Learning to Segment When Experts Disagree

Le Zhang[1,2], Ryutaro Tanno[2,4], Kevin Bronik[2], Chen Jin[2],
Parashkev Nachev[3], Frederik Barkhof[1,2], Olga Ciccarelli[1], Daniel C. Alexander[2]

[1]Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation,
Queen Square Institute of Neurology, Faculty of Brain Sciences,
University College London, London, UK.
[2]Centre for Medical Image Computing, Department of Computer Science,
University College London, London, UK.
[3]High Dimensional Neurology Group, Queen Square Institute of Neurology,
University College London, London, UK.
[4]Healthcare Intelligence, Microsoft Research, Cambridge, UK

Abstract. Recent years have seen an increasing use of supervised learning methods for segmentation tasks. However, the predictive performance of these algorithms depend on the quality of labels, especially in medical image domain, where both the annotation cost and inter-observer variability are high. In a typical annotation collection process, different clinical experts provide their estimates of the "true" segmentation labels under the influence of their levels of expertise and biases. Treating these noisy labels blindly as the ground truth can adversely affect the performance of supervised segmentation models. In this work, we present a neural network architecture for jointly learning, from noisy observations alone, both the reliability of individual annotators and the true segmentation label distributions. The separation of the annotators' characteristics and true segmentation label is achieved by encouraging the estimated annotators to be maximally unreliable while achieving high fidelity with the training data. Our method can also be viewed as a translation of STAPLE, an established label aggregation framework proposed in Warfield et al [1] to the supervised learning paradigm. We demonstrate first on a generic segmentation task using MNIST data and then adapt for usage with MRI scans of *multiple sclerosis* (MS) patients for lesion labelling. Our method shows considerable improvement over the relevant baselines on both datasets in terms of segmentation accuracy and estimation of annotator reliability, particularly when only a single label is available per image. An open-source implementation of our approach can be found at `https://github.com/UCLBrain/MSLS`.

1 Introduction

Segmentation of anatomical structures in medical images is known to suffer from high inter-reader variability [2,3], affecting the performance of downstream supervised machine learning models. For example, 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 [4]. Further aggravated by differences in levels of expertise, annotations of MS lesions suffer from

high annotation variations [5]. 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 [6], that is amenable to machine learning, necessitating an intelligent method to learn robustly from such noisy annotations.

To mitigate inter-reader variations, different pre-processing techniques are commonly used to curate 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 aggregation of brain tumour segmentation labels [2]. A key limitation of such approaches, however, is that all experts are assumed to be equally reliable. Warfield *et al.*[1] proposed a label fusion method, called STAPLE that explicitly models the reliability of individual experts and use such information to "weigh" their opinions in the label aggregation step. After consistent demonstration of its superiority over the standard majority vote preprocessing in multiple applications, STAPLE has become a staple label fusion method in the creation of medical image segmentation datasets e.g., ISLES [7], MSSeg [8], Gleason'19 [9] datasets. Asman *et al.*later extended this approach in [10] by accounting for voxel-wise consensus to address the issue of under-estimation of annotators' reliability. In [11], 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 [12] 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.*[13] who designed a strategy to further incorporate the local morphological similarity between atlases and target images, and different extensions of this approach such as [14,15] have since been considered. However, these previous label fusion approaches share a drawback—they critically lack a mechanism to integrate information across different training images, fundamentally limiting the remit of applications to cases where each image receives a reasonable number of annotations from multiple experts, which may be expensive in practice. Moreover, relatively simplistic functions are employed to model the relations between observed noisy annotations, true labels and reliability of experts, which may fail to capture complex characteristics of human experts.

Our contributions: 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. Specifically, we achieve this by translating the long-standing STAPLE framework [1] to the supervised learning setting. The proposed approach (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., prone to over-segmentation) by estimating the pixel-wise confusion matrices on a per image basis. The parameters of our models are global variables that are optimised across different training samples; this enables the model to infer annotators' characteristics and true labels based on correlations between similar image samples even when only a single annotation is available per image. In contrast, this would not be possible with STAPLE [1] and its variants [11,13,14] where the anno-

tators' parameters are estimated on every image separately. Lastly, unlike the previous label fusion methods, as a supervised approach, our method produces a model that can segment test images without needing to acquire labels from annotators or atlases.

