Anti-interference from Noisy Labels: Mean-Teacher-assisted Confident Learning for Medical Image Segmentation

Zhe Xu, Donghuan Lu, Jie Luo, Yixin Wang, Jiangpeng Yan, Kai Ma, Yefeng Zheng, *Fellow, IEEE*, and Raymond Kai-yu Tong, *Senior Member, IEEE*

***Abstract*— Manually segmenting medical images is expertise-demanding, time-consuming and laborious. Acquiring massive high-quality labeled data from experts is often infeasible. Unfortunately, without sufficient high-quality pixel-level labels, the usual data-driven learning-based segmentation methods often struggle with deficient training. As a result, we are often forced to collect additional labeled data from multiple sources with varying label qualities. However, directly introducing additional data with low-quality noisy labels may mislead the network training and undesirably offset the efficacy provided by those high-quality labels. To address this issue, we propose a Mean-Teacher-assisted Confident Learning (MTCL) framework constructed by a teacher-student architecture and a label self-denoising process to robustly learn segmentation from a small set of high-quality labeled data and plentiful low-quality noisy labeled data. Particularly, such a synergistic framework is capable of simultaneously and robustly exploiting (i) the additional dark knowledge inside the images of low-quality labeled set via perturbation-based unsupervised consistency, and (ii) the productive information of their low-quality noisy labels via explicit label refinement. Comprehensive experiments on left atrium segmentation with simulated noisy labels and hepatic and retinal vessel segmentation with real-world noisy labels demonstrate the superior segmentation performance of our approach as well as its effectiveness on label denoising.**

***Index Terms*— Medical Image Segmentation, Noisy Label, Label Denoising.**

I. Introduction

DEEP learning (DL) has greatly advanced medical image segmentation, where the success of most DL-based methods relies on a large amount of high-quality (HQ) labeled data. However, manually segmenting medical images is extremely expertise-demanding, time-consuming and laborious. Considering the social efficiency, in practice, acquiring massive high-quality labeled data (termed as Set-HQ) from experts is always infeasible. Unfortunately, without sufficient HQ pixel-level labels, the data-hungry learning-based segmentation methods often struggle with overfitting, leading to inferior performance. To relieve this issue, we often resort to collecting additional labeled data with varying label qualities, e.g., crowdsourcing from non-experts or using machine-generated labels without any quality control, as depicted in Fig. 1. However, directly introducing additional data with low-quality (LQ) noisy labels (termed as Set-LQ) may confuse the network training and undesirably offset the efficacy provided by those HQ labels, which easily leads to performance degradation instead [1], [2]. Therefore, how to effectively and robustly exploit the additional information in plentiful LQ noisy labeled data is crucial to the medical image analysis community.

Such a pervasive dilemma motivates several efforts [2]–[6] to alleviate the negative effects brought by LQ labels. However, this challenging topic is still underexplored while existing literature on learning segmentation with noisy labels ignores the clear distinction of their applicable scenarios, causing ambiguous benchmarks. For example, some approaches [4]–[6] assumed that they collected and mixed data from multiple sources, i.e., Set-HQ and Set-LQ are indiscriminate, as shown in Fig. 1 (a). In contrast, other techniques [1]–[3] were developed for another distinct practical scenario—inviting experts to label or perform quality control for a small set of data and thus making Set-HQ and Set-LQ separated, as shown in Fig. 1 (b). To further clarify the problem setting, here, we define the former scenario as **Set-HQ-agnostic**, while the latter one as **Set-HQ-knowable**. In this work, we focus on the latter setting because (i) collaborating with experienced radiologists to acquire a reasonable amount of HQ labeled data from them is feasible in clinical practice; (ii) such a separated strategy implies that we embed prior knowledge to help the network distinguish between HQ labeled data and LQ labeled data.

Tailoring for the Set-HQ-knowable scenario, we propose

This research was done with Tencent Healthcare (Shenzhen) Co., LTD and Tencent Jarvis Lab and supported by General Research Fund from Research Grant Council of Hong Kong (No. 14205419) and the Scientific and Technical Innovation 2030-"New Generation Artificial Intelligence" Project (No. 2020AAA0104100).

Z. Xu and R. Tong are with Department of Biomedical Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China. (e-mail: jackxz@link.cuhk.edu.hk; kytong@cuhk.edu.hk).

D. Lu, K. Ma and Y. Zheng are with Tencent Healthcare (Shenzhen) Co., LTD and Tencent Jarvis Lab, Shenzhen, China.

J. Luo is with Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

Y. Wang is with Department of Bioengineering, Stanford University, Stanford, CA, USA.

J. Yan is with Department of Automation, Tsinghua University, Beijing, China.

Corresponding authors: R. Tong (e-mail: kytong@cuhk.edu.hk) and D. Lu (e-mail: caleblu@tencent.com)

a novel Mean-Teacher-assisted Confident Learning (MTCL) framework, which is constructed by a self-ensembling teacher-student architecture and a label self-denoising process to robustly learn segmentation from a small set of high-quality labeled data and plentiful low-quality noisy labeled data. By encouraging teacher-student consistency under different perturbations for the same input, the network can additionally exploit the dark knowledge inside the images of Set-LQ. Synergistically, based on the classification noise process (CNP) assumption [7] that label noise should be class-conditional, we remold confident learning (CL) [8], which was initially proposed for removing incorrect labels in image-level classification, to consecutively characterize the pixel-wise label error map assisted by a "third party", i.e., the teacher model. Observing different tasks suffer various inherent difficulties, subsequently, we propose a family of label refinement strategies including hard, fixed smooth and uncertainty-aware dynamic smooth to suit various clinical practices, allowing the network to receive more productive and robust guidance from the additional LQ noisy labeled data. We conduct comprehensive experiments on left atrium segmentation with simulated noisy labels, hepatic vessel segmentation with real-world noisy labels and retinal vessel segmentation with real-world machine-generated noisy labels. The results demonstrate the superior segmentation performance of our approach as well as its effectiveness on label denoising.

This work substantially extends our preliminary work [1] on hepatic vessel segmentation at MICCAI'21: (i) we, for the first time, clearly reveal and define two pervasive yet distinct clinical scenarios on learning with noisy labels, i.e., Set-HQ-agnostic and Set-HQ-knowable (our focus), with interesting findings and more insightful discussion; (ii) performing appropriate label refinement is crucial for exploiting informative knowledge in LQ labels. Yet, less attention is paid to discussing the refinement process in [1]. Observing different tasks suffer various inherent challenges, we further generalize original MTCL with a comprehensive family of label refinement strategies from perspectives of human empiricism and model-driven calibration for various clinical practices; (iii) we conduct more comprehensive and insightful experiments on simulated left atrium data, real-world hepatic and retinal vessel data under two distinct clinical scenarios, and the potential of leveraging tiny HQ labeled data to explicitly improve the label quality of other datasets is quantitatively demonstrated.

II. Related Work

A. Label-efficient Medical Image Segmentation

The success of DL-based medical image segmentation methods usually relies on the massive radiologist-examined labeled data. However, labeling volumetric scans is extremely laborious and expertise-demanding. Such predicament motivates many researches on label-efficient learning, including weakly supervised learning [9], self-supervised learning [10], [11] and semi-supervised learning (SSL) [12]–[15]. Specifically, weakly supervised learning aims to learn segmentation from simple annotations, e.g., image-level labels [16] and bounding boxes [17]. Self-supervised learning approaches usually utilize

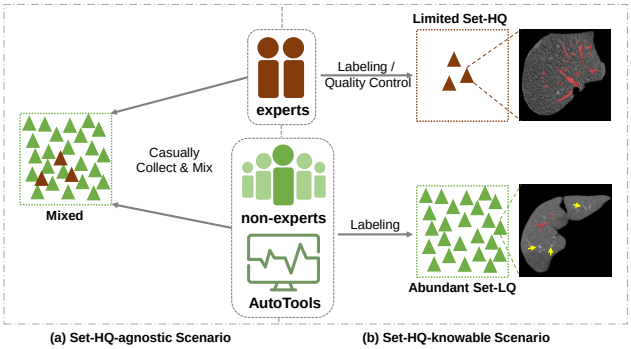


Fig. 1. Illustration of two practical scenarios with real-world examples of labeling hepatic vessels: (a) Set-HQ-agnostic: casually collect and mix the so-called "labeled data" without any quality examination; (b) Set-HQ-knowable: invite experts to involve in majority-voted labeling or extra quality control for a small number of images, so that Set-HQ and Set-LQ can be separated, which is our focused setting in this work.

unlabeled data to train the networks in a supervised-like manner via pretext tasks such as solving a jigsaw puzzle [10] or reconstructing corrupted images [11]. Recent progress leans semi-supervised learning since it is feasible to acquire a small set of labeled data and abundant unlabeled data in the clinical scenario, wherein the consistency regularization methods [12], [13], [15] attract more attention in this community. Apart from the supervised loss for HQ labeled data, existing methods exploit the unlabeled data by encouraging the unsupervised consistency under different perturbations [12]–[14] or the supervised-like consistency [15] via prototypical network. In our work, the backbone shares a similar design to the mean-teacher (MT) model [13], which is a well-known SSL framework. However, beyond the unlabeled data, it is also feasible to obtain annotations for those abundant unlabeled data via crowdsourcing from non-experts or using off-the-shelf tools without laborious quality control in practice. Although these labels are contaminated with heavy noises and may become the encumbrance for typical network training (as shown in Sec. IV), we notice that they can still provide rewarding supervision if treated appropriately. Thus, assisted by the MT-like design, we further tailor a synergistic framework that can simultaneously and robustly exploit the dark knowledge in the images of Set-LQ and their casually collected LQ noisy labels.

