Disentangling Human Error from Ground Truth in Segmentation of Medical Images

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent years have seen increasing use of supervised learning methods for segmentation tasks. However, the predictive performance of these algorithms depends on the quality of labels. This problem is particularly pertinent in the medical image domain, where both the annotation cost and inter-observer variability are high. In a typical label acquisition process, different human experts provide their estimates of the "true" segmentation labels under the influence of their own biases and competence levels. Treating these noisy labels blindly as the ground truth limits the performance that automatic segmentation algorithms can achieve. In this work, we present a method for jointly learning, from purely noisy observations alone, the reliability of individual annotators and the true segmentation label distributions, using two coupled CNNs. The separation of the two is achieved by encouraging the estimated annotators to be maximally unreliable while achieving high fidelity with the noisy training data. We first define a toy segmentation dataset based on MNIST and study the properties of the proposed algorithm. We then demonstrate the utility of the method on three public medical imaging segmentation datasets with simulated (when necessary) and real diverse annotations: 1) MSLSC (multiple-sclerosis lesions); 2) BraTS (brain tumours); 3) LIDC-IDRI (lung abnormalities). In all cases, our method outperforms competing methods and relevant baselines particularly in cases where the number of annotations is small and the amount of disagreement is large. The experiments also show strong ability to capture the complex spatial characteristics of annotators' mistakes, which could be potentially utilised for the purpose of education.

1 Introduction

Segmentation of anatomical structures in medical images is known to suffer from high inter-reader variability [1, 2, 3, 4, 5], affecting limiting the performance of downstream supervised machine learning models. This problem is particularly prominent in the medical domain where the labelled data is commonly scarce due to the high cost of annotations. For instance, accurate identification of multiple sclerosis (MS) lesions in MRIs is difficult even for experienced experts due to variability in lesion location, size, shape and anatomical variability across patients [6]. Another example [4] reports the average inter-reader variability in the range 74-85% for glioblastoma (a type of brain tumour) segmentation. Further aggravated by differences in biases and levels of expertise, segmentation annotations of structures in medical images suffer from high annotation variations [7]. In consequence, despite the present abundance of medical imaging data thanks to over two decades of digitisation, the world still remains relatively short of access to data with curated labels [8], that is amenable to machine learning, necessitating intelligent methods to learn robustly from such noisy annotations.

To mitigate inter-reader variations, different pre-processing techniques are commonly used to curate segmentation annotations by fusing labels from different experts. The most basic yet popular approach is based on the majority vote where the most representative opinion of the experts is treated as the ground truth (GT). A smarter version that accounts for similarity of classes has proven effective in

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

aggregation of brain tumour segmentation labels [4]. A key limitation of such approaches, however, is that all experts are assumed to be equally reliable. Warfield *et al.*[9] proposed a label fusion method, called STAPLE that explicitly models the reliability of individual experts and uses that information to "weigh" their opinions in the label aggregation step. After consistent demonstration of its superiority over the standard majority-vote pre-processing in multiple applications, STAPLE has become the go-to label fusion method in the creation of public medical image segmentation datasets e.g., ISLES [10], MSSeg [11], Gleason'19 [12] datasets. Asman *et al.*later extended this approach in [13] by accounting for voxel-wise consensus to address the issue of under-estimation of annotators' reliability. In [14], another extension was proposed in order to model the reliability of annotators across different pixels in images. More recently, within the context of multi-atlas segmentation problems [15] where image registration is used to warp segments from labeled images ("atlases") onto a new scan, STAPLE has been enhanced in multiple ways to encode the information of the underlying images into the label aggregation process. A notable example is STEP proposed in Cardoso *et al.*[16] who designed a strategy to further incorporate the local morphological similarity between atlases and target images, and different extensions of this approach such as [17, 18] have since been considered. However, these previous label fusion approaches have a common drawback—they critically lack a mechanism to integrate information across different training images. This fundamentally limits the remit of applications to cases where each image comes with a reasonable number of annotations from multiple experts, which can be prohibitively expensive in practice. Moreover, relatively simplistic functions are used to model the relationship between observed noisy annotations, true labels and reliability of experts, which may fail to capture complex characteristics of human annotators.

In this work, we introduce the first instance of an end-to-end supervised segmentation method that jointly estimates, from noisy labels alone, the reliability of multiple human annotators and true segmentation labels. The proposed architecture (Fig. 1) consists of two coupled CNNs where one estimates the true segmentation probabilities and the other models the characteristics of individual annotators (e.g., tendency to over-segmentation, mix-up between different classes, etc) by estimating the pixel-wise confusion matrices (CMs) on a per image basis. Unlike STAPLE [9] and its variants, our method models, and disentangles with deep neural networks, the complex mappings from the input images to the annotator behaviours and to the true segmentation label. Furthermore, the parameters of the CNNs are "global variables" that are optimised across different image samples; this enables the model to disentangle robustly the annotators' mistakes and the true labels based on correlations between similar image samples, even when the number of available annotations is small per image (e.g., a single annotation per image). In contrast, this would not be possible with STAPLE [9] and its variants [14, 16] where the annotators' parameters are estimated on every target image separately.

For evaluation, we first simulate a diverse range of annotator types on the MNIST dataset by performing morphometric operations with Morpho-MNIST framework [19]. Then we demonstrate the potential in several real-world medical imaging datasets, namely (i) MS lesion segmentation dataset (MSLSC) from the ISBI 2015 challenge [20], (ii) Brain tumour segmentation dataset (BraTS) [4] and (iii) Lung nodule segmentation dataset (LIDC-IDRI) [21]. Experiments on all datasets demonstrate that our method consistently leads to better segmentation performance compared to widely adopted label-fusion methods and other relevant baselines, especially when the number of available labels for each image is low and the degree of annotator disagreement is high.

2 Related Work

The majority of algorithmic innovations in the space of *label aggregation for segmentation* have uniquely originated from the medical imaging community, partly due to the prominence of the inter-reader variability problem in the field, and the wide-reaching values of reliable segmentation methods. The aforementioned methods based on the STAPLE-framework such as [9, 13, 14, 16, 22, 17, 17, 18, 23] are based on generative models of human behaviours, where the latent variables of interest are the unobserved true labels and the "reliability" of the respective annotators. Our method can be viewed as an instance of translation of the STAPLE-framework to the supervised learning paradigm. As such, our method produces a model that can segment test images without needing to acquire labels from annotators or atlases unlike STAPLE and its local variants. Another key difference is that our method is jointly trained on many different subjects while the STAPLE-variants are only fitted on a per-subject basis. This means that our method is able to learn from correlations between different subjects, which previous works have not attempted— for example, our method uniquely can estimate the reliability and true labels even when there is only one label available per input image as shown later.

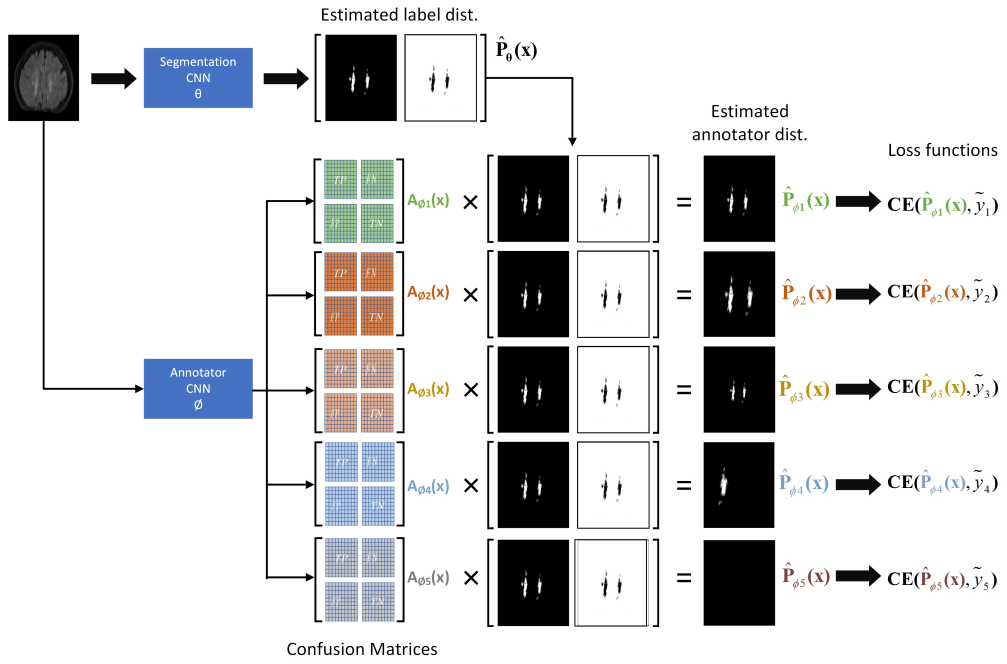

Figure 1: General schematic of the neural fusion network. The method consists of two components: (1) Segmentation network parametrised by θ that generates an estimate of the GT segmentation probabilities, $p_\theta(\mathbf{x})$ for the given input image $\mathbf{x}$; (2) Annotator network, parametrised by ϕ, that estimates the CMs $\{\mathbf{A}_\phi^{(r)}(\mathbf{x})\}_{r=1}^n$ of the annotators. The segmentation probabilities of respective annotators $\hat{\mathbf{p}}_\phi^{(r)}(\mathbf{x}) := \mathbf{A}_\phi^{(r)}(\mathbf{x}) \cdot \mathbf{p}_\theta(\mathbf{x})$ are then computed. The model parameters $\{\theta,\phi\}$ are optimized to minimize the sum of five cross-entropy losses between each estimated annotator distribution $\mathbf{p}_\phi^{(r)}(\mathbf{x})$ and the noisy labels $\tilde{\mathbf{y}}^{(r)}$ observed from each annotator.