For evaluation, we first simulate a diverse range of annotator types on the MNIST dataset by performing morphometric operations with Morpho-MNIST framework [16]. Then we demonstrate the potential in the real-world task of MS lesion segmentation on ISBI 2015 challenge dataset. Experiments on both datasets demonstrate that our method leads to better performance with respect to the widely adopted STAPLE framework and the naive CNNs trained on traditional labels, and is capable of recovering the true label distributions even when there is only one label available per example.

Other related works: 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 [17] and its recent variant, PHiSeg [18] have shown that the aforementioned inter-reader variation 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 true segmentation map of the underlying anatomy, and the variations arise from the characteristics of individual annotators.

We should also note that, in standard supervised classification problems [19,20,21,22], several methods have shown promising results in modelling the label noise of multiple annotators and thereby restoring the true label distribution in medical imaging applications and beyond. By contrast, no attention has been paid to the same problem in more complicated, structured prediction tasks – to our knowledge, our work makes the first attempt to address such problem for segmentation where the outputs are high dimensional and structured. In particular, our approach yields the estimate of reliability in every pixel as a function of the input image, crucial in capturing the complex spatial variations in annotators' characteristics and absent in the classification setting.

2 Method

2.1 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* $p(\boldsymbol{y}|\boldsymbol{x})$ *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).

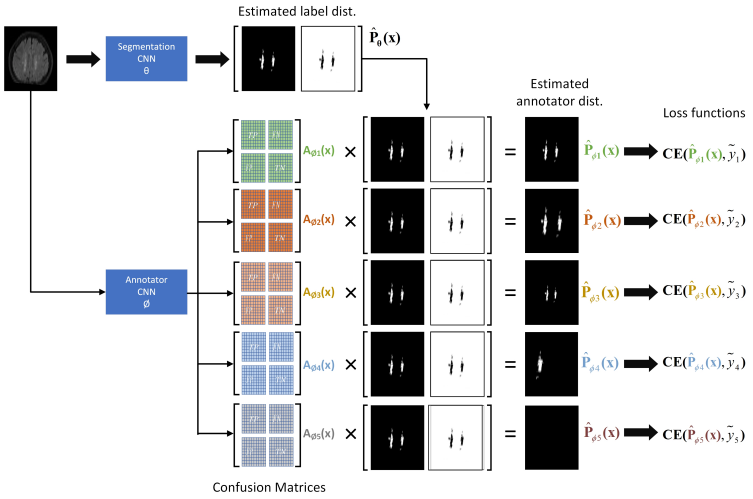

Fig. 1. General schematic of the model in the presence of 5 annotators. The method consists of two components: (1) Segmentation network parametrised by θ that generates an estimate of the GT segmentation probabilities, $p_\theta(\mathbf{x})$ for the given input image $\mathbf{x}$; (2) Annotator network, parametrised by ϕ, that estimates the confusion matrices (CMs) $\{\mathbf{A}_\phi^{(r)}(\mathbf{x})\}_{r=1}^5$ 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.

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 [12].

2.2 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:

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

where $\tilde{y}_{wh}^{(r)} \in [1,\ldots,L]$ 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:

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

where $p(y_{wh}\,|\,\mathbf{x})$ denotes the GT label distribution over the $(w,h)^{\text{th}}$ pixel in the image $\mathbf{x}$, and $p(\tilde{y}_{wh}^{(r)}\,|\,y_{wh},\mathbf{x})$ 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 $\mathbf{a}^{(r)}(\mathbf{x},w,h)_{ij}:=p(\tilde{y}_{wh}^{(r)}=i\,|\,y_{wh}=j,\mathbf{x})$ as the *confusion matrix*(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 $p(\{\tilde{\mathbf{y}}^{(r)}\}_{r\in S(\mathbf{x})}\,|\,\mathbf{x})$ 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 $p(y_{wh}=i\,|\,\mathbf{x})$;(2) *Annotator Network*, parametrised by ϕ, that generate estimates of the pixel-wise confusion matrices 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 $p(\tilde{y}_{wh}^{(r)}=i\,|\,y_{wh}=j,\mathbf{x})$. 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.