B. Learning Segmentation with Noisy Labels

Considerable label noises in Set-LQ can mislead the network training and undesirably offset the efficacy of those HQ labels [1], [2]. Most existing works of noisy supervised learning focus on image-level classification tasks [8], [18], [19], while the more challenging pixel-wise segmentation task is still underexplored. Here, following Sec. I, we categorize existing approaches for the segmentation task into two-folds: (i) Set-HQ-agnostic (Fig. 1 (a)), i.e., simply mixing all the collected data: TriNet [6] designs a tri-network and uses integrated prediction from two networks for the third network training to alleviate the misleading problem. Zhang et al. [5] proposed a two-stage method that utilizes a mixed dataset to pre-train the network and then refine the labels via confidence

estimation to train another network. PNL [4] introduces an image-level label quality evaluation module that selects the good ones to tune the network. (ii) Set-HQ-knowable (Fig. 1 (b)), e.g., inviting experts to label or perform quality control for a small set of data and thus making Set-HQ and Set-LQ separated: Luo et al. [2] proposed to use two decoupled decoders, where one is for Set-HQ and the other is for Set-LQ, to mitigate the negative effects brought by those low-quality labels. KDEM [3] extends [2] by further introducing knowledge distillation and entropy minimization regularization. The two methods share the same spirit that implicitly decouples the learning processes for Set-HQ and Set-LQ, however, explicitly characterizing the locations of those label errors can be more valuable because it can provide the spatial information for label denoising and make approaches much easier to control. Besides, only evaluated on simulated data, the effectiveness of these methods on real-world noisy labels has not been investigated. In this work, we focus on the Set-HQ-knowable scenario. Even so, extensive experiments and discussions are performed on both simulated and real-world data under the two distinct practical settings.

III. Methodology

A. Framework Overview

To ease the description of our methodology, we formulate the segmentation problem under the Set-HQ-knowable scenario setting. In this task, the entire training set contains N samples. However, only M samples have the expert-examined high-quality labels (termed as Set-HQ), while the remaining $N-M$ samples, termed as Set-LQ, are labeled by non-experts or other automatic segmentation tools without any quality control by experts. We denote the Set-HQ as $\mathcal{S}_h = \left\{\left(\mathbf{X}_{(i)}, \mathbf{Y}_{h(i)}\right)\right\}_{i=1}^{M}$ and the Set-LQ as $\mathcal{S}_l = \left\{\left(\mathbf{X}_{(i)}, \mathbf{Y}_{l(i)}\right)\right\}_{i=M+1}^{N}$, where $\mathbf{X}_{(i)} \in \mathbb{R}^{\Omega_i}$ represents the input images and $\mathbf{Y}_{h(i)}, \mathbf{Y}_{l(i)} \in \{0,1\}^{\Omega_i}$ (we focus on binary segmentation task) denotes the given high-quality and low-quality segmentation label of $\mathbf{X}_{(i)}$, respectively.

Fig. 2 depicts the proposed Mean-Teacher-assisted Confident Learning (MTCL) framework, which aims to simultaneously learn segmentation from limited Set-HQ and abundant Set-LQ. Briefly, the MTCL framework consists of a student model and a weight-averaged teacher model, accompanied by a label self-denoising process for LQ labeled data. In such synergistic design, apart from benefiting from the high-quality supervision from those expert-examined labels, our MTCL can further (i) exploit the dark knowledge inside the image-only data of Set-LQ with the help of consistency regularization; (ii) receive rewarding guidance from those LQ noisy labels via explicit class-conditional label noise characterization and subsequent label denoising process. The details of these synergistic processes are elaborated successively in the following sections.

B. Exploit Dark Knowledge inside the Images of Set-LQ

Since the LQ labels of Set-LQ could be the encumbrance for model training, it is natural to follow the typical SSL setting that casts LQ labels away and exploit the dark knowledge inside the image-only data firstly. Here, we follow this intuition and adopt the mean-teacher (MT) model [13], a top-performing SSL framework, as our basic architecture. The basic MT framework consists of a student model (updated by back-propagation) and a parallel weight-averaged teacher model (updated by the weights of the student model in different training stages). We choose the MT-like architecture design for three reasons: (i) instead of simply mixing Set-HQ and Set-LQ, the MT design allows the network to distinguish the two different sub-sets under the Set-HQ-knowable setting as shown in Fig. 2. HQ labels can provide prime guidance to reliably train the student model, while such HQ knowledge can be transferred to exploit images of Set-LQ with the teacher model; (ii) it has superior ability to exploit the dark knowledge in the image-only data via perturbation-based consistency regularization; (iii) the self-ensembling teacher model can serve as a "third party" that continuously provides out-of-sample predicted probabilities to the following class-conditional label noise characterization process during training, as elaborated in Sec. III-C.

Formally, denoting the weights of the student model at training step t as θ_t, we update the teacher model's weights $\tilde{\theta}_t$ using exponential moving average (EMA) strategy, which can be formulated as:

$$\tilde{\theta}_t = \alpha\tilde{\theta}_{t-1} + (1-\alpha)\theta_t, \tag{1}$$

where α is the EMA decay rate and set to 0.99 as recommended by [13]. Based on the smoothness assumption [20], when the same input is fed into both models, we encourage the teacher model's temporal ensemble prediction to be consistent with that of the student model under different perturbations (e.g., adding random Gaussian noise ξ to the input images). As shown in Fig. 2, besides the typical supervised loss $\mathcal{L}_{hs}$ derived from the HQ labels, this process further contributes to an unsupervised consistency loss $\mathcal{L}_c$ that measures the dissimilarity between the predicted probabilities of the teacher model and that of the student model.

C. LQ Label Self-denoising Towards Effective Guidance

Besides the image-only information, further robustly leveraging the concomitant noisy labels of Set-LQ is the key to superior performance. To achieve it, we further propose a synergistic LQ label self-denoising process to alleviate the potential misleading guidance caused by label noises.

Inspired by the arbitration based labeling procedure where a third party, e.g., the experts, is consulted for disputed labeling cases, we resort to mimicking such quality control process to identify the label noises. As a start, we take 2D setting as example and introduce the class-conditional classification noise process (CNP) [7] assumption: (1) besides the given (observed) LQ noisy label $y_{l(w,h)}$, every pixel $x_{(w,h)}$ in the image $\mathbf{X} \in \mathbb{R}^{W\times H}$ from Set-LQ exists a true (latent) label $y^*_{l(w,h)}$; (2) every label in class $j \in \mathcal{C}_m$ may be independently mislabeled as class $i \in \mathcal{C}_m$ with probability $p\left(y_l = i \mid y_l^* = j\right)$ ($\mathcal{C}_m$ indicates the set of m class label).

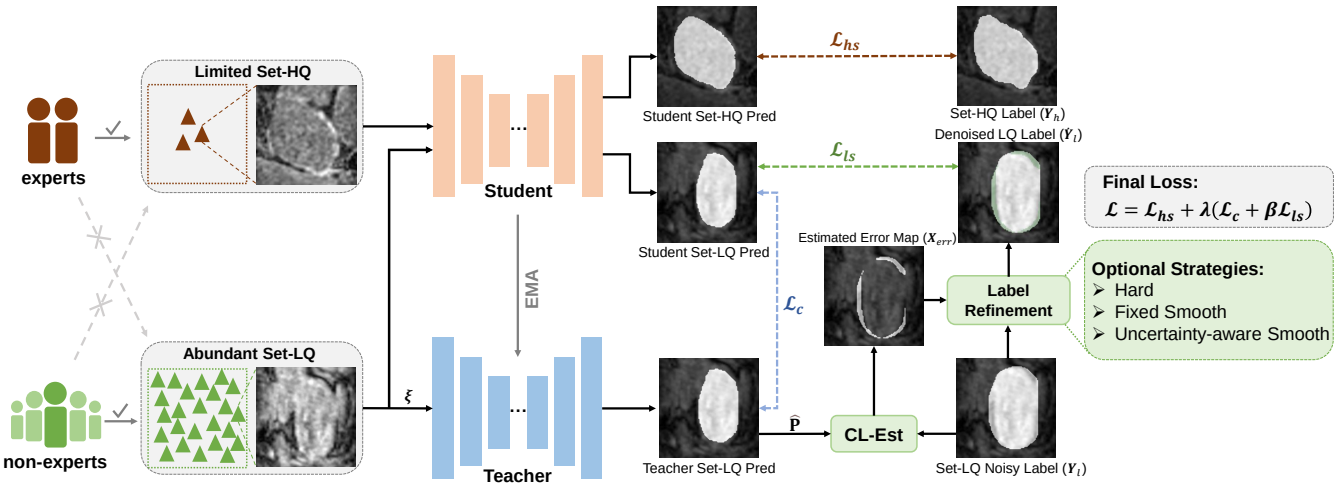


Fig. 2. Illustration of the proposed Mean-Teacher-assisted Confident Learning (MTCL) framework. The student model is updated via gradient back-propagation, while the teacher model is updated as the exponential moving average (EMA) of student weights. CL-Est denotes the Confident Learning-based label noise Estimation module. The total loss is a weighted combination of the supervised loss $\mathcal{L}_{hs}$ on Set-HQ, the perturbation-based consistency loss $\mathcal{L}_c$ on Set-LQ and the calibrated supervised loss $\mathcal{L}_{ls}$ on Set-LQ.