Our work also relates to a recent strand of methods that aim to generate a set of diverse and plausible segmentation proposals on a given image. Notably, probabilistic U-net [24] and its recent variants, PHiSeg [25] have shown that the aforementioned inter-reader variations in segmentation labels can be modelled with sophisticated forms of probabilistic CNNs. Such approaches, however, fundamentally differ from ours in that variable annotations from many experts in the training data are assumed to be all realistic instances of the true segmentation; we assume, on the other hand, that there is a single, unknown, true segmentation map of the underlying anatomy, and each individual annotator produces a noisy approximation to it with variations that reflect their individual characteristics. The latter assumption may be reasonable in the context of segmentation problems since there exists only one true boundary of the physical objects captured in an image while multiple hypothesis can arise from ambiguities in human interpretations.

We also note that, in standard classification problems, a plethora of different works have shown the utility of modelling the labeling process of human annotators in restoring the true label distribution [26, 27, 28]. Such approaches can be categorized into two groups: (1) *two-stage* approach [29, 30, 31, 32, 33], and (2) *simultaneous* approach. In the first category, the noisy labels are first curated through a probabilistic model of annotators, and subsequently, a supervised machine-learning model is trained on the curated labels. The initial attempt [29] was made in the early 1970s, and numerous advances such as [30, 31, 32, 33] since built upon this work e.g. by estimating sample difficulty and human biases. In contrast, models in the second category aim to curate labels and learn a supervised model jointly in an end-to-end fashion [34, 35, 36, 37, 27, 28] so that the two components inform each other. Although the evidence still remains limited to the simple classification task, these *simultaneous* approaches have shown promising improvements over the methods in the first category in terms of the predictive performance of the supervised model and the sample efficiency (i.e., fewer labels are required per input). However, to date very little attention has been paid to the same problem in more complicated, structured prediction tasks where the outputs are high dimensional. In this work, we propose the first *simultaneous* approach to addressing such a problem for image segmentation, while drawing inspirations from the STAPLE framework [9] which would fall into the *two-stage* approach category.

3 Method

3.1 Problem Set-up

In this work, we consider the problem of learning a supervised segmentation model from noisy labels acquired from multiple human annotators. Specifically, we consider a scenario where set of images $\{\mathbf{x}_n \in \mathbb{R}^{W\times H\times C}\}_{n=1}^{N}$ (with W,H,C denoting the width, height and channels of the image) are assigned with noisy segmentation labels $\{\tilde{\mathbf{y}}_n^{(r)} \in \mathcal{Y}^{W\times H}\}_{n=1,...,N}^{r\in S(\mathbf{x}_i)}$ from multiple annotators where $\tilde{\mathbf{y}}_n^{(r)}$ denotes the label from annotator $r \in \{1,...,R\}$ and $S(\mathbf{x}_n)$ denotes the set of all annotators who labelled image $\mathbf{x}_i$ and $\mathcal{Y}=[1,2,...,L]$ denotes the set of classes.

Here we assume that every image $\mathbf{x}$ annotated by at least one person i.e., $|S(\mathbf{x})| \geq 1$, and no GT labels $\{\mathbf{y}_n \in \mathcal{Y}^{W\times H}\}_{n=1,...,N}$ are available. The problem of interest here is to *learn the unobserved true segmentation distribution* $p(\boldsymbol{y} \mid \boldsymbol{x})$ *from such noisy labelled dataset* $\mathcal{D} = \{\boldsymbol{x}_n, \tilde{\boldsymbol{y}}_n^{(r)}\}_{n=1,...,N}^{r\in S(\mathbf{x}_n)}$ i.e., the combination of images, noisy annotations and experts' identities for labels (which label was obtained from whom).

We also emphasise that *the goal at inference time is to segment a given unlabelled test image* but not to fuse multiple available labels as is typically done in multi-atlas segmentation approaches [15].

3.2 Probabilistic Model and Proposed Architecture

Here we describe the probabilistic model of the observed noisy labels from multiple annotators. We make two key assumptions: (1) annotators are statistically independent, (2) annotations over different pixels are independent given the input image. Under these assumptions, the probability of observing noisy labels $\{\tilde{\mathbf{y}}^{(r)}\}_{r\in S(\mathbf{x})}$ on $\mathbf{x}$ factorises as:

$$p(\{\tilde{\mathbf{y}}^{(r)}\}_{r\in S(\mathbf{x})} \mid \mathbf{x}) = \prod_{r\in S(\mathbf{x})} p(\tilde{\mathbf{y}}^{(r)} \mid \mathbf{x}) = \prod_{r\in S(\mathbf{x})} \prod_{\substack{w\in\{1,...,W\}\\ h\in\{1,...,H\}}} p(\tilde{y}_{wh}^{(r)} \mid \mathbf{x}) \tag{1}$$

where $\tilde{y}_{wh}^{(r)} \in [1,...,L]$ denotes the $(w,h)^{\text{th}}$ elements of $\tilde{\mathbf{y}}^{(r)} \in \mathcal{Y}^{W\times H}$. Now we rewrite the probability of observing each noisy label on each pixel (w,h) as:

$$p(\tilde{y}_{wh}^{(r)} \mid \mathbf{x}) = \sum_{y_{wh}=1}^{L} p(\tilde{y}_{wh}^{(r)} \mid y_{wh}, \mathbf{x}) \cdot p(y_{wh} \mid \mathbf{x}) \tag{2}$$

where $p(y_{wh} \mid \mathbf{x})$ denotes the GT label distribution over the $(w,h)^{\text{th}}$ pixel in the image $\mathbf{x}$, and $p(\tilde{y}_{wh}^{(r)} \mid y_{wh}, \mathbf{x})$ describes the noisy labelling process by which annotator r corrupts the true segmentation label. In particular, we refer to the $L\times L$ matrix whose each $(i,j)^{\text{th}}$ element is defined by the second term $\mathbf{a}^{(r)}(\mathbf{x},w,h)_{ij} := p(\tilde{y}_{wh}^{(r)} = i \mid y_{wh} = j, \mathbf{x})$ as the CM of annotator r at pixel (w,h) in image $\mathbf{x}$.

We introduce a CNN-based architecture which models the different constituents in the above joint probability distribution $p(\{\tilde{\mathbf{y}}^{(r)}\}_{r\in S(\mathbf{x})} \mid \mathbf{x})$ as illustrated in Fig. 1. The model consists of two components: (1) *Segmentation Network*, parametrised by θ, which estimates the GT segmentation probability map, $\hat{\mathbf{p}}_\theta(\mathbf{x}) \in \mathbb{R}^{W\times H\times L}$ whose each $(w,h,i)^{\text{th}}$ element approximates $p(y_{wh} = i \mid \mathbf{x})$;(2) *Annotator Network*, parametrised by ϕ, that generate estimates of the pixel-wise CMs of respective annotators as a function of the input image, $\{\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x}) \in [0,1]^{W\times H\times L\times L}\}_{r=1}^{R}$ whose each $(w,h,i,j)^{\text{th}}$ element approximates $p(\tilde{y}_{wh}^{(r)} = i \mid y_{wh} = j, \mathbf{x})$. Each product $\hat{\mathbf{p}}_\phi^{(r)}(\mathbf{x}) := \hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x}) \cdot \hat{\mathbf{p}}_\theta(\mathbf{x})$ represents the estimated segmentation probability map of the corresponding annotator. Note that here "$\cdot$" denotes the element-wise matrix multiplications in the spatial dimensions W,H. At inference time, we use the output of the segmentation network $\hat{\mathbf{p}}_\theta(\mathbf{x})$ to segment test images.

