6	61.7	66.7	65.7	**24.5**

TABLE II

EVALUATION OF PANCREAS AND PANCREATIC TUMOR MASS SEGMENTATION WITH DIFFERENT TYPES OF CONTEXT LABEL GENERATION APPROACHES. THE BEST AND SECOND BEST RESULTS ARE IN **BOLD** WITH THE BEST ALSO **UNDERLINED**

Model	t	Pancreas				Tumor mass			
		DSC	SEN	PRC	HD	DSC	SEN	PRC	HD
w/o context labels	1	82.9	82.9	85.1	13.9	48.9	**49.2**	59.3	**52.0**
K-means [1]	2	82.1	82.3	84.6	14.6	44.8	45.4	54.8	64.4
Dilated masks [26]	2	**83.5**	**84.6**	84.4	11.4	48.1	48.2	58.0	58.2
CoLab	2	**83.3**	83.3	**85.3**	**7.7**	**51.1**	**50.8**	**63.2**	**44.8**
CoLab	4	82.7	81.7	**86.0**	**7.3**	48.7	45.9	**62.0**	57.3
CoLab	6	82.7	82.9	84.7	8.0	**49.6**	48.4	61.5	52.8

1) Performance of Baseline Methods: K-means based context labels divide the background samples into different classes based on pixel intensity. Those context labels seem to help the segmentation model separating the abdominal organs and bones from background such as fat and air for the abdominal CT images while in the case of brain MR, it may help segmenting the images into different regions such as white matter and ventricles. We find these k-means based context labels are effective for liver and kidney tumor segmentation, while being less effective for colon tumor, brain tumor, brain lesion and pancreas tumor segmentation. This may be due to the distinct intensities of large organs and background in CT where the generated labels have meaningful semantic information, benefiting the model training by enhancing the context representations of kidney and liver tumor. Because the appearance of colon tumor is heterogeneous, its segmentation depends more on spatial information and cannot benefit from purely intensity based context labels. As the pancreas label already provides enriched context information for pancreatic tumor segmentation, we find baseline context label techniques such as k-means and dilated masks cannot provide additional benefits. The structure of the brain anatomy in MR is more complex and the classes generated by k-means would be noisy and disconnected. Here, the k-means generated labels are less beneficial for as they provide insufficient context for the ROI.

The context labels based on dilated masks make the segmentation model aware of regions neighbouring the ROIs. We find those context labels are effective for liver tumor, kidney tumor,
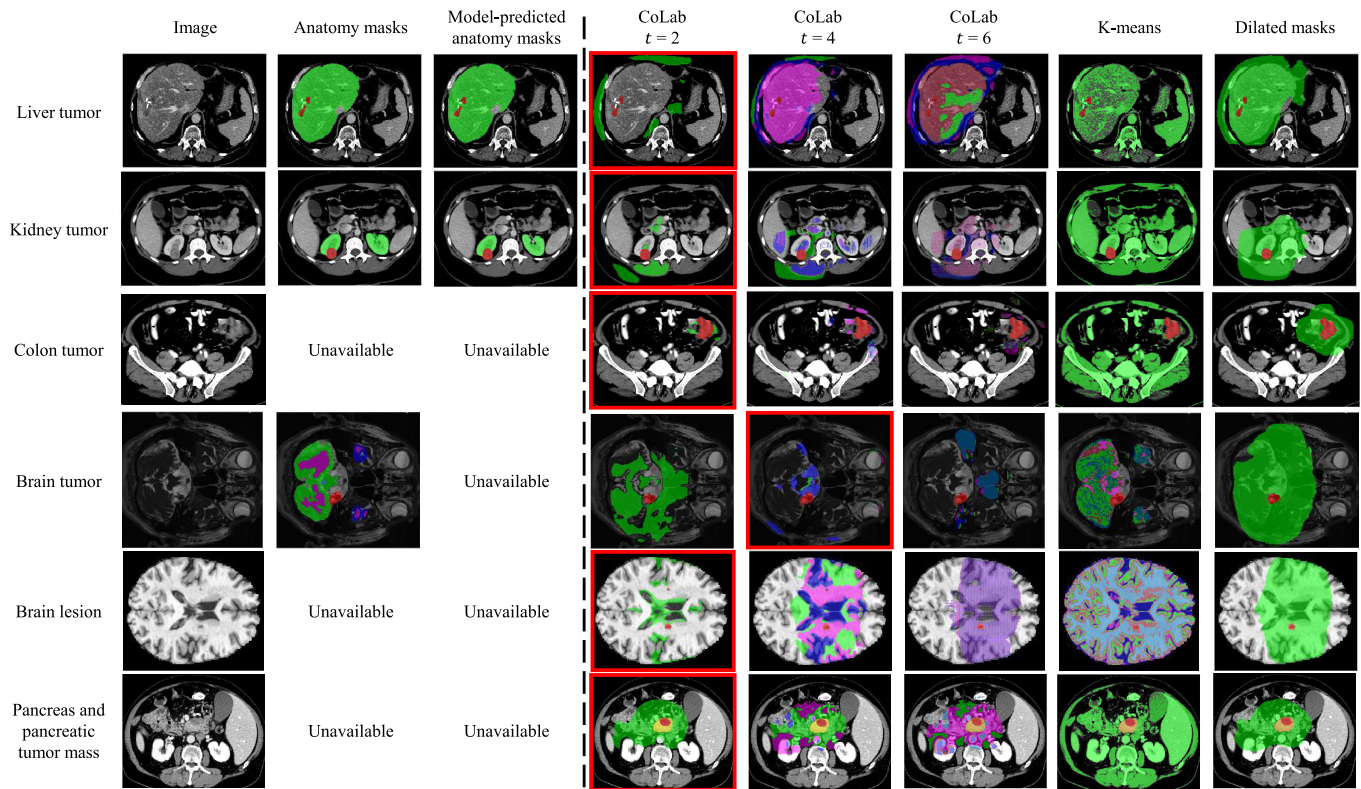
Fig. 8. Examples of the results $\tilde{f}_\theta(x_i)$ of training data with different types of context labels. The binary ROI masks are shown in red, while multi-class ROI masks are shown in red and orange. The context labels are visualized using other colors The context labels which benefit ROI segmentation most and generated without human input are marked with red boxes. CoLab learns to generate specific structure masks for different tasks. This helps the segmentation network to learn better feaure representations for the background class.