1) Characterize the pixel-wise class-conditional label errors: Based on CNP assumption, an assembling work [8] in image-level classification tasks, termed as confident learning (CL), further reveals that confident joint matrix [21] is effective in partitioning and counting label errors, while its normalized statistics, i.e., joint distribution matrix, can robustly characterize aleatoric uncertainty from latent label noises. Inspired by it, a CL-based pixel-level label noise estimation module (i.e., CL-Est in Fig. 2) is subtly tailored on top of the mean-teacher architecture to characterize the error map that can indicate the specific pixel-level location (i.e., $(w,h)^{th}$) of each mislabeled pixel. Here, the teacher model serves as the "third party" to provide out-of-sample predicted probabilities $\hat{\mathbf{P}}$. Note that we use the clean input for $\hat{\mathbf{P}}$ instead of the aforementioned perturbed one. Ideally, such a third party is collaboratively enhanced during training. Then, we denote the predicted probability that a pixel x belongs to its given label $y_l = i$ from the teacher model as its self-confidence, formulated as $\hat{p}_i(x)$. Low self-confidence can be a heuristic likelihood of being a mislabeled pixel. If x with label $y_l = i$ has *large enough* $\hat{p}_j(x) \geq t_j$, the true (latent) label y_l^* of x can be suspected to be j instead of i. Here, the per-class threshold t_j is obtained empirically by calculating the average (expected) self-confidence $\hat{p}_j(x)$ of the pixels labeled with $y_l = j$, formulated as $t_j = \frac{1}{|\mathbf{X}_{y_l=j}|}\sum_{x\in\mathbf{X}_{y_l=j}} \hat{p}_j(x)$. Then, with such threshold strategy, the confident joint matrix $\mathbf{C}_{y_l,y_l^*}$ for each image $\mathbf{X}$ can be defined as:

$$\mathbf{C}_{y_l,y_l^*}[i][j] := \left|\hat{\mathbf{X}}_{y_l=i,y_l^*=j}\right|, \text{ where}$$
$$\hat{\mathbf{X}}_{y_l=i,y_l^*=j} := \left\{ x \in \mathbf{X}_{y_l=i} : \hat{p}_j(x) \geq t_j, j = \underset{c\in\mathcal{C}_m:\hat{p}_c(x)\geq t_c}{\arg\max}\ \hat{p}_c(x) \right\}. \quad (2)$$

Compared to the naive confusion matrix [22], such confident joint matrix enables more robustness to class-imbalance and overconfident predicted probabilities [8]. Then, we calibrate $\mathbf{C}_{y_l,y_l^*}$ as:

$$\tilde{\mathbf{C}}_{y_l,y_l^*}[i][j] = \frac{\mathbf{C}_{y_l,y_l^*}[i][j]}{\sum_{j\in\mathcal{C}_m}\mathbf{C}_{y_l,y_l^*}[i][j]} \cdot |\mathbf{X}_{y_l=i}|\,. \quad (3)$$

By further normalizing the calibrated confident joint matrix $\tilde{\mathbf{C}}_{y_l,y_l^*}$, we can estimate the joint distribution matrix $\hat{\mathbf{Q}}_{y_l,y_l^*}$:

$$\hat{\mathbf{Q}}_{y_l,y_l^*}[i][j] = \frac{\tilde{\mathbf{C}}_{y_l,y_l^*}[i][j]}{\sum_{i\in\mathcal{C}_m,j\in\mathcal{C}_m}\tilde{\mathbf{C}}_{y_l,y_l^*}[i][j]}. \quad (4)$$

Then, we adopt the prune by class (PBC) [8] strategy to identify the label noises. For each class $i \in \mathcal{C}_m$, PBC selects the $n\cdot\sum_{j\in\mathcal{C}_m:j\neq i}\left(\hat{\mathbf{Q}}_{y_l,y_l^*}[i][j]\right)$ pixels with the lowest self-confidence $\hat{p}_i(x)$ (n denotes the number of pixels in image $\mathbf{X}$) as the mislabeled ones, thereby obtaining the binary estimated error map $\mathbf{X}_{err}$, where "1" denotes that this pixel is identified as a mislabeled one and vice versa. Note that the CL-Est module is computationally efficient and model-agnostic (only working on the predicted probabilities $\hat{\mathbf{P}}$). No trainable parameters are additionally introduced. Such pixel-level error map $\mathbf{X}_{err}$ can explicitly guide the subsequent label refinement process.

2) Label refinement towards rewarding supervision: Performing appropriate label refinement is crucial for exploiting informative knowledge in LQ labels. Due to the diversity of medical images, different segmentation tasks suffer various inherent difficulties, e.g., ambiguous boundaries, intrinsic noises in images and complex organ morphology, which can affect the accuracy of the estimated error map. Here, distinct from our preliminary version [1] with limited focus on refinement process, we generalize original MTCL with a family of label refinement strategies from perspectives of human empiricism and model-driven calibration, including hard refinement, fixed smooth refinement (FS), and uncertainty-aware dynamic smooth refinement (UDS), to suit various clinical practices:

- **Hard Refinement**: The option #1, i.e., hard refinement, indicates that we highly trust the accuracy of estimated

error map $\mathbf{X}_{err}$, and therefore impose hard refinement on the given noisy masks $\mathbf{Y}_l$. This strategy may suit tasks with simple segmentation targets. We denote $\dot{\mathbf{Y}}_l$ as the denoised LQ label. The hard refinement operation is formulated as:

$$\dot{\mathbf{Y}}_l = \mathbf{Y}_l + \mathbf{X}_{err} \cdot (-1)^{\mathbf{Y}_l}. \tag{5}$$

- **Fixed Smooth Refinement (FS)**: Although the CL-based label noise estimation can be effective in finding label errors, perfect disambiguation of model errors (epistemic uncertainty) from intrinsic label noises (aleatoric uncertainty) can hardly be achieved. Therefore, we provide an option #2 that imposes fixed smooth refinement on the noisy masks of Set-LQ, which is formulated as:

$$\dot{\mathbf{Y}}_l = \mathbf{Y}_l + \mathbf{X}_{err} \cdot (-1)^{\mathbf{Y}_l} \cdot \tau, \tag{6}$$

where $\tau \in [0, 1]$ is the fixed smooth factor, which is empirically set as 0.8.

- **Uncertainty-aware Dynamic Smooth Refinement (UDS)**: Rather than human empiricism, we further introduce adaptive model-driven calibration for very challenging tasks, e.g., segmenting hepatic vessels. Intuitively, for each pixel, if the "third party", i.e., the teacher model, yields more certain prediction, the smooth factor could be raised up to close to 1, and vice versa. Therefore, we further provide option #3 by introducing epistemic uncertainty of the teacher model to dynamically adjust the smooth factor for each pixel. Specifically, we estimate the uncertainty via Monte Carlo dropout [23]. Firstly, T stochastic forward passes are performed on the teacher model under random dropout and Gaussian noise perturbation injected into the input. Thereby, a set of softmax probability vector $\{\mathbf{p}_t\}_{t=1}^{T}$ can be obtained, and then we adopt the predictive entropy to estimate the uncertainty for each pixel, as calculated by:

$$u = -\sum_c (\frac{1}{T}\sum_t \mathbf{p}_t^c) \log(\frac{1}{T}\sum_t \mathbf{p}_t^c), \tag{7}$$

where $\mathbf{p}_t^c$ denotes the t-th predicted probability of the c-th class. Note that u is the pixel-wise uncertainty while the whole image uncertainty is denoted as $\mathbf{U}$. Since the predictive entropy has a fixed range [23], divided by the maximum value $log(2)$, we can further obtain the normalized uncertainty map $\hat{\mathbf{U}}$. Intuitively, an uncertain pixel should be allocated a small weight, therefore the uncertainty-aware dynamic smooth refinement can be formulated as:

$$\dot{\mathbf{Y}}_l = \mathbf{Y}_l + \mathbf{X}_{err} \cdot (-1)^{\mathbf{Y}_l} \cdot (1 - \hat{\mathbf{U}}). \tag{8}$$

Note that the specific choice of refinement strategy depends on the specific application, no priority is given to a specific strategy. For example, if the estimated error map $\mathbf{X}_{err}$ can accurately characterize the label errors, choosing the simple hard refinement may provide more productive information. Particularly, the choice of the strategy is further discussed in Sec. V-A. With the help of the label refinement process, the negative effects brought by label noises can be eliminated to some extent, and thereby the LQ labels can provide more rewarding guidance (termed as $\mathcal{L}_{ls}$ as shown in Fig. 2) to optimize the student model.

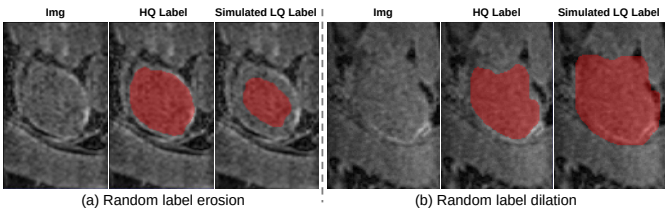


Fig. 3. Visualization of the simulated low-quality noisy labels for the LA dataset [26]: (a) imposing random erosion on the labels, and (b) imposing random dilation on the labels.

D. Final Loss Function

The total loss is a weighted combination of the supervised loss $\mathcal{L}_{hs}$ on Set-HQ, the perturbation-based consistency loss $\mathcal{L}_c$ on Set-LQ and the calibrated supervised loss $\mathcal{L}_{ls}$ on Set-LQ, calculated by:

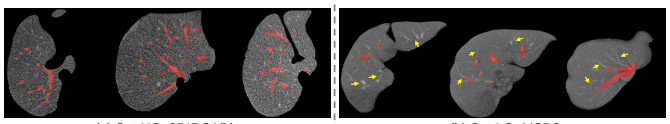
$$\mathcal{L} = \mathcal{L}_{hs} + \lambda(\mathcal{L}_c + \beta\mathcal{L}_{ls}), \tag{9}$$

where $\mathcal{L}_c$ is calculated by the pixel-wise mean squared error (MSE); β is empirically set to 5; λ is a ramp-up trade-off weight commonly scheduled by the time-dependent Gaussian function [24] $\lambda(t) = w_{max} \cdot e^{\left(-5\left(1-\frac{t}{t_{max}}\right)^2\right)}$, where w_{max} is the maximum weight commonly set as 0.1 [25] and $t_{\max}$ is the maximum training iteration. Such weighting schedule for λ can avoid the domination by misleading targets at the beginning of network training. The specific supervised losses for different tasks are elaborated in Sec. IV.