We note that each spatial CM $\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x})$ contains WHL^2 variables, and calculating the corresponding annotator's prediction $\hat{\mathbf{p}}_\phi^{(r)}(\mathbf{x})$ requires $WH(2L-1)L$ floating-point operations, potentially incurring a large time/space cost when the number of classes is large. Although not the focus of this work (as we are concerned with medical imaging applications for which the number of classes are mostly limited to less than 10), we also consider a low-rank approximation (rank$=1$) scheme to alleviate this issue wherever appropriate. More details are provided in the supplementary.

3.3 Learning Spatial Confusion Matrices and True Segmentation

Next, we describe how we jointly optimise the parameters of segmentation network, θ and the parameters of annotator network, ϕ. In short, we minimise the negative log-likelihood of the probabilistic model plus a regularisation term via stochastic gradient descent. A detailed description is provided below.

Given training input $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ and noisy labels $\tilde{\mathbf{Y}}^{(r)} = \{\tilde{\mathbf{y}}_n^{(r)} : r \in S(\mathbf{x}_n)\}_{n=1}^N$ for $r = 1, ..., R$, we optimaize the parameters $\{\theta, \phi\}$ by minimizing the negative log-likelihood (NLL), $-\log p(\tilde{\mathbf{Y}}^{(1)},...,\tilde{\mathbf{Y}}^{(R)}|\mathbf{X})$. From eqs. (1) and (2), this optimization objective equates to the sum of cross-entropy losses between the observed noisy segmentations and the estimated annotator label distributions:

$$-\log p(\tilde{\mathbf{Y}}^{(1)},...,\tilde{\mathbf{Y}}^{(R)}|\mathbf{X}) = \sum_{n=1}^{N}\sum_{r=1}^{R} \mathbb{1}(\tilde{\mathbf{y}}_n^{(r)} \in \mathcal{S}(\mathbf{x}_n)) \cdot \mathrm{CE}\big(\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x}) \cdot \hat{\mathbf{p}}_\theta(\mathbf{x}_n), \tilde{\mathbf{y}}_n^{(r)}\big) \tag{3}$$

Minimizing the above encourages each annotator-specific prediction $\hat{\mathbf{p}}^{(r)}(\mathbf{x}) := \hat{\mathbf{A}}_\phi^{(r)} \hat{\mathbf{p}}_\theta(\mathbf{x})$ to be as close as possible to the true noisy label distribution of the annotator $\mathbf{p}^{(r)}(\mathbf{x})$. However, this loss function alone is not capable of separating the annotation noise from the true label distribution; there are many combinations of pairs $\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x})$ and segmentation model $\hat{\mathbf{p}}_\theta(\mathbf{x})$ such that $\hat{\mathbf{p}}^{(r)}(\mathbf{x})$ perfectly matches the true annotator's distribution $\mathbf{p}^{(r)}(\mathbf{x})$ for any input $\mathbf{x}$ (e.g., permutation of rows in the CMs). To combat this problem, inspired by Tanno *et al.*[28], which addressed an analogous issue for the classification task, we add the trace of the estimated CMs to the loss function in Eq. (3) as a regularisation term (see Sec 3.4). We thus optimize the combined loss:

$$\sum_{n=1}^{N}\sum_{r=1}^{R} \mathbb{1}(\tilde{\mathbf{y}}_n^{(r)} \in \mathcal{S}(\mathbf{x}_i)) \cdot \Big[\mathrm{CE}\big(\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x}) \cdot \hat{\mathbf{p}}_\theta(\mathbf{x}_n), \tilde{\mathbf{y}}_n^{(r)}\big) + \lambda \cdot \mathrm{tr}\big(\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x}_n)\big)\Big] \tag{4}$$

where $\mathcal{S}(\mathbf{x}))$ denotes the set of all labels available for image $\mathbf{x}$, and $\mathrm{tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. The mean trace represents the average probability that a randomly selected annotator provides an accurate label. Intuitively, minimising the trace encourages the estimated annotators to be maximally unreliable while minimising the cross entropy ensures fidelity with observed noisy annotators. We minimise this combined loss via stochastic gradient descent to learn both $\{\theta,\phi\}$.

3.4 Justification for the Trace Norm

Here we provide a further justification for using the trace regularisation. Tanno *et al.*[28] showed that if the average CM of annotators is *diagonally dominant*, and the cross-entropy term in the loss function is zero, minimising the trace of the estimated CMs uniquely recovers the true CMs. However, their results concern properties of the average CMs of both the annotators and the classifier over the data population, rather than individual data samples. We show a similar but slightly weaker result in the sample-specific regime, which is more relevant as we estimate CMs of respective annotators on every input image.

First, let us set up the notations. For brevity, for a given input image $\mathbf{x} \in \mathbb{R}^{W \times H \times C}$, we denote the estimated CM of annotator r at $(i,j)^{\text{th}}$ pixel by $\hat{\mathbf{A}}^{(r)} := [\mathbf{A}^{(r)}(\mathbf{x})_{ij}] \in [0,1]^{L \times L}$. We also define the mean CM $\mathbf{A}^* := \sum_{r=1}^R \pi_r \hat{\mathbf{A}}^{(r)}$ and its estimate $\hat{\mathbf{A}}^* := \sum_{r=1}^R \pi_r \hat{\mathbf{A}}^{(r)}$ where $\pi_r \in [0,1]$ is the probability that the annotator r labels image $\mathbf{x}$. Lastly, as we stated earlier, we assume there is a single GT segmentation label per image — thus the true L-dimensional probability vector at pixel (i,j) takes the form of a one-hot vector i.e., $\mathbf{p}(\mathbf{x}) = \mathbf{e}_k$ for, say, class $k \in [1,...,L]$. Then, the followings result motivates the use of the trace regularisation:

Theorem 1. *If the annotator's segmentation probabilities are perfectly modelled by the model for the given image i.e., $\hat{\boldsymbol{A}}^{(r)}\hat{\boldsymbol{p}}_\theta(\boldsymbol{x}) = \boldsymbol{A}^{(r)}\boldsymbol{p}(\boldsymbol{x}) \forall r = 1, ..., R$, and the average true CM $\boldsymbol{A}^*$ at a given pixel and its estimate $\hat{\boldsymbol{A}}^*$ are diagonally dominant ($a^*_{ii} > a^*_{ij}$, $\hat{a}^*_{ii} > \hat{a}^*_{ij}$ for all $i \neq j$), then $\boldsymbol{A}^{(1)},...,\boldsymbol{A}^{(R)} = \mathit{argmin}_{\hat{\boldsymbol{A}}^{(1)},...,\hat{\boldsymbol{A}}^{(R)}} \Big[tr(\hat{\boldsymbol{A}}^*)\Big]$ and such solutions are **unique** up to the k^{th} column where k is the correct pixel class.*

The corresponding proof is provided in the supplementary material. The above result shows that if each estimated annotator's distribution $\hat{\mathbf{A}}^{(r)}\hat{\mathbf{p}}_\theta(\mathbf{x})$ is very close to the true noisy distribution $\mathbf{p}^{(r)}(\mathbf{x})$

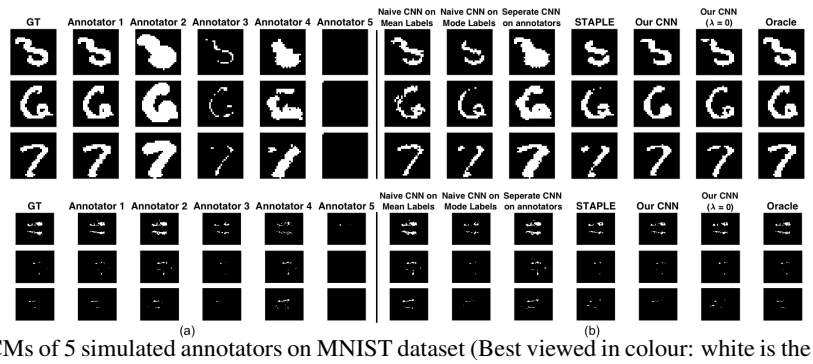

Figure 2: CMs of 5 simulated annotators on MNIST dataset (Best viewed in colour: white is the true positive, green indicates the false negative, red is the false positive and black is the true negative).