colon tumor, brain lesion and pancreas segmentation but yield limited improvements for brain tumor segmentation. This may be because liver tumors, kidney tumors, colon tumors and brain lesions appear in varying parts of the primary organ. If trained as binary tasks, the segmentation model predicts FP outside the liver, kidney and colon regions or in anatomically unrelated parts of the brain because these samples share similar characteristics with the ROIs, as shown in Fig. 2(a). In this case, the dilated masks could enhance the spatial relationship between the ROI and the surrounding context. For example, the dilated masks would connect the liver tumor segmentation with liver context and avoid FP outside the liver regions. However, the samples of VS dataset only contain one ROI per case and VS always appear around the superior vestibular area. Therefore, the segmentation model does not need more spatial information to locate the ROIs and does not benefit from the dilated masks.

As shown in the third column of Fig. 8, we obtain high-quality automatic segmentations of liver and kidney regions as the segmentation of large organs is relatively easy and robust in CT images. The model-predicted masks show similar improvements compared to using the manually labeled anatomy masks. However, this approach is not applicable when we do not have prior knowledge available about the ROIs or when we do not have access to external, manually labelled datasets to train supervised models for context label generation.

We summarize the results with component-based post-processing in supplementary material. We find post-processing always decreases the overall performance indicated by DSC

for liver tumor, colon tumor and brain lesion segmentation as there always exist multiple separate ROIs in those tasks. For kidney tumor, pancreas tumor and brain tumor segmentation, we find that post-processing is effective in improving the overall segmentation performance of most results while the models trained with context labels are always better than the ones trained with binary tasks. This indicates that the FP next to ROIs cannot be eliminated with simple post-processing.

2) CoLab: As illustrated in Fig. 2(c), CoLab can effectively reduce different types of false positive predictions and improve precision while preserving sensitivity. CoLab can be applied to different tasks with similar hyper-parameters and show consistent improvements ($+2.2$ to $+8.1$ points of DSC). We find CoLab yields better segmentation results than all the other context label approaches that do not use expert knowledge. Moreover, CoLab shows similar improvements compared to using organ masks for liver and kidney tumor segmentation and even better performance for brain tumor segmentation, compared with using tissue masks. It is particularly interesting to find that CoLab can improve the segmentation performance of pancreatic tumor mass, on the top of its strong context label of pancreas. It indicates that there is still room to improve the context representation for segmentation tasks with human defined context labels. CoLab is a flexible and generic method and can be directly applied to boost performance for a variety of segmentation tasks with heterogeneous background classes. We further analyze the context labels generated by CoLab based on intensity histograms and summarize the results in supplementary material. We find the context labels

generated by CoLab can highlight the regions which have different intensity distributions from the background samples. The intensity of one of the generated context labels is always similar to that of ROIs. This might indicates that CoLab can generate useful and semantic context labels which can help the segmentation model better distinguish ROIs from similar background samples.