IV. Experiments

A. Datasets and Experimental Setup

1) 3D Left Atrium Segmentation Dataset with Simulated Noisy Label: The left atrium (LA) segmentation dataset [26] contains 100 3D gadolinium-enhanced magnetic resonance images (GE-MRIs) with expert-examined labels. The images have the isotropic resolution of $0.625 \times 0.625 \times 0.625\ mm^3$. Following the same data division and preprocessing procedure in [25], 80 samples are used for training and the remaining 20 samples are used for testing. All the scans are cropped to the center of heart region and the intensities are normalized to zero mean and unit variance. To simulate the practical scenario, three settings with different portions of samples annotated are investigated. Specifically, we randomly select 10%, 20% and 30% training samples as HQ labeled data, and those remaining samples undergo random erosion and dilation with 3-15 pixels as Set-LQ, as exampled in Fig. 3. For example, the 10% setting simulates the scenario where we collect 80 training samples and invite experts to spend reasonable time to perform HQ labeling for 8 cases, while the remaining 72 (90%) cases are labeled by non-experts without any quality control (simulated by intentionally introducing label noise for this specific dataset).



(a) Set-HQ: 3DIRCADb (b) Set-LQ: MSD8

Fig. 4. Visualization of real-world hepatic vessel segmentation datasets: (a) 3DIRCADb dataset [27] with high-quality annotations, and (b) MSD8 dataset [28] with numerous mislabeled and unlabeled pixels. Red represents the labeled vessels, while the yellow arrows in (b) point at some outrageous unlabeled pixels.

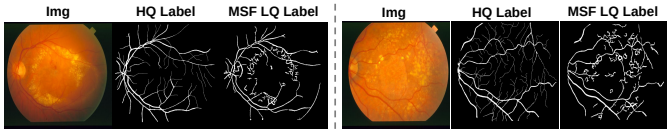


Fig. 5. Examples of the STARE retinal vessel segmentation dataset [29]. HQ labels are provided by an expert. Their matched spatial filter (MSF) algorithm [29] based segmentation results can serve as our LQ labels (i.e., AutoTool-generated labels as depicted in Fig. 1).

2) Real-world Hepatic Vessel Segmentation Datasets: Despite great effort to simulate the real-world scenario, the simulated noisy labels, which most existing studies use, may not be representative enough for the varying quality of labels in reality. Therefore, further contributing a real-world benchmark is also crucial in this field. Meanwhile, we observe that due to low contrast and complex morphology of hepatic vessels, manually segmenting those vessels from computed tomography (CT) is far more error-prone, expertise-demanding and laborious than other structures, e.g., the whole liver, which despondently results in the extreme difficulty of obtaining HQ labels of the hepatic vessel. Reviewing the publicly available datasets, we found two hepatic vessel segmentation datasets can serve as the real-world practice: (i) **3DIRCADb** dataset [27], maintained by the French Institute of Digestive Cancer Treatment, can serve as **Set-HQ**. It contains only 20 abdominal CT scans but with high-quality hepatic vessel annotation, as exampled in Fig. 4 (a). In this dataset, different images share the same axial slice size (512×512 pixels) while the pixel spacing varies from 0.57 to 0.87 mm, the slice thickness varies from 1 to 4 mm, and the slice number is between 74 and 260; (ii) **MSD8** dataset [28], which was collected from Memorial Sloan Kettering Cancer Center, consists of 443 abdominal CT scans. The slice thickness varies from 2.5 to 5 mm. Although this dataset provides the hepatic vessel annotations, the labels are semi-automatically obtained via the Scout application [30], wherein the label quality is obviously worse than that of the 3DIRCADb dataset, as exampled in Fig. 4 (b). Thus, we regard MSD8 as a real-world **Set-LQ**. According to the statistics in [31], around 65.5% of the vessel pixels are unlabeled and approximately 8.5% non-vessel pixels are mislabeled as vessels in the MSD8 dataset, resulting in the necessity of laborious manual refinement in previous work [31] to avoid the biased testing. In our experiments, for the 3DIRCADb dataset, we randomly split 10 cases for testing, and further select 10 or 5 cases from the remaining samples as two HQ labeled settings for training. All the samples in MSD8 (Set-LQ) are used for training since their low-quality labels are not appropriate for unbiased evaluation.

3) Real-world Retinal Vessel Segmentation Datasets: Besides the above hepatic vessel segmentation datasets, we further notice that the **STARE** dataset [29] for the challenging retinal vessel segmentation task can also serve as another real-world practice. Specifically, the STARE dataset contains 20 color retinal fundus images captured at a field of view of 35 degrees with the size of 700×605 pixels. Two careful manual annotations for each image are provided by STARE, wherein we choose the one from expert Valentina Kouznetsova as our HQ labels, as shown in Fig. 5. Besides, STARE also provides the third annotation generated by their matched spatial filter (MSF) algorithm [29], as shown in Fig. 5. However, since the MSF-based labels are obviously noisy, few previous efforts have been made to utilize them but it suits our scenario (i.e., AutoTool-generated labels as depicted in Fig. 1). Thus, we regard them as LQ labels. We randomly select 2 and 4 samples with HQ labels as two Set-HQ settings. Due to limited training data, all the remaining images are used for training and regarded as Set-LQ with the MSF-based LQ labels. Then, we perform cross-database testing on the commonly used DRIVE-test [32] dataset. This test set contains 20 color fundus images captured at a field of view of 45 degrees (each with a size of 565×584 pixels) with two HQ expert-examined labels, wherein the labels from their first expert are often utilized as ground truths [33]. All the images are converted into grayscale to alleviate the interference of hue and saturation, followed by commonly used preprocessing strategies [33] including contrast limited adaptive histogram equalization (CLAHE) [34] and gamma correction.

4) Baseline Approaches: Despite limited works on noisy supervised segmentation, especially under the Set-HQ-knowable scenario, we seek to include as many baselines under different scenarios as possible for comparison and provide insights to future research in this field. Specifically, the baselines can be categorized as follows:

- Fully supervised baselines: (i) **H-Sup**: only using limited Set-HQ to train the backbone network; (ii) **HL-Sup**: mixing both Set-HQ and Set-LQ to train the network.
- Set-HQ-agnostic setting: (i) **TriNet** [6]: a tri-network based noise-tolerant method extended from co-teaching [35] strategy; (ii) **2SRnT** [5]: a two-stage method that utilizes mixed dataset to pre-train a network and then employs a similar error identification strategy to refine the labels to train another network; (iii) **PNL** [4]: an image-level quality evaluation based noise-tolerant strategy.
- Set-HQ-knowable setting (i.e., our focus): (i) **Decoupled** [2]: using two decoupled decoders, one for Set-HQ and the other for Set-LQ, to mitigate the negative effects brought by LQ labels; (ii) **KDEM** [3]: an extended work of [2] that further introduces HQ knowledge distillation and additional entropy minimization regularization. Note that both Decoupled [2] and KDEM [2] introduce more trainable parameters because of the additional decoder.

All the baseline methods share the same backbone and data partition protocols to ensure fairness.

5) Implementation and Evaluation Metrics: The framework is implemented in Python with PyTorch, using an NVIDIA

GeForce RTX 3090 GPU with 24GB memory. The network is trained using the SGD optimizer (weight decay=0.0001, momentum=0.9). Standard data augmentation, including randomly cropping, flipping and rotating, is also applied. The learning rate is initialized as 0.01 and decayed with a power of 0.9 after each iteration. Considering the training efficiency, T is set to 6 for uncertainty estimation. Besides, only the student model will be utilized at the inference stage, which can ensure the computational efficiency. Other task-related details are elaborated later. For LA segmentation and hepatic vessel segmentation, we adopt four well-known metrics for a comprehensive evaluation, including Dice score, Jaccard, average surface distance (ASD) and 95% Hausdorff distance (95HD). Following previous evaluation in retinal vessel segmentation [33], we adopt four commonly used metrics, including Dice score, area under receiver operation characteristic curve (AUC), sensitivity (SEN) and accuracy (ACC). The implementation is publicly available at `https://github.com/lemoshu/MTCL`.

B. Experiments on Simulated LA Dataset

1) Task-related Implementation Details: For 3D LA segmentation, following the setting in [25], we adopt the same 3D V-Net [36] as the backbone. During training, we randomly crop patches of $112 \times 112 \times 80$ voxels as the network input. The batch size is set to 4, including 2 HQ labeled images and 2 LQ labeled images in each mini-batch. Both $\mathcal{L}_{hs}$ and $\mathcal{L}_{ls}$ use a combination of cross-entropy loss and Dice loss with equal weights of 0.5. The maximum training iteration is set to 8,000. For a fair comparison, no extra post-processing is utilized. We use a sliding window strategy with a stride of $18 \times 18 \times 4$ voxels for the inference stage.