(which is encouraged by minimizing the cross-entropy loss), and for a given pixel, the average CM has diagonal entries larger than any other entries in each row [1], then minimizing its trace will drive the estimates of the k^{th} ('correct class') columns in the respective annotator's CMs to match the true values. Although this result is weaker than what was shown in [28] for the population setting rather than the individual samples, the single-ground-truth assumption means that the remaining values of the CMs are uniformly equal to $1/L$, and thus it suffices to recover the column of the correct class.

To encourage $\{\hat{\mathbf{A}}^{(1)},...,\hat{\mathbf{A}}^{(R)}\}$ to be also diagonally dominant, we initialize them with identity matrices by training the *annotation network* to maximise the trace for sufficient iterations as a warm-up period. Intuitively, the combination of the trace term and cross-entropy separates the true distribution from the annotation noise by finding the maximal amount of confusion which explains the noisy observations well.

4 Experiments

We evaluate our method on a variety of datasets including both synthetic and real-world scenarios:1) for MNIST segmentation and ISBI2015 MS lesion segmentation challenge dataset [38], we apply morphological operations to generate synthetic noisy labels in binary segmentation tasks; 2) for BraTS 2019 dataset [4], we apply similar simulation to create noisy labels in a multi-class segmentation task; 3) we also consider the LIDC-IDRI dataset which contains multiple annotations per input acquired from different clinical experts as the evaluation in practice. Details of noisy label simulation can be found in Appendix A.1.

Our experiments are based on the assumption that no ground-truth (GT) label is not known a priori, hence, we compare our method against multiple label fusion methods. IN particular, we consider four label fusion baselines: a) mean of all of the noisy labels; b) mode labels by taking the "majority vote"; c) label fusion via the original STAPLE method [9]; d) Spatial STAPLE, a more recent extension of c) that accounts for spatial variations in CMs. After curating the noisy annotations via above methods, we train the segmentation network and report the results. For c) and d), we used the toolkit[2]. In addition, we also include a recent method called Probabilistic U-net as another baseline, which has been shown to capture inter-reader variations accurately. The details are presented in Appendix A.2.

For evaluation metrics, we use: 1) root-MSE between estimated CMs and real CMs; 2) Dice coefficient (DICE) between estimated segmentation and true segmentation; 3) The generalized energy distance proposed in [24] to measure the quality of the estimated annotator's labels.

4.1 MNIST and MS lesion segmentation datasets

MNIST dataset consists of 60,000 training and 10,000 testing examples, all of which are 28×28 grayscale images of digits from 0 to 9, and we derive the segmentation labels by thresholding the intensity values at 0.5. The MS dataset is publicly available and comprises 21 3D scans from 5 subjects. All scans are split into 10 for training and 11 for testing. We hold out 20% of training images as a validation set for both datasets. On both datasets, our proposed model achieves a higher dice similarity coefficient than STAPLE on the dense label case and, even more prominently, on the single label (i.e., 1 label per image) case (shown in Tables. 1&2 and Fig. 2). In addition, our model outperforms

[1] For the standard "majority vote" or the mean label to capture the correct true labels, one requires each diagonal element in the average CM to be larger than the sum of the remaining elements in the same row, which is a more strict condition.

[2] https://www.nitrc.org/projects/masi-fusion/

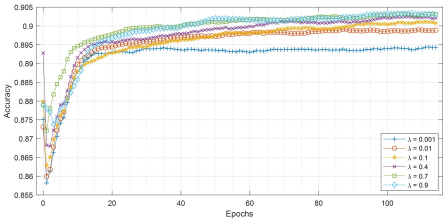

Figure 3: Curves of validation accuracy during training of our model for a range of hyperparameters. For our method, the scaling of trace regularizer is varied in [0.001, 0.01, 0.1, 0.4, 0.7, 0.9].)

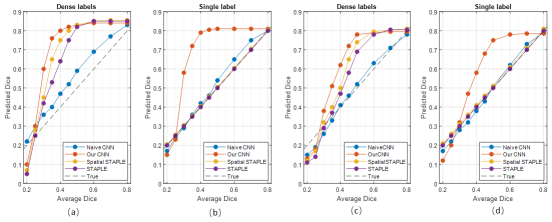

Figure 4: Segmentation accuracy of different models on MNIST (a, b) and MS (c, d) dataset for a range of annotation noise (measured in averaged Dice with respect to GT.

STAPLE without or with trace norm, in terms of CM estimation, specifically, we could achieve an increase at 6.3%. Additionally, we include the performance on different regularisation coefficient, which is presented in Fig. 3. Fig. 4 compares the segmentation accuracy on MNIST and MS lesion for a range of average dice where labels are generated by a group of 5 simulated annotators. Fig. 5 illustrates our model can capture the patterns of mistakes for each annotator.

Models	MNIST DICE (%) (testing)	MNIST CM estimation (validation)	MSLesion DICE (%) (testing)	MSLesion CM estimation (validation)
Naive CNN on mean labels	38.36 ± 0.41	n/a	46.55 ± 0.53	n/a
Naive CNN on mode labels	62.89 ± 0.63	n/a	47.82 ± 0.76	n/a
Probabilistic U-net [24]	65.12 ± 0.83	n/a	46.15 ± 0.59	n/a
Separate CNNs on annotators	70.44 ± 0.65	n/a	46.84 ± 1.24	n/a
STAPLE [9]	78.03 ± 0.29	0.1241 ± 0.0011	55.05 ± 0.53	0.1502 ± 0.0026
Spatial STAPLE [14]	78.96 ± 0.22	0.1195 ± 0.0013	58.37 ± 0.47	0.1483 ± 0.0031
Ours without Trace	79.63 ± 0.53	0.1125 ± 0.0037	65.77 ± 0.62	0.1342 ± 0.0053
Ours	82.92 ± 0.19	0.0893 ± 0.0009	67.55 ± 0.31	0.0811 ± 0.0024
Oracle (Ours but with known CMs)	83.29 ± 0.11	0.0238 ± 0.0005	78.86 ± 0.14	0.0415 ± 0.0017

Table 1: Comparison of segmentation accuracy and error of CM estimation for different methods with dense labels (mean $\pm$ standard deviation).

Models	MNIST DICE (%) (testing)	MNIST CM estimation (validation)	MSLesion DICE (%) (testing)	MSLesion CM estimation (validation)
Naive CNN on mean & mode labels	32.79 ± 1.13	n/a	27.41 ± 1.45	n/a
STAPLE [9]	54.07 ± 0.68	0.2617 ± 0.0064	35.74 ± 0.84	0.2833 ± 0.0081
Spatial STAPLE [14]	56.73 ± 0.53	0.2384 ± 0.0061	38.21 ± 0.71	0.2591 ± 0.0074
Ours without Trace	74.48 ± 0.37	0.1538 ± 0.0029	54.76 ± 0.66	0.1745 ± 0.0044
Ours	76.48 ± 0.25	0.1329 ± 0.0012	56.43 ± 0.47	0.1542 ± 0.0023

Table 2: Comparison of segmentation accuracy and error of CM estimation for different methods with one label per image (mean $\pm$ standard deviation).

Generalised Energy Distance (Dice)	MNIST	MS	BraTS	LIDC-IDRI
Probabilistic U-net [24]	1.46 ± 0.04	1.91 ± 0.03	3.23 ± 0.07	1.97 ± 0.03
Ours	$\mathbf{1.24 \pm 0.02}$	$\mathbf{1.67 \pm 0.03}$	$\mathbf{3.14 \pm 0.05}$	$\mathbf{1.87 \pm 0.04}$

Table 3: Comparison of Generalised Energy Distance on different datasets (mean $\pm$ standard deviation). The distance metric used here is Dice.

4.2 BraTS Dataset and LIDC-IDRI Dataset

We also evaluate our model on a multi-class segmentation task, using all of the 259 high grade glioma (HGG) cases in training data from 2019 multi-modal Brain Tumour Segmentation Challenge (BraTS). We extract each slice as 2D images and split them at case-wise to have, 1600 images for training, 300 for validation and 500 for testing. Pre-processing includes: concatenation of all of available modalities; centre cropping to 192 x 192; normalisation for each case at each modality. To create synthetic noisy labels in multi-class scenario, we first choose a target class and then apply morphological operations on the provided GT mask to create 4 synthetic noisy labels at different patterns, namely, over-segmentation, under-segmentation, wrong segmentation and good segmentation. Details of noisy label simulation are in Appendix A.3.