We evaluate CoLab with different number context classes t. We find that when the segmentation model tends to make FP far from the ROI such as liver tumor, kidney tumor, colon tumor, brain lesion and pancreas tumor, $t = 2$ is the most effective. This may indicate that the segmentation would need the context labels to provide mostly spatial information and $t = 2$ makes CoLab focus on the location of the ROI as the task generator only needs to separate one class of context from the remaining background. This single class would enhance the spatial relationship between the ROI and its surrounding objects. This is consistent with the observations of the results as shown in Fig. 8. Specifically, the context labels generated with $t = 2$ highlight the region next to the liver, kidney and colon regions and could implicitly help the segmentation model focus on the vicinity of the ROI boundaries.

When the baseline segmentation model is making FP around the ROIs such as in the case of VS segmentation, $t = 4$ and $t = 6$ seem more effective. This is probably because the segmentation requires more structured information of surrounding objects and the complicated context of brain anatomy could be better represented when divided into more than 2 classes. As illustrated in Fig. 8, context labels generated with $t = 4$ highlight structure information around the ROIs. Specifically, CoLab generates segmentations of the superior vestibular area for VS segmentation where the baseline had predicted FP. Similarly, with the task of brain lesion segmentation, the context labels highlight the regions around ventricles to avoid FP in this area. By doing so, the model learns spatial information by building specific representations for the ventricle regions.

We should note that CoLab may not be very effective with large t, specifically we find CoLab with $t = 6$ would not bring much improvements for kidney tumor, colon tumor and brain lesion segmentation. As shown in Fig. 8, we observe that the generated context labels would yield checkerboard patterns. This might be because further decomposing the background class cannot make the task easier to learn. In this case, we find the task generator cannot generate coherent labels and produce similar probabilities for the different classes in those regions, making the context labels trivial and hard to learn.

In general, we find $t = 2$ is the most effective choice for most cases in terms of DSC, especially for tumor segmentation in abdominal CT images. This might indicate that the background representation is not so complex and could be improved by adding refined spatial information. When trained with t larger than 2, the models could have better results in terms of HD, compared with the models trained with $t = 2$. In this case, the models learn the background class better, as they make less FP away from the ROIs. However, choosing t larger than 2 seems to make the models fail to learn the foreground class well, resulting in worse DSC and sensitivity.

This may be because increasing t could make the prediction of foreground class more difficult.

We suggest the use of $t = 2$ when the baseline model makes FP distant from the ROIs, and $t = 4$ when over-segmentation is observed closer to the ROIs. In addition, we look into the intensity histograms of ROIs for different datasets and summarize the results in supplementary material. We find the intensity distributions of most ROIs are not distinct from that of the background samples, apart from brain tumor. Perhaps for brain tumor segmentation, the model can benefit from knowing the regions which share the most similar intensity distributions with the ROIs, thus requiring more classes of context labels. This can be a potential factor due to which the optimal class of context labels t for brain tumor is 4 instead of 2.

D. Effect on Logit Distributions

We also show the effect of CoLab ($t = 2$ for liver and kidney tumor segmentation and $t = 4$ for brain tumor segmentation) on the logit distributions in Fig. 3(c, f, i). We find CoLab has a similar regularization effect than using anatomy masks. By decomposing the background class, the segmentation model can robustly map the background samples to different locations away from the decision boundary, reducing the logit shifts, and thus yielding fewer FP.

VI. Conclusion

In this study, we find that segmentation models are prone to over-segment the ROIs in the presence of heterogeneous background classes. With the observation of network behaviour across multiple segmentation tasks, we conclude that this is due to the model being unable to learn discriminative representations from the background samples and the models can be significantly improved by incorporating context labels during training. We present CoLab, a generic method to automatically generate effective context labels through a meta-learning scheme. We show CoLab improves overall segmentation performance substantially for several challenging segmentation tasks, without the need of expert knowledge and laborious labeling efforts. In future work, we will evaluate CoLab in settings such as domain shifts. It will also be interesting to explore automatic context label generation in multi-task settings where a shared backbone model is trained to automatically identify interesting subclasses across tasks and even datasets.

References

[1] D. Arthur and S. Vassilvitskii, "K-Means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*. New Orleans, LA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[2] M. A. Attiyeh et al., "Survival prediction in pancreatic ductal adeno-carcinoma by quantitative computed tomography image analysis," *Ann. Surgical Oncol.*, vol. 25, no. 4, pp. 1034–1042, Apr. 2018.

[3] M. A. Attiyeh et al., "Preoperative risk prediction for intraductal papillary mucinous neoplasms by quantitative CT image analysis," *HPB*, vol. 21, no. 2, pp. 212–218, Feb. 2019.