2) Comparison Study: Table I presents the quantitative comparison results under 10%, 20% and 30% HQ labeled data settings. Firstly, under the typical supervised setting (i.e., H-Sup and HL-Sup), when trained with 10% HQ labeled data, the network performs poorly and can even further benefit from additional Set-LQ despite their noisy labeling. However, the severe performance degradation can be found under 20% and 30% HQ labeled data settings, which reveals that the additional LQ labels mislead the training process. As for the methods under Set-HQ-agnostic setting, the three baselines, i.e., TriNet, 2SRnT and PNL, can effectively alleviate the negative effects brought by those agnostic noisy labels. When it turns to the same scenario as ours, i.e., Set-HQ-knowable, unstable performances appear in Decoupled [2] and KDEM [3]. As mentioned in Sec. II-B, such implicit decoupling strategy is hard-to-control. For example, KDEM performs poorly in leveraging the LQ labeled data under 10% HQ labeled data setting, yet achieving very competitive results when we increase the proportion of Set-HQ to 20% or 30%. Overall, the proposed MTCL with three different refinement strategies substantially surpass the state-of-the-art methods. Especially, in this LA segmentation task, using hard refinement achieves the best results, implying that the estimated error map can satisfactorily characterize the location of label noises so that the network is prone to make full use of these informative denoised labels. Thus, we adopt hard refinement in the following LA segmentation experiments. Fig. 6 presents the LA segmentation results of our MTCL and other approaches under 10% HQ labeled data setting. Consistently, the predicted mask of MTCL fits more accurately with the ground truth, which further demonstrates the effectiveness of our method.

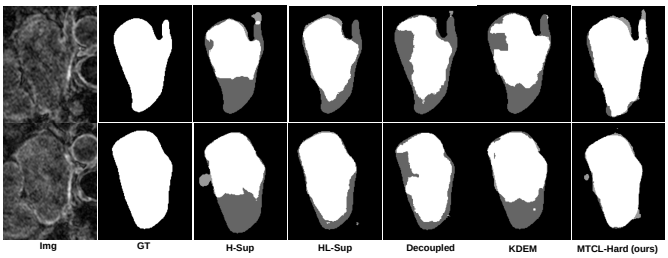


Fig. 6. Exemplar LA segmentation results of the proposed MTCL and other approaches under 10% HQ labeled data setting. Grey color represents the inconsistency between the predicted result and the ground truth.

3) Analytical Ablation Study: To verify the effectiveness of each component, we propose different variants to perform an ablation study under different HQ labeled data settings: (i) **MT** [13]: a typical MT model that additionally uses the image-only information of Set-LQ, i.e., $\mathcal{L} = \mathcal{L}_{hs} + \lambda\mathcal{L}_c$; (ii) **MTNL**: extended MT by further introducing the original noisy labels (NL) of Set-LQ into supervised loss; (iii) **MTCL w/o** $\mathcal{L}_c$: ignoring additional dark knowledge inside the image-only information of Set-LQ, i.e., $\mathcal{L} = \mathcal{L}_{hs} + \lambda\mathcal{L}_{ls}$.

As shown in Table II, perturbation-based consistency learning can effectively exploit the image-only information of Set-LQ. Interestingly, when additionally introducing those LQ labels of Set-LQ into the supervised process, such perturbation-tolerant SSL training scheme has potential in alleviating the performance degradation caused by the label noises, particularly under the scarce Set-HQ setting (MTNL vs. MT under 10% HQ labeled data setting). It can also be observed that, without exploiting the image-only information of Set-LQ (i.e., MTCL w/o $\mathcal{L}_c$), our method is also able to achieve respectable performance gains, revealing that the synergistic LQ label denoising process contributes more in our framework. Nonetheless, superior segmentation can be achieved by simultaneously exploiting the image-only information and the denoised LQ labels of Set-LQ.

4) Impact of weight β: As shown in our loss function (Eqn. 9), a fixed factor β is utilized to control the trade-off between the consistency loss $\mathcal{L}_c$ and the calibrated supervised loss $\mathcal{L}_{ls}$ on Set-LQ. Here, we investigate the impact of different β. As quantitatively shown in Table III, it can be observed that the proposed MTCL is not particularly sensitive to β ($p > 0.05$), except for the 95HD metric. Overall, we set β to 5 since it performs optimally in terms of most mean metrics.

5) Extended Comparison with the Set-HQ-agnostic Scenario: The proposed MTCL framework is tailored for the Set-HQ-knowable scenario. To further investigate the efficacy of this separated strategy, we simultaneously feed the student and the teacher models with the mixed-quality data, i.e., mixing Set-HQ and Set-LQ. As shown in Table IV, such Set-HQ-agnostic

TABLE I

QUANTITATIVE COMPARISON STUDY ON THE LEFT ATRIUM SEGMENTATION TASK [26]. $*$ INDICATES $p \leq 0.05$ FROM A TWO-SIDED PAIRED T-TEST WHEN COMPARING RESULTS OF OUR BEST MODEL WITH THE BEST-PERFORMING METHOD UNDER THE SET-HQ-KNOWABLE SETTING. THE BEST MEAN RESULTS ARE IN BOLD.

Methods	Settings			Metrics			
	Set-HQ	Set-LQ	Set-HQ-knowable?	Dice [%] ↑	Jaccard [%] ↑	95HD [mm] ↓	ASD [mm] ↓
H-Sup	10%	0%	-	79.99	68.12	21.11	5.48
HL-Sup	10%	90%	-	80.55	67.78	12.94	3.86
TriNet [6]	10%	90%	×	85.02	73.95	13.96	2.68
2SRnT [5]	10%	90%	×	84.62	73.88	14.84	2.53
PNL [4]	10%	90%	×	84.97	72.53	14.09	2.50
Decoupled [2]	10%	90%	✓	85.67	75.13	15.81	2.51
KDEM [3]	10%	90%	✓	80.20	69.51	14.59	5.41
MTCL-Hard (ours)	10%	90%	✓	**88.34***	**79.27***	10.85*	**2.33**
MTCL-FS (ours)	10%	90%	✓	86.53	76.71	11.01	2.42
MTCL-UDS (ours)	10%	90%	✓	87.41	77.93	**9.53**	2.49
H-Sup	20%	0%	-	86.03	76.06	14.26	3.51
HL-Sup	20%	80%	-	81.32	68.85	16.21	4.14
TriNet [6]	20%	80%	×	86.04	76.87	14.08	2.69
2SRnT [5]	20%	80%	×	87.05	77.34	13.32	2.61
PNL [4]	20%	80%	×	86.98	77.42	13.83	2.77
Decoupled [2]	20%	80%	✓	84.66	73.92	12.12	3.37
KDEM [3]	20%	80%	✓	88.31	79.47	**8.32**	2.69
MTCL-Hard (ours)	20%	80%	✓	**89.25***	**80.72***	10.15	2.24*
MTCL-FS (ours)	20%	80%	✓	87.44	77.88	10.46	**2.13**
MTCL-UDS (ours)	20%	80%	✓	88.89	80.13	10.30	2.21
H-Sup	30%	0%	-	87.08	77.45	13.23	2.30
HL-Sup	30%	70%	-	84.04	72.74	11.54	3.38
TriNet [6]	30%	70%	×	86.07	76.24	9.34	2.73
2SRnT [5]	30%	70%	×	88.21	78.22	8.74	2.26
PNL [4]	30%	70%	×	86.28	77.51	9.11	2.24
Decoupled [2]	30%	70%	✓	86.46	76.33	9.72	2.80
KDEM [3]	30%	70%	✓	88.59	79.76	9.61	2.33
MTCL-Hard (ours)	30%	70%	✓	**89.98***	**81.89***	**6.35***	1.89*
MTCL-FS (ours)	30%	70%	✓	89.74	81.48	7.49	1.75
MTCL-UDS (ours)	30%	70%	✓	89.76	81.41	9.87	**1.69**
H-Sup (upper bound)	100%	0%	-	91.14	83.82	5.75	1.52

TABLE II

ABLATION STUDY ON LEFT ATRIUM SEGMENTATION UNDER VARYING HQ LABELED DATA SETTINGS. NOTE THAT WE ADOPT MTCL-HARD IN THIS TASK FOR ITS SUPERIOR PERFORMANCE. $*$ INDICATES SIGNIFICANT DIFFERENCE ($p \leq 0.05$) FROM OUR COMPLETE VERSION (MTCL (OURS)). BEST RESULTS ARE IN BOLD.

Methods	Settings	Metrics			
	HQ / LQ	Dice [%] ↑	Jaccard [%] ↑	95HD [mm] ↓	ASD [mm] ↓
MT [13]	10% / 90%	84.24*	73.26*	19.41*	2.71
MTNL	10% / 90%	86.12*	75.63*	10.68	2.66
MTCL w/o $\mathcal{L}_c$	10% / 90%	87.92	78.67	**8.43***	2.37
MTCL (ours)	10% / 90%	**88.34**	**79.27**	10.85	**2.33**
MT [13]	20% / 80%	88.42*	79.45*	13.07*	2.73
MTNL	20% / 80%	86.37*	76.25*	11.66*	2.76*
MTCL w/o $\mathcal{L}_c$	20% / 80%	89.02	80.39	10.21	2.33
MTCL (ours)	20% / 80%	**89.25**	**80.72**	**10.15**	**2.24**
MT [13]	30% / 70%	89.10	80.45*	8.80*	2.32
MTNL	30% / 70%	86.48*	76.49*	12.96*	2.81*
MTCL w/o $\mathcal{L}_c$	30% / 70%	89.17	80.55*	10.91*	**1.73**
MTCL (ours)	30% / 70%	**89.98**	**81.89**	**6.35**	1.89

TABLE III

RESULTS OF 3D LEFT ATRIUM SEGMENTATION WITH DIFFERENT β UNDER 10% HQ LABELED DATA SETTING. NOTE THAT WE ADOPT MTCL-HARD IN THIS TASK FOR ITS SUPERIOR PERFORMANCE. BEST RESULTS ARE IN BOLD.