The LIDC-IDRI dataset contains 1018 lung CT scans from 1010 lung patients with manual lesion segmentations from four experts. For each scan, 4 radiologists provided annotation masks for lesions that they independently detected and considered to be abnormal. For our experiments, we use the same method in [24] to pre-process all scans. We split the dataset at case-wise into a training (722 patients),

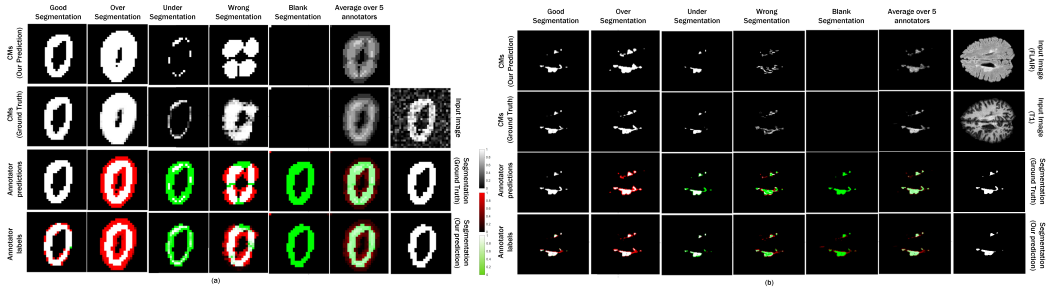

Figure 5: Visualisation of estimated true labels and confusion matrices on MNIST/MS datasets (Best viewed in colour: white is the true positive, green is the false negative, red is the false positive and black is the true negative).

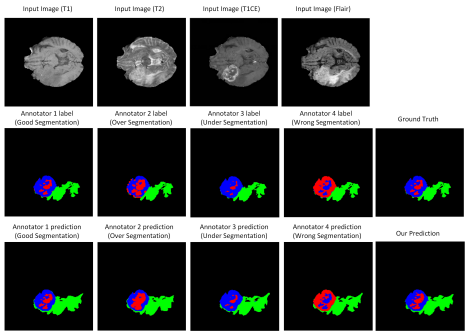

Figure 6: The final segmentation of our model on BraTS and each annotator network predictions visualization. (Best viewed in colour: the target label is red.)

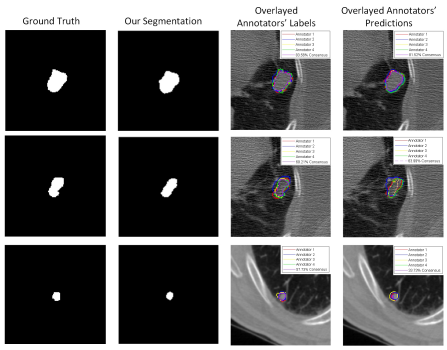

Figure 7: Segmentation results on LIDC-IDRI dataset and the visualization of each annotator contours and the consensus.

validation (144 patients) and testing (144 patients). We then resampled the CT scans to $1mm \times 1mm$ in-plane resolution. We also centre cropped 2D images (180×180 pixels) around lesion positions, in order to focus on the annotated lesions. The lesion positions are those where at least one of the experts segmented a lesion. We hold 5000 images in the training set, 1000 images in the validation set and 1000 images in the test set.

On both BraTS and LIDC-IDRI dataset, our proposed model achieves a higher dice similarity coefficient than STAPLE and Spatial STAPLE on both of the dense labels and single label scenarios (shown in Table. 4 and Table. 5 in Appendix A.3). In addition, our model (with trace) outperforms STAPLE in terms of CM estimation by a large margin at 14.4% on BraTS. In Fig. 6, we visualized the segmentation results on BraTS and the corresponding annotators' predictions. Fig. 7 presents three examples of the segmentation results and the corresponding four annotator contours, as well as the consensus. As shown in both figures, our model successfully predicts the both the segmentation of lesions and the variations of each annotator in different cases. Additionally, as shown in Table.3, our model consistently outperforms Probabilistic U-Net on generalized energy distance across the four test different datasets, which indicates that our method is better at capturing the inter-annotator variability than the baseline Probabilistic U-Net.

5 Conclusion

We introduced the first learning method based on CNNs for simultaneously recovering the label noise of multiple annotators and the GT label distribution for supervised segmentation problems. We demonstrated this method on real-world datasets with synthetic annotations and real-world annotations. Our method is capable of estimating individual annotators and thereby improving robustness against label noise. Experiments have shown our model achieves considerable improvement over the traditional label fusion approaches including averaging, the majority vote and the widely used STAPLE framework and spatially varying versions, in terms of both segmentation accuracy and the quality of CM estimation.

In the future, we plan to accommodate meta-information of annotators (e.g., number of years of experience), and non-image data (e.g., genetics) that may influence the pattern of the underlying segmentation label such as lesion appearance, in our framework. We are also interested in assessing the downstream utility of our approach in active data collection schemes where the segmentation model $\hat{\mathbf{p}}_\theta(\mathbf{x})$ is used to select which samples to annotate ("active learning"), and the annotator models $\{\hat{\mathbf{A}}_\phi^{(r)}(\mathbf{x})\}_{r=1}^R$ are used to decide which experts to label them ("active labelling").

Boarder Impact Statement

Image segmentation has been one of the main challenges in modern medical image analysis, and describes the process of assigning each pixel or voxel in images with biologically meaningful discrete labels, such as anatomical structures and tissue types (e.g. pathology and healthy tissues). The task is required in many clinical and research applications, including surgical planning [39, 40], and the study of disease progression, aging or healthy development [41, 42, 43]. However, there are often cases in practice where the correct delineation of structures is challenging; this is also reflected in the well-known presence of high inter- and intra-reader variability in segmentation labels obtained from trained experts [9, 23, 5].

Although expert manual annotations of lesions is feasible in practice, this task is time consuming. It usually takes 1.5 to 2 hours to label a MS patient with average 3 visit scans. Meanwhile, the long-established gold standard for segmentation of medical images has been manual voxel-by-voxel labeling by an expert anatomist. Unfortunately, this process is fraught with both interand intra-rater variability (e.g., on the order of approximately 10% by volume [44, 45]). Thus, developing an automatic segmentation technique to fix the variability among inter- and intra-readers could be meaningful not only in terms of the accuracy in delineating MS lesions but also in the related reductions in time and economic costs derived from manual lesion labeling. The lack of consistency in labelling is also common to see in other medical imaging applications, e.g., in lung abnormalities segmentation from CT images. A lesion might be clearly visible by one annotator, but the information about whether it is cancer tissue or not might not be clear to others. However, a potential point of criticism could be that our work in the current form has only been demonstrated on medical images. We would like to convince AC/PCs that the medical imaging domain alone offers a considerably broad range of opportunities for impact; e.g., diagnosis/prognosis in radiology, surgical planning and study of disease progression and treatment, etc. In addition, the annotator information could be potentially utilised for the purpose of education. Another potential opportunity is to integrate such information into the data/label acquisition scheme in order to train reliable segmentation algorithms in a data-efficient manner.

References

[1] Elizabeth Lazarus, Martha B Mainiero, Barbara Schepps, Susan L Koelliker, and Linda S Livingston. Bi-rads lexicon for us and mammography: interobserver variability and positive predictive value. *Radiology*, 239(2):385–391, 2006.

[2] Takeyuki Watadani, Fumikazu Sakai, Takeshi Johkoh, Satoshi Noma, Masanori Akira, Kiminori Fujimoto, Alexander A Bankier, Kyung Soo Lee, Nestor L Müller, Jae-Woo Song, et al. Interobserver variability in the ct assessment of honeycombing in the lungs. *Radiology*, 266(3):936–944, 2013.

[3] Andrew B Rosenkrantz, Ruth P Lim, Mershad Haghighi, Molly B Somberg, James S Babb, and Samir S Taneja. Comparison of interreader reproducibility of the prostate imaging reporting and data system and likert scales for evaluation of multiparametric prostate mri. *American Journal of Roentgenology*, 201(4):W612–W618, 2013.

[4] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[5] Leo Joskowicz, D Cohen, N Caplan, and J Sosna. Inter-observer variability of manual contour delineation of structures in ct. *European radiology*, 29(3):1391–1399, 2019.

[6] Huahong Zhang, Alessandra M Valcarcel, Rohit Bakshi, Renxin Chu, Francesca Bagnato, Russell T Shinohara, Kilian Hett, and Ipek Oguz. Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 338–346. Springer, 2019.

[7] Eytan Kats, Jacob Goldberger, and Hayit Greenspan. A soft staple algorithm combined with anatomical knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 510–517. Springer, 2019.

[8] Hugh Harvey and Ben Glocker. A standardised approach for preparing imaging data for machine learning tasks in radiology. In *Artificial Intelligence in Medical Imaging*, pages 61–72. Springer, 2019.