2.3 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:

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

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.*[21], which addressed a similar issue for simple classification, we add the trace of the estimated CMs to the

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
Separate CNNs on annotators	70.44 ± 0.65	n/a	46.84 ± 1.24	n/a
STAPLE [1]	78.03 ± 0.29	0.1241 ± 0.0011	55.05 ± 0.53	0.1502 ± 0.0026
Spatial STAPLE [11]	78.96 ± 0.22	0.1195 ± 0.0013	58.37 ± 0.47	0.1483 ± 0.0031
Ours without Trace (λ=0)	79.63 ± 0.53	0.1125 ± 0.0037	65.77 ± 0.62	0.1342 ± 0.0053
Our method (λ=0.001)	82.02 ± 0.21	0.0979 ± 0.0016	66.23 ± 0.39	0.0956 ± 0.0031
Our method (λ=0.01)	82.73 ± 0.21	0.0913 ± 0.0014	66.42 ± 0.37	0.0939 ± 0.0027
Our method (λ=0.1)	82.92 ± 0.19	0.0893 ± 0.0009	67.55 ± 0.31	0.0811 ± 0.0024
Our method (λ=0.7)	82.97 ± 0.14	0.0887 ± 0.0008	67.58 ± 0.29	0.0805 ± 0.0021
Our method (λ=0.9)	82.94 ± 0.18	0.0891 ± 0.0009	67.56 ± 0.33	0.0809 ± 0.0023
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 ± standard deviation).

loss in Eq. (3) as a regularisation term. 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 $\mathbf{A}$. 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 $\{\theta,\phi\}$.

While it still remains unclear whether the theoretical justification for the trace regularisation in Tanno *et al.*[21] holds in this sample-specific setting, we demonstrate empirically that such regularisation consistently improves the performance of both segmentation and the estimation of confusion matrices.

3 Experiments and Analysis

Experimental Settings. In this work, we evaluate our method on two datasets: MNIST segmentation dataset [23] and ISBI 2015 MS lesion segmentation challenge dataset [24]. The 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 each with T1w (voxel size = 0.82 × 0.82 × 1.17 mm^3) and FLAIR (voxel size = 0.82 × 0.82 × 2.2 mm^3) MRIs. 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.

For both datasets, 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 [16] (see Fig. 2(a) for examples). In particular, the first annotator provides faithful segmentation ("good-segmentation") with approximate GT, the second tends over-segment

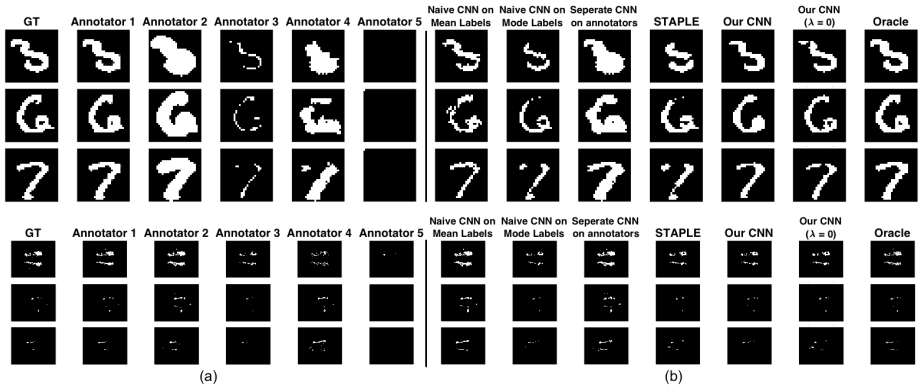

Fig. 2. Visualisation of segmentation labels on two datasets: (a) GT and simulated annotator's segmentations (Annotator 1 - 5); (b) the predictions from the supervised models.

("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. Based on the simulated segmentation annotations, we compute the corresponding CMs of respective annotators for evaluation by comparing against the GT labels. We quantify the segmentation accuracy using the dice similarity coefficient (DICE) and the error of CM estimation by the root-mean-squared-error between each CM and its estimate over the annotators.