β	Metrics			
	Dice [%] ↑	Jaccard [%] ↑	95HD [mm] ↓	ASD [mm] ↓
1	88.01	78.71	13.53	2.42
3	87.94	78.70	12.11	2.35
5	**88.34**	**79.27**	10.85	**2.33**
10	88.18	79.09	**8.95**	2.34
15	88.16	78.99	16.93	2.35

TABLE IV

COMPARISON BETWEEN THE SET-HQ-AGNOSTIC AND THE SET-HQ-KNOWABLE SCENARIOS ON LA SEGMENTATION UNDER DIFFERENT HQ LABELED DATA SETTINGS. NOTE THAT WE ADOPT MTCL-HARD IN THIS TASK FOR ITS SUPERIOR PERFORMANCE. $*$ INDICATES SIGNIFICANT IMPROVEMENT FROM THE SET-HQ-AGNOSTIC VERSION ($p \leq 0.05$). BEST RESULTS ARE IN BOLD.

Methods	Settings	Metrics			
	HQ / knowable?	Dice [%] ↑	Jaccard [%] ↑	95HD [mm] ↓	ASD [mm] ↓
MTCL (ours)	10% / ×	86.06	75.74	11.16	2.66
MTCL (ours)	10% / ✓	**88.34***	**79.27***	**10.85**	**2.33**
MTCL (ours)	20% / ×	85.33	74.66	15.85	2.43
MTCL (ours)	20% / ✓	**89.25***	**80.72***	**10.15***	**2.24**
MTCL (ours)	30% / ×	86.94	77.06	9.51	2.94
MTCL (ours)	30% / ✓	**89.98***	**81.89***	**6.35***	**1.89***

input greatly interferes with the network training, resulting in significant performance degradation. As a result, we strongly advise using the proposed MTCL in conjunction with the Set-HQ-knowable scenario, i.e., making feasible efforts to obtain a reasonable amount of HQ labeled data.

C. Experiments on Real-world Hepatic Vessel Datasets

1) Task-related Implementation Details: Experimentally, the task of hepatic vessel segmentation is far more challenging than the former LA segmentation task. Following our preliminary version [1], we perform this vessel segmentation task in 2D with 2D U-Net [37] as the backbone since we experimentally found the performance of 3D networks is depressingly worse than the 2D ones in this task, which may be because of large imaging thickness variation (in and between the two datasets) and imbalanced and sparse objects. The batch size is set to 4, including 2 HQ labeled images and 2 LQ labeled images in each mini-batch. $\mathcal{L}_{hs}$ is a combination of cross-entropy loss, Dice loss, focal loss [38] and boundary loss [39], as such a combination performs the best in our exploratory fully-supervised experiments. $\mathcal{L}_{ls}$ shares the similar form as

TABLE V

EXPLORATORY EXPERIMENTAL RESULTS FOR DIFFERENT INPUT TYPES ON SET-HQ (I.E., 3DIRCADB [27]) UNDER THE 10 HQ LABELED DATA SETTING. "$\boldsymbol{i}$", "$\boldsymbol{p}$" AND "$\boldsymbol{c}$" REPRESENT THE CT IMAGE, THE VESSEL PROBABILITY MAP AND THE CONCATENATED ONE, RESPECTIVELY. $*$ INDICATES $\boldsymbol{p \leq 0.05}$ FROM A TWO-SIDED PAIRED T-TEST WHEN COMPARING H-SUP (C) WITH OTHERS. THE BEST MEAN RESULTS ARE IN BOLD.

Methods	Metrics			
	Dice [%] ↑	Jaccard [%] ↑	95HD [mm] ↓	ASD [mm] ↓
H-Sup (i)	60.93	45.24	10.31	2.72
H-Sup (p)	60.82	46.08	10.29	2.36
H-Sup (c)	**66.85***	**51.50***	**9.21***	**2.05***

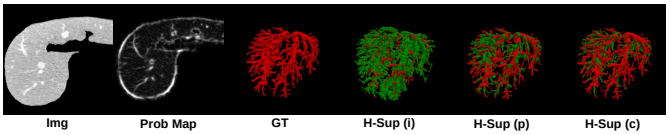


Fig. 7. Visualization of the fused hepatic vessel segmentation results of different input types on Set-HQ under the 10 HQ labeled data setting. "$\boldsymbol{i}$", "$\boldsymbol{p}$" and "$\boldsymbol{c}$" represent the CT image, the vessel probability map and the concatenated one, respectively. Red represents the ground truth, while green denotes the difference between the ground truth and the predicted vessel.

$\mathcal{L}_{hs}$ but without the boundary loss as it was not designed for soft labels. The maximum training iteration is set to 30,000. For inference, the trained model segments each volume slice-by-slice and the 2D predictions are concatenated back into 3D volume. A post-processing step that removes very small regions (less than 0.1% of the volume size) is performed.

2) Robust Preprocessing Strategy for Hepatic Vessel: We perform several exploratory fully supervised experiments on the Set-HQ. For image preprocessing, a standard preprocessing strategy is firstly applied. We mask and center-crop images to the liver region with the size of $320 \times 320 \times D$ voxels, where D denotes the total slice number of the volume. Note that for the MSD8 dataset, the liver masks are obtained with the off-the-shelf trained H-DenseUNet model [40] because no liver annotation is provided in the original dataset. Then, the intensities are truncated to $[-100, 250]$ HU, followed by min-max normalization.

However, we observe that many images have intrinsic image noises [41] and heterologous intensity distribution (also known as domain shift), which could drive the model to be over-sensitive to the high-intensity regions and unstable for the prediction, as shown in Table V and Fig. 7. To alleviate the above challenges, we further introduce the vessel probability map based on the Sato tubeness filter [42] as another modality to provide auxiliary information. By calculating the eigenvectors of the Hessian matrix, the similarities of the pixel to a tube can be obtained, which can enhance the potential vessel regions with high probability (illustrated in Fig. 7). Following the input-level fusion strategy commonly used in brain tumor segmentation [43], we concatenate the vessel probability maps with the images in the input space. By jointly considering such dual-information (H-Sup (c)), the network could perceive more robust vessel signals towards better segmentation performance (shown in Table V and Fig. 7). Therefore, the following experiments will be conducted with the fused input.

TABLE VI

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON HEPATIC VESSEL SEGMENTATION. $*$ INDICATES $\boldsymbol{p \leq 0.05}$ FROM A TWO-SIDED PAIRED T-TEST WHEN COMPARING THE RESULTS OF OUR BEST MODEL WITH THE BEST-PERFORMING METHOD UNDER THE SET-HQ-KNOWABLE SETTING. BEST MEAN RESULTS ARE SHOWN IN BOLD.

Methods	Settings	Metrics			
	HQ / knowable?	Dice [%] ↑	Jaccard [%] ↑	95HD [mm] ↓	ASD [mm] ↓
H-Sup	10 / -	66.85	51.50	9.21	2.05
HL-Sup	10 / -	63.38	48.17	9.20	1.61
TriNet [6]	10 / ×	66.78	51.94	9.46	2.35
2SRnT [5]	10 / ×	68.24	52.18	8.84	1.59
PNL [4]	10 / ×	66.73	51.02	9.33	2.01
Decoupled [2]	10 / ✓	70.35	54.12	7.58	1.33
KDEM [3]	10 / ✓	70.89	54.91	7.82	1.35
MTCL-Hard (ours)	10 / ✓	71.88	56.13	7.39	1.26
MTCL-FS (ours)	10 / ✓	72.16	56.73	7.32	1.29
MTCL-UDS (ours)	10 / ✓	**72.82***	**57.26***	**7.28**	**1.25**
H-Sup	5 / -	63.37	49.21	10.02	2.83
HL-Sup	5 / -	62.09	48.86	9.68	2.85
TriNet [6]	5 / ×	65.32	51.24	9.52	2.47
2SRnT [5]	5 / ×	66.47	51.83	8.67	1.98
PNL [4]	5 / ×	64.06	50.39	9.63	2.35
Decoupled [2]	5 / ✓	67.72	51.63	8.31	1.61
KDEM [3]	5 / ✓	68.98	53.46	8.04	1.55
MTCL-Hard (ours)	5 / ✓	70.08	54.76	7.85	1.48
MTCL-FS (ours)	5 / ✓	70.78	55.01	7.77	1.49
MTCL-UDS (ours)	5 / ✓	**71.45***	**55.57***	**7.68***	**1.43***

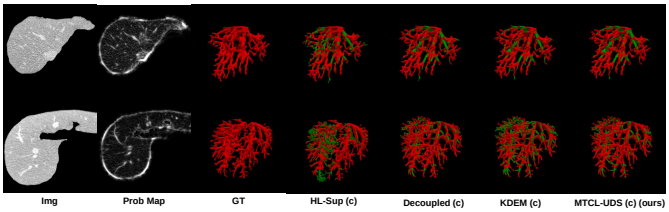


Fig. 8. Visualization of the fused hepatic vessel segmentation results from different methods under the 10 HQ labeled data setting. The red voxels represent the ground truth, while the green voxels denote the difference between the ground truth and the predicted vessel.

3) Experimental Results: Table VI presents the quantitative comparison results for hepatic vessel segmentation with additional real-world Set-LQ, i.e., the MSD8 dataset. As predicted, the LQ noisy labels of Set-LQ cause unavoidable performance degradation. Especially, compared with H-Sup, HL-Sup's Dice score and Jaccard significantly drop from 66.85% to 63.38%, and from 51.50% to 48.17%, respectively, under the 10 HQ labeled data setting. Among the Set-HQ-agnostic methods, TriNet and PNL show limited ability in handling this challenging task, while 2SRnT achieves more incremental gains from Set-LQ. In contrast, the Set-HQ-knowable baseline methods, i.e., Decoupled and KDEM, show excellent ability in exploiting the Set-LQ, resulting in significant improvement. Even so, the proposed MTCL approaches with three different refinement strategies achieve more appealing improvement in terms of all four metrics under the two HQ labeled data settings. Particularly, MTCL-UDS performs better in terms of Dice, Jaccard and 95HD metrics, implying that accurately estimating the error map in this task is quite difficult so that the model resorts to calibrating the refinement process based on its epistemic uncertainty. Therefore, we prefer MTCL-UDS as our final model in this task. Fig. 8 further presents the 3D rendering of the segmentation results obtained by our MTCL-UDS and other approaches under the 10 HQ labeled data setting. Consistently, the predicted vessels of MTCL-UDS achieve more appealing visual results.