[9] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

[10] Stefan Winzeck, Arsany Hakim, Richard McKinley, José AADSR Pinto, Victor Alves, Carlos Silva, Maxim Pisov, Egor Krivov, Mikhail Belyaev, Miguel Monteiro, et al. Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral mri. *Frontiers in neurology*, 9:679, 2018.

[11] Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports*, 8(1):1–17, 2018.

[12] Gleason 2019 challenge. `https://gleason2019.grand-challenge.org/Home/`. Accessed: 2020-02-30.

[13] Andrew J Asman and Bennett A Landman. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (collate). *IEEE transactions on medical imaging*, 30(10):1779–1794, 2011.

[14] Andrew J Asman and Bennett A Landman. Formulating spatially varying performance in the statistical fusion framework. *IEEE transactions on medical imaging*, 31(6):1326–1336, 2012.

[15] Juan Eugenio Iglesias, Mert Rory Sabuncu, and Koen Van Leemput. A unified framework for cross-modality multi-atlas segmentation of brain mri. *Medical image analysis*, 17(8):1181–1191, 2013.

[16] M Jorge Cardoso, Kelvin Leung, Marc Modat, Shiva Keihaninejad, David Cash, Josephine Barnes, Nick C Fox, Sebastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Steps: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Medical image analysis*, 17(6):671–684, 2013.

[17] Andrew J Asman and Bennett A Landman. Non-local statistical label fusion for multi-atlas segmentation. *Medical image analysis*, 17(2):194–208, 2013.

[18] Alireza Akhondi-Asl, Lennox Hoyte, Mark E Lockhart, and Simon K Warfield. A logarithmic opinion pool based staple algorithm for the fusion of segmentations with associated reliability weights. *IEEE transactions on medical imaging*, 33(10):1997–2009, 2014.

[19] Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20, 2019.

[20] Martin Styner, Joohwi Lee, Brian Chin, M Chin, Olivier Commowick, H Tran, S Markovic-Plese, V Jewells, and S Warfield. 3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation. *Midas Journal*, 2008:1–6, 2008.

[21] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[22] Neil I Weisenfeld and Simon K Warfield. Learning likelihoods for labeling (l3): a general multi-classifier segmentation algorithm. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 322–329. Springer, 2011.

[23] Leo Joskowicz, D Cohen, N Caplan, and Jacob Sosna. Automatic segmentation variability estimation with segmentation priors. *Medical image analysis*, 50:54–64, 2018.

[24] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018.

[25] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019.

[26] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[27] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.

[28] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. *arXiv preprint arXiv:1902.03680*, 2019.

[29] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

[30] Padhraic Smyth, Usama M Fayyad, Michael C Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in neural information processing systems*, pages 1085–1092, 1995.

[31] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043, 2009.

[32] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432, 2010.

[33] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12):1428–1436, 2013.

[34] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual international conference on machine learning*, pages 889–896. ACM, 2009.

[35] Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy. Modeling annotator expertise: Learning when everybody knows a bit of something. In *AISTATs*, pages 932–939, 2010.

[36] Steve Branson, Grant Van Horn, and Pietro Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7474–7483, 2017.

[37] Grant Van Horn, Steve Branson, Scott Loarie, Serge Belongie, Cornell Tech, and Pietro Perona. Lean multiclass crowdsourcing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2723, 2018.

[38] Andrew Jesson and Tal Arbel. Hierarchical mrf and random forest segmentation of ms lesions and healthy tissues in brain mri. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[39] David T Gering, Arya Nabavi, Ron Kikinis, Noby Hata, Lauren J O'Donnell, W Eric L Grimson, Ferenc A Jolesz, Peter M Black, and William M Wells III. An integrated visualization system for surgical planning and guidance using image fusion and an open mr. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 13(6):967–975, 2001.

[40] Gloria P Mazzara, Robert P Velthuizen, James L Pearlman, Harvey M Greenberg, and Henry Wagner. Brain tumor target volume determination for radiation treatment planning through automated mri segmentation. *International Journal of Radiation Oncology* Biology* Physics*, 59(1):300–312, 2004.

[41] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[42] Marcel Prastawa, John H Gilmore, Weili Lin, and Guido Gerig. Automatic segmentation of mr images of the developing newborn brain. *Medical image analysis*, 9(5):457–466, 2005.

[43] Alex P Zijdenbos, Reza Forghani, Alan C Evans, et al. Automatic" pipeline" analysis of 3-d mri data for clinical trials: application to multiple sclerosis. *IEEE Trans. Med. Imaging*, 21(10):1280–1291, 2002.

[44] Edward A Ashton, Chihiro Takahashi, Michel J Berg, Andrew Goodman, Saara Totterman, and Sven Ekholm. Accuracy and reproducibility of manual and semiautomated quantification of ms lesions by mri. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 17(3):300–308, 2003.

[45] Bonnie N Joe, Melanie B Fukui, Carolyn Cidis Meltzer, Qing-shou Huang, Roger S Day, Phil J Greer, and Michael E Bozik. Brain tumor volume measurement: comparison of manual and semiautomated methods. *Radiology*, 212(3):811–816, 1999.

[46] Sergi Valverde, Mostafa Salem, Mariano Cabezas, Deborah Pareto, Joan C. Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, page 101638, 2018.

[47] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

A Additional results

A.1 Data Simulation

We generate synthetic annotations from an assumed GT to generate efficacy of the approach in an idealised situation where the GT is known. We simulate a group of 5 annotators of disparate characteristics by performing morphological transformations (e.g., thinning, thickening, fractures, etc) on the ground-truth (GT) segmentation labels, using Morpho-MNIST software [19]. In particular, the first annotator provides faithful segmentation ("good-segmentation") with approximate GT, the second tends over-segment ("over-segmentation"), the third tends to under-segment ("under-segmentation"), the fourth is prone to the combination of small fractures and over-segmentation ("wrong-segmentation") and the fifth always annotates everything as the background ("blank-segmentation"). We create training data by deriving labels from the simulated annotators.

A.2 MNIST and MS Dataset

We examine the ability of our method to learn the CMs of annotators and the true label distribution. We compared the performance of our method against several baselines and the original STAPLE algorithm [9] and Spatial STAPLE [14]. The first baseline is the naive CNN trained on the mean labels and the majority vote labels across the 5 annotators. The second baseline is the separate CNNs trained on 5 annotator labels and evaluate on their mean output. The "oracle" model is the idealistic scenario where CMs of the annotators are a priori known to the model while "annotators" indicate the average labeling accuracy of each annotator group. All the baselines and the annotator CNN, the segmentation CNN in our model are implemented with the NicMSlesions architecture described in [46]. We also evaluate on the validation set the effects of regularisation coefficient $\lambda \in \{0, 0.001, 0.01, 0.1, 0.4, 0.7, 0.9\}$ of the trace-norm in Eq. 4 on the accuracy of segmentation and CM estimation. Results are shown in Fig. 3.

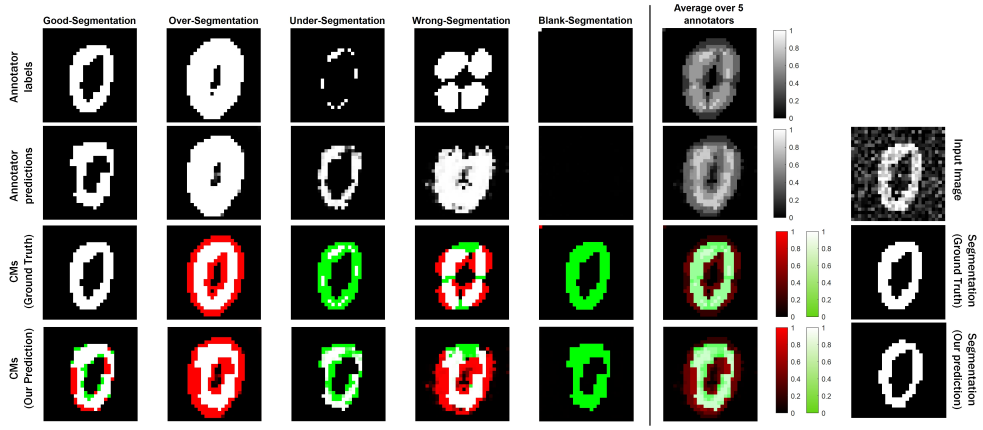

Figure 7: Visualisation of estimated true labels and confusion matrices on MNIST datasets (Best viewed in colour: white is the true positive, green is the false negative, red is the false positive and black is the true negative).