[4] Y. Bengio, "Gradient-based optimization of hyperparameters," *Neural Comput.*, vol. 12, no. 8, pp. 1889–1900, 2000.

[5] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*.

[6] F. Bragman, R. Tanno, S. Ourselin, D. Alexander, and J. Cardoso, "Stochastic filter groups for multi-task CNNs: Learning specialist and generalist convolution kernels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1385–1394.

[7] J. Cai, Y. Wang, and J.-N. Hwang, "ACE: Ally complementary experts for solving long-tailed recognition in one-shot," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 112–121.

[8] J. Chakraborty et al., "CT radiomics to predict high-risk intraductal papillary mucinous neoplasms of the pancreas," *Med. Phys.*, vol. 45, no. 11, pp. 5019–5029, Nov. 2018.

[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[10] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1568–1577.

[11] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, "Joint classification-regression forests for spatially structured multi-object segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2012, pp. 870–881.

[12] N. Heller et al., "The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," 2019, *arXiv:1904.00445*.

[13] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020.

[14] M. Islam and B. Glocker, "Spatially varying label smoothing: Capturing uncertainty from expert annotations," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2021, pp. 677–688.

[15] B. Kang et al., "Decoupling representation and classifier for long-tailed recognition," 2019, *arXiv:1910.09217*.

[16] B. A. Landman, Z. Xu, J. E. Igelsias, M. Styner, T. R. Langerak, and A. Klein. (2015). *MICCAI Multi-Atlas Labeling Beyond the Cranial Vault—Workshop and Challenge*. Accessed: Dec. 2020. [Online]. Available: https://www.synapse.org/#!Synapse:syn3193805, doi: 10.7303/syn3193805.

[17] C. Ledig et al., "Robust whole-brain segmentation: Application to traumatic brain injury," *Med. Image Anal.*, vol. 21, no. 1, pp. 40–58, 2015.

[18] Z. Li, K. Kamnitsas, and B. Glocker, "Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 402–410.

[19] Z. Li, K. Kamnitsas, and B. Glocker, "Analyzing overfitting under class imbalance in neural networks for image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 1065–1077, Mar. 2020.

[20] Z. Li, J. Yu, Y. Wang, H. Zhou, H. Yang, and Z. Qiao, "DeepVolume: Brain structure and spatial connection-aware network for brain MRI super-resolution," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3441–3454, Jul. 2019.

[21] S.-L. Liew et al., "A large, open source dataset of stroke anatomical brain images and manual lesion segmentations," *Sci. Data*, vol. 5, no. 1, Feb. 2018, Art. no. 180011.

[22] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," 2018, *arXiv:1806.09055*.

[23] S. Liu, A. J. Davison, and E. Johns, "Self-supervised generalisation with meta auxiliary learning," 2019, *arXiv:1901.08933*.

[24] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2537–2546.

[25] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[26] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" 2019, *arXiv:1906.02629*.

[27] A. Navon, I. Achituve, H. Maron, G. Chechik, and E. Fetaya, "Auxiliary learning by implicit differentiation," 2020, *arXiv:2007.02693*.

[28] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Sensitivity and generalization in neural networks: An empirical study," 2018, *arXiv:1802.08760*.

[29] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8026–8037.

[30] F. Pedregosa, "Hyperparameter optimization with approximate gradient," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 737–746.

[31] T. Qaiser et al., "Multiple instance learning with auxiliary task weighting for multiple myeloma classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 786–796.

[32] A. A. A. Setio et al., "Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1160–1169, Mar. 2016.

[33] J. Shapey et al., "An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI," *J. Neurosurgery*, vol. 134, no. 1, pp. 171–179, Jan. 2019.

[34] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv:1902.09063*.

[35] V. V. Valindria et al., "Small organ segmentation in whole-body MRI using a two-stage FCN and weighting schemes," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, Cham, Switzerland: Springer, 2018, pp. 346–354.

[36] L. Xiang, G. Ding, and J. Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 247–263.

[37] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[38] D. Zhang et al., "Exploring task structure for brain tumor segmentation from multi-modality MR images," *IEEE Trans. Image Process.*, vol. 29, pp. 9032–9043, 2020.

[39] D. Zhang, J. Zhang, Q. Zhang, J. Han, S. Zhang, and J. Han, "Automatic pancreas segmentation based on lightweight DCNN modules and spatial prior propagation," *Pattern Recognit.*, vol. 114, Jun. 2021, Art. no. 107762.

[40] Y. Zhou et al., "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10672–10681.

[41] 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 *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 693–701.