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, the original STAPLE algorithm [1] and the Spatial STAPLE algorithm [11]. 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. All the baselines and the

Models	MNIST DICE (%) (testing)	MNIST CM estimation (validation)	MSLesion DICE (%) (testing)	MSLesion CM estimation (validation)
Naive CNN	32.79 ± 1.13	n/a	27.41 ± 1.45	n/a
STAPLE [1]	54.07 ± 0.68	0.2617 ± 0.0064	35.74 ± 0.84	0.2833 ± 0.0081
Spatial STAPLE [11]	56.73 ± 0.53	0.2384 ± 0.0061	38.21 ± 0.71	0.2591 ± 0.0074
Ours without Trace ($\lambda=0$)	74.48 ± 0.37	0.1538 ± 0.0029	54.76 ± 0.66	0.1745 ± 0.0044
Our method ($\lambda=0.001$)	75.42 ± 0.28	0.1402 ± 0.0015	55.67 ± 0.50	0.1623 ± 0.0028
Our method ($\lambda=0.01$)	75.93 ± 0.27	0.1394 ± 0.0014	55.81 ± 0.49	0.1581 ± 0.0027
Our method ($\lambda=0.1$)	76.48 ± 0.25	0.1329 ± 0.0012	56.43 ± 0.47	0.1542 ± 0.0023
Our method ($\lambda=0.7$)	76.51 ± 0.22	0.1324 ± 0.0011	56.49 ± 0.45	0.1538 ± 0.0022
Our method ($\lambda=0.9$)	76.49 ± 0.24	0.1326 ± 0.0011	56.45 ± 0.43	0.1540 ± 0.0024

Table 2. Comparison of segmentation accuracy and error of CM estimation for different methods with one label per image (mean ± std). We note that 'Naive CNN' is trained on randomly selected annotations for each image.

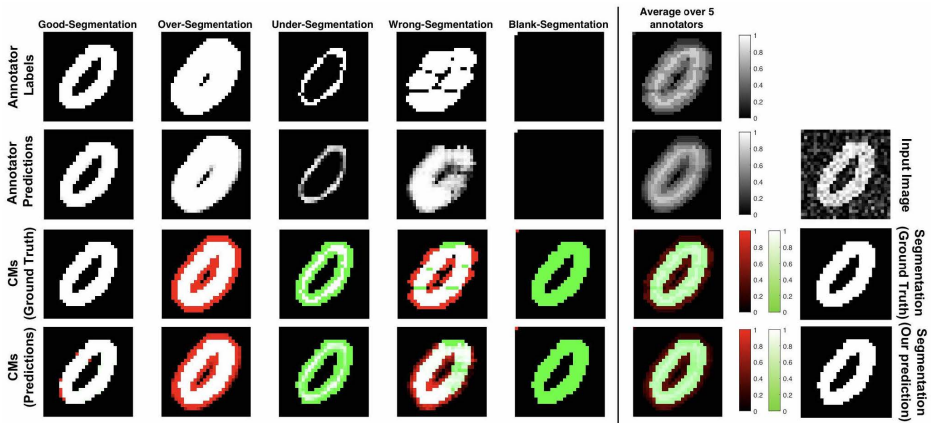

Fig. 3. Confusion matrices (CMs) of 5 simulated annotators on the MNIST dataset (Best viewed in colour: white is the true positive, green indicates the false negative, red is the false positive and black is the true negative).

annotator CNN, the segmentation CNN in our model are implemented with the NicMSlesions architecture described in [25]. We also evaluate on the validation set the effects of regularisation coefficient $\lambda \in \{0,0.001,0.01,0.1,0.7,0.9\}$ of the trace-norm in Eq. 4 on the accuracy of segmentation and CM estimation. 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.