A.3 BraTS and LIDC-IDRI

We also evaluate our model on a multi-class segmentation task, using training data from 2019 Brain Tumour Segmentation Challenge (BraTS). In training data of BraTS 2019, there are 259 cases with high grade (HG) and 76 cases with low grade (LG) glioma. For each case, four MRI modalities are available, FLAIR, T1, T1-contrast and T2. The datasets are pre-processed by the organizers and co-registered to the same anatomical template, interpolated to the same resolution (1 mm^3) and skull-stripped. We centre cropped 2D images (192 $\times$ 192 pixels) and hold 1600 2D images for training, 300 images for validation, 500 images for testing, we apply Gaussian normalization on each case of each modality, to have zero-mean and unit variance. Fig. 6 shows one such tumor case in four different modality. To create synthetic noisy labels in multi-class scenario, we first choose a target class and then apply morphological operations on the provided GT mask to create 4 synthetic noisy labels at different patterns, namely, over-segmentation, under-segmentation, wrong segmentation and good segmentation. Details of noisy label simulation are in Appendix A.3.

The LIDC-IDRI dataset contains 1018 lung CT scans from 1010 lung patients with manual lesion segmentations from four experts. For each scan, 4 radiologists provided annotation masks for lesions that they independently detected and considered to be abnormal. For our experiments, we use the same method in [24] to pre-process all scans. We split the dataset at case-wise into a training (722 patients), validation (144 patients) and testing (144 patients). We then resampled the CT scans to $1mm \times 1mm$ in-plane resolution. We also centre cropped 2D images (180×180 pixels) around lesion positions, in order to focus on the annotated lesions. The lesion positions are those where at least one of the experts segmented a lesion. We hold 5000 images in the training set, 1000 images in the validation set and 1000 images in the test set.

On both BraTS and LIDC-IDRI dataset, our proposed model achieves a higher dice similarity coefficient than STAPLE on both of the dense labels and single label scenarios (shown in Table. 4 and Table. 5 in Appendix A.3). In addition, our model (with trace) outperforms STAPLE in terms of CM estimation by a large margin at 14.4% on BraTS. In Fig. 6, we visualized the segmentation results on BraTS and the corresponding annotators' predictions. Fig. 7 presents three examples of the segmentation results and the corresponding four annotator contours, as well as the consensus. As shown in both figures, our model successfully predicts the both the segmentation of lesions and the variations of each annotator in different cases. Additionally, as shown in Table.3, our model consistently outperforms Probabilistic U-Net on generalized energy distance across the four test different datasets, which indicates that our method is better at capturing the inter-annotator variability than the baseline Probabilistic U-Net.

Models	BraTS DICE (%) (testing)	BraTS CM estimation (validation)	LIDC-IDRI DICE (%) (testing)	LIDC-IDRI CM estimation (validation)
Naive CNN on mean labels	29.42 ± 0.58	n/a	56.72 ± 0.61	n/a
Naive CNN on mode labels	34.12 ± 0.45	n/a	58.64 ± 0.47	n/a
Probabilistic U-net [24]	40.53 ± 0.75	n/a	61.26 ± 0.69	n/a
STAPLE [9]	46.73 ± 0.17	0.2147 ± 0.0103	69.34 ± 0.58	0.0832 ± 0.0043
Spatial STAPLE [14]	47.31 ± 0.21	0.1871 ± 0.0094	70.92 ± 0.18	0.0746 ± 0.0057
Ours without Trace	49.03 ± 0.34	0.1569 ± 0.0072	71.25 ± 0.12	0.0482 ± 0.0038
Ours	53.47 ± 0.24	0.1185 ± 0.0056	74.12 ± 0.19	0.0451 ± 0.0025
Oracle (Ours but with known CMs)	67.13 ± 0.14	0.0843 ± 0.0029	79.41 ± 0.17	0.0381 ± 0.0021

Table 4: Comparison of segmentation accuracy and error of CM estimation for different methods with dense labels (mean $\pm$ standard deviation). (ryu): On LIDC-IDRI, it is rather surprising that our method performs better than SpatialSTAPLE when the GT were created by SpatialSTAPLE! We should mention this clearly in the results section.

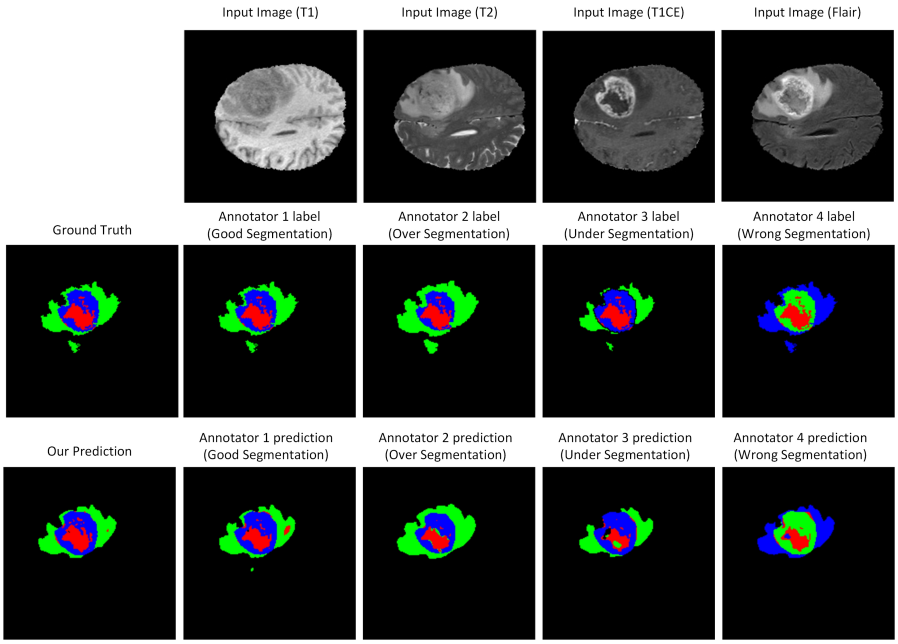

Figure 8: The final segmentation of our model on BraTS and each annotator network predictions visualization. (Best viewed in colour: the target label is red.)

Models	BraTS DICE (%) (testing)	BraTS CM estimation (validation)	LIDC-IDRI DICE (%) (testing)	LIDC-IDRI CM estimation (validation)
Naive CNN on mean & mode labels	36.12 ± 0.93	n/a	48.36 ± 0.79	n/a
STAPLE [9]	38.74 ± 0.85	0.2956 ± 0.1047	57.32 ± 0.87	0.1715 ± 0.0134
Spatial STAPLE [14]	41.59 ± 0.74	0.2543 ± 0.0867	62.35 ± 0.64	0.1419 ± 0.0207
Ours without Trace	43.74 ± 0.49	0.1825 ± 0.0724	66.95 ± 0.51	0.0921 ± 0.0167
Ours	46.21 ± 0.28	0.1576 ± 0.0487	68.12 ± 0.48	0.0587 ± 0.0098

Table 5: Comparison of segmentation accuracy and error of CM estimation for different methods with one label per image (mean $\pm$ standard deviation).

A.4 Low-rank approximation

In particular, we parametrise the spatial CM $\hat{\mathbf{A}}_{\phi}^{(r)}(\mathbf{x}) = \mathbf{B}_{1,\phi}^{(r)}(\mathbf{x}) \cdot \mathbf{B}_{2,\phi}^{T,(r)}(\mathbf{x})$ where both $\mathbf{B}_{1,\phi}^{(r)}$ and $\mathbf{B}_{2,\phi}^{(r)}$ are smaller matrices of size $W \times H \times L \times l$ where $l << L$. Two separate rectangular matrices are used since the confusion matrices are not necessarily symmetric. Such low-rank approximation reduces the total number of variables to $2WHLl$ and the FLOPs to $WH(4L(l-0.25)-l)$. Still need to decide whether to include this paragraph depending on the results on DICE.

Rank	Dice	CM estimation	GPU Memory	No. Parameters & FLOPS
Default	53.47 ± 0.24	0.1185 ± 0.0056	2.68GB	192×192
rank 1	50.56 ± 2.00	-	2.57GB	

Table 6: Segmentation performance of low-rank approximation on BraTS. GPU memory is when batch size 1 is used (mean $\pm$ standard deviation).