D. Experiments on Real-world Retinal Vessel Datasets

1) Task-related Implementation Details: Following the above hepatic vessel segmentation, the same 2D U-Net [37] is also adopted here. Previous fully-supervised work in retinal vessel segmentation [33] implied that patch-based segmentation has more appealing performance. Similarly, we randomly extract 7,500 partly overlapped patches for each sample with the size of 48×48 pixels, resulting in 150,000 patches from the STARE dataset. The batch size is set to 64, including 32 HQ labeled patches and 32 LQ labeled patches in each mini-batch. The maximum training iteration is set to 70,000. Both $\mathcal{L}_{hs}$ and $\mathcal{L}_{ls}$ use the standard cross-entropy loss. In the test stage, ordered patches are extracted at the stride of 16 and the final result is obtained by stitching the corresponding patch predictions together. The probability maps of the partly overlapped patches are averaged.

2) Experimental Results: Table VII presents the quantitative comparison results for retinal vessel segmentation. Under both HQ labeled data settings (i.e., two and four HQ samples, respectively), the MSF-based LQ noisy labels of Set-LQ cause unavoidable performance degradation. Among the Set-HQ-agnostic methods, PNL shows limited ability in handling this challenging task, while TriNet and 2SRnT can marginally benefit from Set-LQ. Regarding the Set-HQ-knowable methods, Decoupled achieves incremental improvement with Set-LQ, whereas we found that KDEM cannot achieve satisfactory results here. Overall, the proposed MTCL approaches with three different refinement strategies achieve more appealing improvement in terms of all four metrics. Particularly, MTCL-UDS achieves the best performance, implying that accurately estimating the error map in this task is also challenging so that the model uncertainty-driven refinement is more beneficial. Therefore, we select MTCL-UDS as our final model for this task. Fig. 9 presents the exemplar segmentation results of our MTCL-UDS and other approaches under the 4 HQ labeled data setting. As observed, the predicted vessels of MTCL-UDS are more consistent with the ground truth.

TABLE VII

QUANTITATIVE COMPARISON ON RETINAL VESSEL SEGMENTATION. $*$ INDICATES $p \leq 0.05$ FROM A TWO-SIDED PAIRED T-TEST WHEN COMPARING THE RESULTS OF OUR BEST MODEL WITH THE BEST-PERFORMING METHOD UNDER THE SET-HQ-KNOWABLE SETTING. BEST RESULTS ARE IN BOLD.

Methods	Settings	Metrics			
	HQ / knowable?	Dice [%] ↑	AUC [%] ↑	SEN [%] ↑	ACC [%] ↑
H-Sup	2 / -	64.95	93.50	49.58	93.18
HL-Sup	2 / -	63.93	93.48	48.71	92.99
TriNet [6]	2 / ×	64.32	93.39	48.01	92.06
2SRnT [5]	2 / ×	65.57	93.72	50.71	92.17
PNL [4]	2 / ×	61.99	92.28	47.37	92.62
Decoupled [2]	2 / ✓	65.21	92.56	52.02	92.08
KDEM [3]	2 / ✓	63.73	92.24	49.56	91.45
MTCL-Hard (ours)	2 / ✓	67.84	93.99	54.34	93.44
MTCL-FS (ours)	2 / ✓	69.67	93.90	57.19	**93.68**
MTCL-UDS (ours)	2 / ✓	**69.97***	**94.26***	**57.95***	93.67*
H-Sup	4 / -	67.69	95.43	52.88	93.57
HL-Sup	4 / -	66.63	94.61	52.66	93.28
TriNet [6]	4 / ×	68.21	93.38	57.39	93.52
2SRnT [5]	4 / ×	68.53	94.73	58.33	93.97
PNL [4]	4 / ×	65.08	93.06	52.61	92.16
Decoupled [2]	4 / ✓	68.37	94.02	59.22	93.85
KDEM [3]	4 / ✓	66.52	93.65	57.67	93.29
MTCL-Hard (ours)	4 / ✓	71.89	93.72	64.44	93.89
MTCL-FS (ours)	4 / ✓	71.94	93.79	67.02	93.21
MTCL-UDS (ours)	4 / ✓	**72.02***	**95.33***	**67.37***	**94.26**
H-Sup (upper bound)	20 (100%) / -	78.19	97.23	69.08	95.09

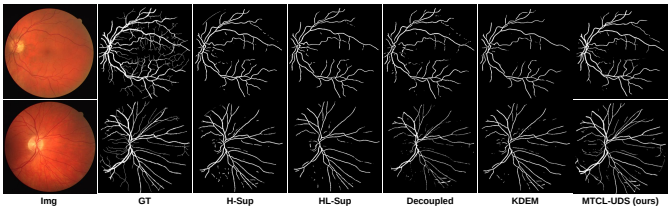


Fig. 9. Exemplar retinal vessel segmentation results under the 4 HQ labeled data setting.

V. DISCUSSION

A. Choice of Specific Label Refinement Strategy

Noteworthily, all the proposed label refinement strategies do not introduce additional trainable parameters. From our experiments, for the easier case of left atrium segmentation with relatively clear and large objects, the network can well-segment most of the organ regions when trained with limited HQ labeled training data, wherein the estimated error map $\mathbf{X}_{err}$ can also accurately characterize most of the label errors. Thus, in this case, choosing the simple hard refinement may make full use of the satisfying refined labels, while the smooth refinement may reduce the efficacy of these labels. In contrast, for more challenging tasks like segmenting complex, class-imbalanced and sparse objects (e.g., the hepatic vessel in CT images and retinal vessel in fundus images), it is much more difficult for the network to learn discriminative representations for these complex topologies with limited available HQ labeled data, resulting in heavy noises in the estimated error map as well. In fact, these tasks are also difficult for human experts. In these cases, we found that the smooth refinement strategies, including FS and UDS, provide better results. Particularly, the model-driven calibration (i.e., MTCL-UDS) performs better in both aforementioned tasks. Since UDS requires T forward passes ($T = 6$ in this work) for uncertainty estimation at each iteration, training each MTCL-UDS is experimentally $\sim$1.4x slower than that of MTCL-Hard and MTCL-FS but it is still acceptable considering its superior performance. Despite FS being more computationally efficient for training each model compared to MTCL-UDS, it treats all pixels equally and introduces an extra fixed hyper-parameter τ that additionally requires laborious tuning. For reference, τ is set to 0.8 here since we found that this value provides the best results in both vessel segmentation tasks after grid search from 0.5 to 0.9. Overall, regarding the difficult objects, we recommend MTCL-UDS for less human efforts and its superior performance on both real-world vessel segmentation tasks in case of suboptimal performance of MTCL-FS (with our reference smooth factor).

B. Visualization of Label Self-denoising

As shown in Sec. IV-B.3, the LQ label self-denoising scheme contributes significantly to superior performance, wherein the teacher model serves as a "third party" to help estimate the joint distribution matrix between the LQ noisy labels and its true (latent) labels to explicitly locate the errors. We further visualize the estimated error maps ($\mathbf{X}_{err}$) for a dilated LQ annotation of LA, a real-world LQ noisy annotation

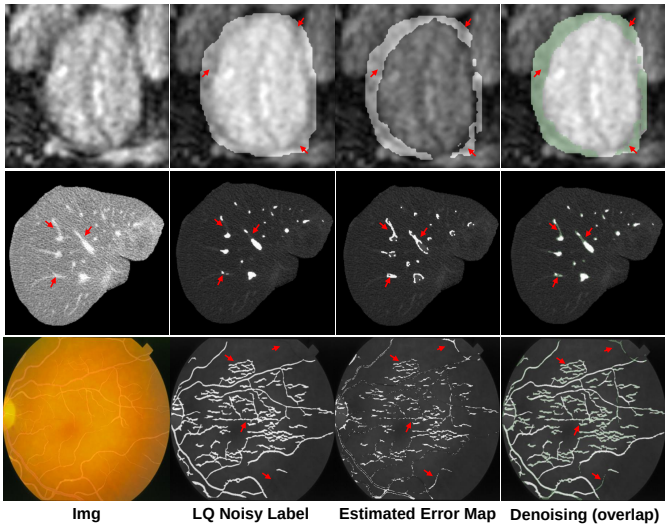


Fig. 10. Visualization of the label self-denoising performance. The overlaid green color represents the denoising areas. The red arrows indicate regions with obvious wrong labels.

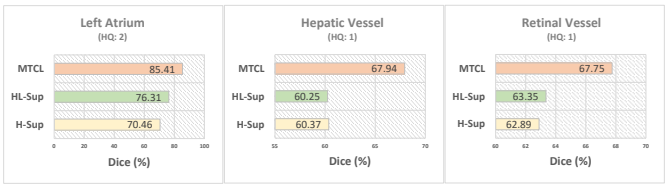


Fig. 11. Dice score comparison under the extremely scarce Set-HQ. Note that we adopt MTCL-Hard for LA segmentation, and MTCL-UDS for hepatic vessel and retinal vessel segmentation.