Performance on MNIST Segmentation Dataset. The segmentation results of several examples are given in Fig. 2(b). Overall, models utilizing CMs are more effective than the naive CNN trained on traditional labels. Except the results from oracle model trained on GT, our proposed model achieves a higher dice similarity coefficient than STAPLE and Spatial STAPLE on both of the dense labels and single label (i.e., 1 label per image) scenarios (shown in Table. 1 and Table. 2). In addition, our model outperforms STAPLE and Spatial STAPLE in terms of CM estimation by a large margin, even removing the trace norm can achieve reasonably high segmentation accuracy and low CM estimation error. Fig. 3 illustrates that our method can estimate CMs of the 5 very different annotators. We can see our method clearly capturing the patterns of mistakes for each annotator.

Results on MS Dataset. Table. 1 and Table. 2 also show a strong correlation between the segmentation accuracy and the error of CM estimation on MS dataset. We observe that our model displays consistently better performance in terms of both segmentation accuracy and estimation of CMs with dense labels and single label. Examples of the different lesion segmentation results are shown in Fig. 2b. Our proposed algorithm shows comparable performance because of the benefits from the pixel information of the image, which provides additional lesion level information. Although the MS lesion is more diverse in shape and size than the MNIST digital data, Fig. 4 shows that our model can still recover the CMs of the 5 different annotators, and presents high segmentation consistency between the GT and our prediction.

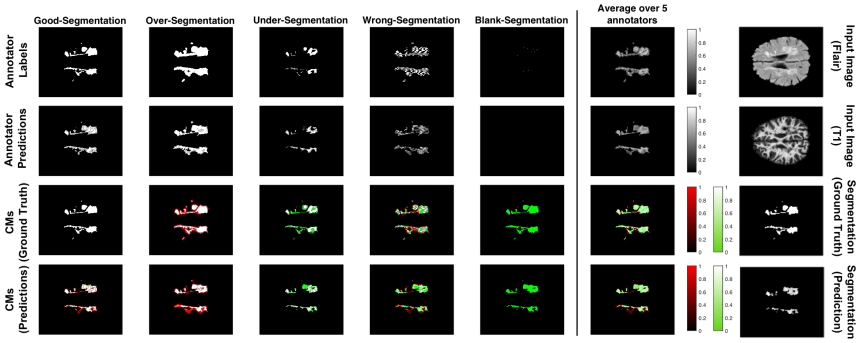

Fig. 4. Confusion matrices (CMs) of 5 simulated annotators on MS dataset (Best viewed in colour: white is the true positive, green is the false negative, red is the false positive and black is the true negative).

4 Conclusion

We introduced the first method for simultaneously recovering the label noise of multiple annotators and the GT label distribution for supervised segmentation problem. We demonstrated this method on the MNIST segmentation dataset and MS lesions dataset respectively. Our method is capable of estimating individual annotators and thereby improving robustness to label noise. Experiments have shown considerable improvement over the common CNNs trained on aggregated labels based on averaging, the majority vote and the widely used STAPLE and Spatial STAPLE framework in terms of both segmentation accuracy and the quality of confusion matrix estimation.

In the future, we plan to extend to accommodate knowledge about the meta-information of annotators (e.g., number of years of experience) and also the non-image data (e.g., genetics) that may influence the pattern of the underlying segmentation label such as lesion appearance. 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").

Acknowledge funding sources: EPSRC grants EP/R006032/1, EP/M020533/1, CRUK/EPSRC grant NS/A000069/1, and the NIHR UCLH Biomedical Research Centre all support this work.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Gleason 2019 challenge. `https://gleason2019.grand-challenge.org/Home/`. Accessed: 2020-02-30.
10. 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.
11. 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.
12. 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.
13. 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.
14. Andrew J Asman and Bennett A Landman. Non-local statistical label fusion for multi-atlas segmentation. *Medical image analysis*, 17(2):194–208, 2013.
15. 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.
16. 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.
17. 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.

18. 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.
19. 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.
20. Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*, 2017.
21. 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.
22. Carole H Sudre, Beatriz Gomez Anson, Silvia Ingala, Chris D Lane, Daniel Jimenez, Lukas Haider, Thomas Varsavsky, Ryutaro Tanno, Lorna Smith, Sébastien Ourselin, et al. Let's agree to disagree: Learning highly debatable multirater labelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 665–673. Springer, 2019.
23. Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
24. 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.
25. 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.