A.5 Algorithm

(Copy from CVPR): Here we provide pseudo-codes of our method (Algorithm 1), generalized EM [34] (Algorithm 2) and model-bootstrapped EM [27] (Algorithm 3) to clarify the differences between different methods for jointly learning the true label distribution and confusion matrices of annotators in eq. 2 in the main text. Given the training set $\mathcal{D} = \{\mathbf{x}_n, \tilde{y}_n^{(1)}, ..., \tilde{y}_n^{(R)}\}_{n=1}^{N}$, each example may not be labelled by

all the annotators. In such cases, for ease of notation, we assign pseudo class $\tilde{y}_n^{(r)} = -1$ to fill the missing labels. The comparison between these three algorithms illustrates the implementational simplicity of our method, despite the comparable or superior performance demonstrated on all three datasets.

Algorithm 1 Our method

Inputs: $\mathcal{D} = \{\mathbf{x}_n, \tilde{y}_n^{(1)}, \ldots, \tilde{y}_n^{(R)}\}_{n=1}^N$, λ: scale of trace regularizer

Initialize the confusion matrices $\{\hat{\mathbf{A}}^{(r)}\}_{r=1}^R$ **to identity matrices**

Initialize the parameters of the base classifier θ

Learn θ **and** $\{\hat{\mathbf{A}}^{(r)}\}_{r=1}^R$ **by performing minibatch SGD on the combined loss**:

$$\theta, \{\hat{\mathbf{A}}^{(r)}\}_{r=1}^R \longleftarrow \text{argmin}_{\theta, \{\hat{\mathbf{A}}^{(r)}\}} \Big[\sum_{i=1}^N \sum_{r=1}^R \mathbb{1}(\tilde{y}_i^{(r)} \neq -1) \cdot \text{CE}(\hat{\mathbf{A}}^{(r)} \hat{\mathbf{p}}_\theta(\mathbf{x}_i), \tilde{y}_i^{(r)}) + \lambda \sum_{r=1}^R \text{tr}(\hat{\mathbf{A}}^{(r)}) \Big]$$

Return: $\hat{\mathbf{p}}_\theta$ and $\{\hat{\mathbf{A}}^{(r)}\}_{r=1}^R$

B Proof of Theorem 2

(Ryu): still need to change the statement of the lemma and the proof.

Here we intend to motivate the addition of the trace regularizer in eq. (4). In the last section, we saw that minimizing cross-entropy loss alone encourages $\hat{\mathbf{A}}^{(r)}\mathbf{P} \rightarrow \mathbf{A}^{(r)}$. Therefore, if we could devise a regularizer which, when minimized, uniquely ensures the convergence $\hat{\mathbf{A}}^{(r)} \rightarrow \mathbf{A}^{(r)}$, then this would make $\mathbf{P}$ tend to the identity matrix, implying that the base model fully captures the true label distribution i.e. $\text{argmax}_k[\mathbf{p}(\hat{\mathbf{x}})_\theta]_k = y \forall \mathbf{x}$. We describe below the trace regularizer is indeed a such regularizer when both $\hat{\mathbf{A}}^{(r)}$ and $\mathbf{A}^{(r)}$ satisfy some conditions. We first show this result assuming that there is a single annotator, and then extend to the scenario with multiple annotators.

Lemma 1 (Single Annotator)**.** *Let $\boldsymbol{P}$ be the CM of the estimated true labels $\hat{\boldsymbol{p}}_\theta$ and $\hat{\boldsymbol{A}}$ be the estimated CM of the annotator. If the model matches the noisy label distribution of the annotator i.e. $\hat{\boldsymbol{A}}\boldsymbol{P} = \boldsymbol{A}$, and both $\hat{\boldsymbol{A}}$ and $\boldsymbol{A}$ are diagonally dominant ($a_{ii} > a_{ij}$, $\hat{a}_{ii} > \hat{a}_{ij}$) for all $i \neq j$, then $\hat{\boldsymbol{A}}$ with the minimal trace uniquely coincides with the true $\boldsymbol{A}$.*

Proof. We show that each diagonal element in the true CM $\mathbf{A}$ forms a lower bound to the corresponding element in its estimation.

$$a_{ii} = \sum_j \hat{a}_{ij} p_{ji} \leq \sum_j \hat{a}_{ii} p_{ji} = \hat{a_{ii}} (\sum_j p_{ji}) = \hat{a}_{ii} \tag{5}$$

for all $i \in \{1,...,L\}$. It therefore follows that $\text{tr}(\mathbf{A}) \leq \text{tr}(\hat{\mathbf{A}})$. We now show that the equality $\hat{\mathbf{A}} = \mathbf{A}$ is uniquely achieved when the trace is the smallest i.e. $\text{tr}(\mathbf{A}) = \text{tr}(\hat{\mathbf{A}}) \Rightarrow \mathbf{A} = \hat{\mathbf{A}}$. From (5), if the trace of $\mathbf{A}$ and $\hat{\mathbf{A}}$ are the same, we see that their diagonal elements also match i.e. $a_{ii} = \hat{a}_{ii} \forall i \in \{1,...,L\}$. Now, the non-negativity of all elements in CMs $\mathbf{P}$ and $\hat{\mathbf{A}}$, and the equality $a_{ii} = \sum_j \hat{a}_{ij} p_{ji}$ imply that $p_{ji} = \mathbb{1}[i = j]$ i.e. $\mathbf{P}$ is the identity matrix.

First, let us set up the notations. For brevity, for a given input image $\mathbf{x} \in \mathbb{R}^{W \times H \times C}$, we denote the estimated CM of annotator r at $(i,j)^{\text{th}}$ pixel by $\hat{\mathbf{A}}^{(r)} := [\mathbf{A}^{(r)}(\mathbf{x})_{ij}] \in [0,1]^{L \times L}$. We also define the mean CM $\mathbf{A}^* := \sum_{r=1}^R \pi_r \hat{\mathbf{A}}^{(r)}$ and its estimate $\hat{\mathbf{A}}^* := \sum_{r=1}^R \pi_r \hat{\mathbf{A}}^{(r)}$ where $\pi_r \in [0,1]$ is the probability that the annotator r labels image $\mathbf{x}$. Lastly, as we stated earlier, we assume there is a single GT segmentation label per image — thus the true L-dimensional probability vector at pixel (i,j) takes the form of a one-hot vector i.e., $\mathbf{p}(\mathbf{x}) = \mathbf{e}_k$ for, say, class $k \in [1,...,L]$. Then, the followings result motivates the use of the trace regularisation:

Theorem 2. *If the annotator's segmentation probabilities are perfectly modelled by the model for the given image i.e., $\hat{\boldsymbol{A}}^{(r)} \hat{\boldsymbol{p}}_\theta(\boldsymbol{x}) = \boldsymbol{A}^{(r)} \boldsymbol{p}(\boldsymbol{x}) \forall r = 1,...,R$, and the average true CM $\boldsymbol{A}^*$ at a given pixel and its estimate $\hat{\boldsymbol{A}}^*$ are diagonally dominant ($a^*_{ii} > a^*_{ij}$, $\hat{a}^*_{ii} > \hat{a}^*_{ij}$ for all $i \neq j$), then $\boldsymbol{A}^{(1)},...,\boldsymbol{A}^{(R)} = \text{argmin}_{\hat{\boldsymbol{A}}^{(1)},...,\hat{\boldsymbol{A}}^{(R)}} \left[tr(\hat{\boldsymbol{A}}^*) \right]$ and such solutions are **unique** up to the k^{th} column where k is the correct pixel class.*

□

We note that the above result was also mentioned in [47] in a more general context of label noise modelling (that neglects annotator information). Here we further augment their proof by showing the uniqueness of solutions (i.e. $\text{tr}(\mathbf{A}) = \text{tr}(\hat{\mathbf{A}}) \Rightarrow \mathbf{A} = \hat{\mathbf{A}}$). In addition, the trace regularization was never used in practice in [47] — for implementation reason, the Frobenius norm was used in all their experiments. We now extend this to the multiple annotator regime.

(Ryu): need to adapt this result too to the new setting.

Proof. As the average CMs $\mathbf{A}^*$ and $\hat{\mathbf{A}}^*$ are diagonally dominant and we have $\mathbf{A}^* = \hat{\mathbf{A}}^* \mathbf{P}$, Lemma 1 yields that $\text{tr}(\mathbf{A}^*) \leq \text{tr}(\hat{\mathbf{A}}^*)$ with equality if and only if $\mathbf{A}^* = \hat{\mathbf{A}}^*$. Therefore, when the trace of the

average CM of annotators is minimized i.e. $\text{tr}(\hat{\mathbf{A}}^*) = \text{tr}(\mathbf{A}^*)$, the estimated CM of the true label distribution $\mathbf{P}$ reduces to identity, giving $\hat{\mathbf{A}}^{(r)} = \mathbf{A}^{(r)}$ for all $r \in \{1,...,R\}$. □