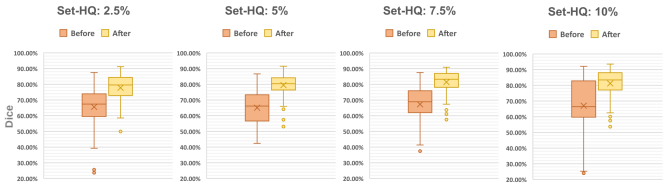


Fig. 12. Improvement of label quality of Set-LQ (indicated by Dice) before and after MTCL-based self-denoising with varying percentages of HQ labeled data. Data are interpreted as box plots. The central lines indicate median Dice values, × the average Dice values, boxes the interquartile range, whiskers the smallest and largest values, and data points ○ outliers.

of the hepatic vessel (from MSD8 [28]) and a machine-generated LQ annotation of retinal vessel (from STARE [29]). As shown in Fig. 10, for the LA example, the dilated area is well-characterized in the estimated error map, while some noticeable mislabeled pixels can also be identified in the real-world noisy annotations of the hepatic vessel and retinal vessel. In practice, the estimated error map can provide guidance for experts to facilitate the laborious label quality control process or for the junior labelers as a self-teaching reference.

C. Extended Study with Extremely Scarce Set-HQ

Besides the above experiments, we further investigate the performance of the proposed MTCL under the most extreme HQ labeled condition. Under this setting, only 2 HQ labeled volumes (the minimum HQ batch size) are available in LA segmentation, while only one HQ labeled sample is available in both hepatic and retinal vessel segmentation. As observed in Fig. 11, regarding the fully supervised baselines, when trained with extremely scarce HQ labeled images, H-Sup performs poorly in three tasks and can further benefit from the additional LQ labeled data in both LA segmentation and retinal vessel segmentation tasks. Based on the results in Sec. IV, we adopt MTCL-Hard for LA segmentation and MTCL-UDS for hepatic vessel and retinal vessel segmentation, respectively. Encouragingly, the proposed MTCL still performs admirably under the most extreme condition. Such experiments indicate that our MTCL is still powerful when confronted with extremely scarce HQ labeled data, resulting in more clinical values.

D. Potential in Improving Public Datasets with Tiny HQ Labeled Data

Furthermore, while publicly available medical image datasets are becoming abundant, they are often from different institutes with different labeling guidelines/criteria, resulting in varying levels of label quality. This self-denoising scheme also suggests a promising application that explicitly and explicably improves the label quality of the public datasets by taking advantage of limited HQ labeled data. Taking the simulated LA dataset as an example, we assume the scenario that only very limited HQ labeled cases (2.5%, 5%, 7.5% and 10%) can be acquired from the radiologists (i.e., Set-HQ), while the remaining cases (97.5%, 95%, 92.5% and 90%) in the training set (i.e., Set-LQ) are labeled by non-experts. Since the noisy labels of Set-LQ are corrupted from the original HQ labels, we can measure the difference between the simulated LQ labels and the original expert labels (i.e., "Before"). Here, we adopt the Dice score as the metric. Then, we train our MTCL with the tiny Set-HQ and the remaining Set-LQ, and subsequently derive the denoised LQ labels of Set-LQ from the optimal model so that the Dice score between the denoised LQ labels and the original labels (i.e., "After") can also be obtained. Note that the LQ labels are generated separately in the four experiments. As observed in the box plots of Fig. 12, MTCL can significantly improve the label quality assisted by very limited HQ labeled data ($p < 0.05$ for all). Particularly, even with the extremely scarce Set-HQ (i.e., 2.5%), our MTCL can still effectively improve the label quality.

E. Limitation and Future Works

Despite achieving exciting performance, one limitation of this work is that we assume both HQ labeled data and LQ unlabeled data are from the same domain distribution. Such a limitation also applies to most of the current semi-supervised methods. Considering this, the proposed method is more suitable for the scenario that all images are acquired from similar imaging protocols. As a result, this will limit the availability of LQ labeled data. Intuitively, besides the inferior label qualities in the LQ labeled data, a more generalized yet challenging scenario is that the notorious domain shift between Set-HQ and Set-LQ also exists. In other words, we have to

simultaneously combat the LQ noisy labels and the domain gap during training. Therefore, further investigating how to take advantage of recent domain adaptation techniques to deal with potential domain shifts in our framework is an interesting future direction with fruitful clinical values.

VI. Conclusion

In this work, we first categorized existing noisy supervised learning approaches by two clinically practical scenarios, i.e., Set-HQ-agnostic and Set-HQ-knowable settings. Targeting at Set-HQ-knowable setting, we proposed a novel Mean-Teacher-assisted Confident Learning (MTCL) framework constructed by a teacher-student architecture and a synergistic label self-denoising process to simultaneously learn segmentation from a small set of high-quality labeled data and plentiful low-quality noisy labeled data. Comprehensive experiments on different segmentation tasks demonstrated the superiority of our method over the state-of-the-art approaches under both simulated and real-world label noise settings. Besides, the extended study showed that our framework has great potential in improving the label quality of noisy labeled data with only a small amount of high-quality labeled data.

References

[1] Z. Xu, D. Lu, Y. Wang, J. Luo, J. Jagadeesan, K. Ma, Y. Zheng, and X. Li, "Noisy labels are treasure: Mean-teacher-assisted confident learning for hepatic vessel segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2021, pp. 3–13.

[2] W. Luo and M. Yang, "Semi-supervised semantic segmentation via strong-weak dual-branch network," in *European Conference on Computer Vision*. Springer, 2020, pp. 784–800.

[3] J. Dolz, C. Desrosiers, and I. B. Ayed, "Teach me to segment with mixed supervision: Confident students become masters," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 517–529.

[4] H. Zhu, J. Shi, and J. Wu, "Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 576–584.

[5] M. Zhang, J. Gao, Z. Lyu, W. Zhao, Q. Wang, W. Ding, S. Wang, Z. Li, and S. Cui, "Characterizing label errors: Confident learning for noisy-labeled image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2020, pp. 721–730.

[6] T. Zhang, L. Yu, N. Hu, S. Lv, and S. Gu, "Robust medical image segmentation from non-expert annotations with tri-network," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2020, pp. 249–258.

[7] D. Angluin and P. Laird, "Learning from noisy examples," *Machine Learning*, vol. 2, no. 4, pp. 343–370, 1988.

[8] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.

[9] G. Yang, C. Wang, J. Yang, Y. Chen, L. Tang, P. Shao, J.-L. Dillenseger, H. Shu, and L. Luo, "Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal CTA images," *BMC Medical Imaging*, vol. 20, no. 1, pp. 1–12, 2020.

[10] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3D medical images by playing a Rubik's cube," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 420–428.

[11] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical Image Analysis*, vol. 58, p. 101539, 2019.

[12] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[13] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017, pp. 1195–1204.

[14] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *arXiv preprint arXiv:1903.03825*, 2019.

[15] Z. Xu, Y. Wang, D. Lu, L. Yu, J. Yan, J. Luo, K. Ma, Y. Zheng, and R. K.-y. Tong, "All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[16] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2017, pp. 568–576.

[17] M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz *et al.*, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 2, pp. 674–683, 2016.

[18] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *International Conference on Learning Representations*, 2016.

[19] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, "Robust learning at noisy labeled medical images: Applied to skin lesion classification," in *16th IEEE International Symposium on Biomedical Imaging*. IEEE, 2019, pp. 1280–1283.

[20] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang, "Smooth neighbors on teacher graphs for semi-supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8896–8905.

[21] C. Elkan, "The foundations of cost-sensitive learning," in *International Joint Conference on Artificial Intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[22] P. Chen, B. B. Liao, G. Chen, and S. Zhang, "Understanding and utilizing deep neural networks trained with noisy labels," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1062–1070.

[23] "What uncertainties do we need in Bayesian deep learning for computer vision?, author=Kendall, Alex and Gal, Yarin," *arXiv preprint arXiv:1703.04977*, 2017.

[24] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 554–565.

[25] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 605–613.

[26] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 67, p. 101832, 2021.

[27] 3DIRCADb Dataset. [Online]. Available: https://www.ircad.fr/research/3d-ircadb-01/

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

[29] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.

[30] B. M. Dawant, R. Li, B. Lennon, and S. Li, "Semi-automatic segmentation of the liver and its evaluation on the MICCAI 2007 grand challenge data set," *3D Segmentation in The Clinic: A Grand Challenge*, pp. 215–221, 2007.

[31] L. Liu, J. Tian, C. Zhong, Z. Shi, and F. Xu, "Robust hepatic vessels segmentation model based on noisy dataset," in *SPIE Conference on Medical Imaging: Computer-Aided Diagnosis*, vol. 11314. International Society for Optics and Photonics, 2020, p. 113140L.

[32] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[33] Y. Wu, Y. Xia, Y. Song, Y. Zhang, and W. Cai, "NFN+: a novel network followed network for retinal vessel segmentation," *Neural Networks*, vol. 126, pp. 153–162, 2020.

[34] A. W. Setiawan, T. R. Mengko, O. S. Santoso, and A. B. Suksmono, "Color retinal image enhancement using CLAHE," in *International Conference on ICT for Smart Society*. IEEE, 2013, pp. 1–3.

[35] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *arXiv preprint arXiv:1804.06872*, 2018.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 234–241.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[39] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, pp. 285–296.

[40] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[41] X. Duan, J. Wang, S. Leng, B. Schmidt, T. Allmendinger, K. Grant, T. Flohr, and C. H. McCollough, "Electronic noise in CT detectors: impact on image noise and artifacts," *American Journal of Roentgenology*, vol. 201, no. 4, pp. W626–W632, 2013.

[42] Y. Sato, S. Nakajima, N. Shiraga, H. Atsumi, S. Yoshida, T. Koller, G. Gerig, and R. Kikinis, "Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images," *Medical Image Analysis*, vol. 2, no. 2, pp. 143–168, 1998.

[43] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, 2019.