Emerging Semantic Segmentation from Positive and Negative Coarse Label Learning

Le Zhang[5] *, Fuping Wu[1,2], Arun Thirunavukarasu[3], Kevin Bronik[4], Thomas Nichols[1], and Bartłomiej W. Papież[1,2]

[1] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[2] Nuffield Department of Population Health, University of Oxford, Oxford, UK
[3] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[4] Department of Engineering Science, University of Oxford, Oxford, UK
[5] School of Engineering, College of Engineering and Physical Sciences, University of Birmingham, Birmingham, UK

Abstract. Large annotated datasets are vital for training segmentation models, but pixel-level labeling is time-consuming, error-prone, and often requires scarce expert annotators, especially in medical imaging. In contrast, coarse annotations are quicker, cheaper, and easier to produce, even by non-experts. In this paper, we propose to use coarse drawings from both positive (target) and negative (background) classes in the image, even with noisy pixels, to train a convolutional neural network (CNN) for semantic segmentation. We present a method for learning the true segmentation label distributions from purely noisy coarse annotations using two coupled CNNs. The separation of the two CNNs is achieved by high fidelity with the characters of the noisy training annotations. We propose to add a complementary label learning that encourages estimating negative label distribution. To illustrate the properties of our method, we first use a toy segmentation dataset based on MNIST. We then present the quantitative results of experiments using publicly available datasets: Cityscapes dataset for multi-class segmentation, and retinal images for medical applications. In all experiments, our method outperforms state-of-the-art methods, particularly in the cases where the ratio of coarse annotations is small compared to the given dense annotations.

Keywords: Segmentation · Coarse Label · Weakly-Supervise Learning

1 Introduction

Thanks to the availability of large datasets with accurate annotations, fully supervised learning (FSL), especially deep supervised learning, has been translated from theoretical algorithms to practice [23] [1]. However, it is generally expensive, time-consuming, and often infeasible to collect pixel-level labels for large-scale

* This research was carried out at the University of Oxford, where the author was affiliated during the period in which this work was completed.

datasets. This problem is particularly prominent in the clinical domain where labeled data are scarce due to the high cost of annotations [22]. For instance, accurate segmentation of vessels in fundus retinal images is difficult even for experienced experts due to variability of vessel's location, size, and shape across population or disease. The labeling process is prone to errors, almost inevitably leading to noisy datasets as seen in machine learning benchmark datasets [13]. Labeling errors can occur due to automated label extraction, ambiguities in input and output spaces, or human errors (e.g. lack of expertise). As a consequence, despite the availability of large imaging repositories, the generation of the curated labels that are available to machine learning remains a challenging issue, necessitating the development of methods that learn robustly from noisy annotations.

To reduce the workload of pixel-level annotation, there has been a considerable effort to exploit weakly-supervised strategies. Weakly-supervised learning (WSL) uses annotations that are cheaper to produce than pixel-wise labels such as bounding boxes [4, 6], coarse annotation [18, 15], scribbles [9, 20], or image-level labels [11, 12] to train the segmentation models. However, the information from weak annotations is of lower precision and usually suffers from noisy information when curating the labels. For example, image-level labels cannot provide the position information of the object of interest (OOI). Using bounding boxes helps to indicate the rough positions of the OOI, but the pixels inside the bounding box may belong to multiple classes if the box size is large. A more annotator-friendly method of efficient supervision is scribble-based annotation, which only requires the annotator to draw a few lines to mark a small part of the OOI. Coarse annotations can provide much more information than scribble and avoid large non-target pixels being grabbed into the bounding box. Meanwhile, drawing coarse annotations on images needs only similar effort and time as the scribble and box-level labeling, and can be conducted by non-experts. Therefore, learning from those coarse annotations, and then correcting the noisy pixels with computational methods, may represent an optimally efficient means of enriching labeled large-scale data with minimal effort.

Our contribution: We introduce an end-to-end supervised segmentation method that estimates true segmentation labels from noisy coarse annotations. The proposed architecture (see Fig. 1) consists of two coupled CNNs where the first CNN estimates the true segmentation probabilities, and the second CNN models the characteristics of two different coarse annotations by estimating the pixel-wise confusion matrices (CMs) on a per-image basis. Unlike previous WSL methods using coarse annotations, our method models and disentangles the complex mappings from the input images to the noisy coarse annotations and to the true segmentation label simultaneously. Specifically, we model the noisy coarse annotation for the objective along with the complementary label learning for the background or non-objective to enable our model to disentangle robustly the errors of the given annotations and the true labels, even when the ratio of coarse annotation is small (e.g., given scribble for each class). In contrast,

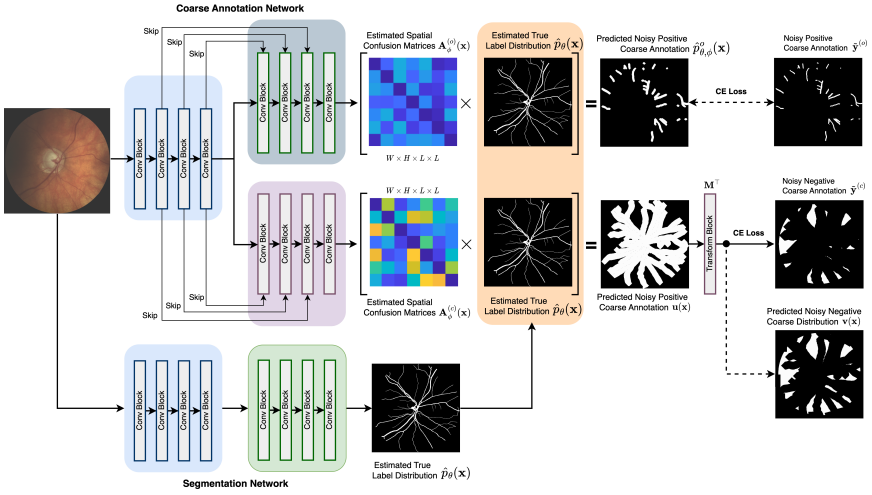

Fig. 1. General schematic of the model supervised by noisy coarse annotations. The method consists of two components: (1) Segmentation network parameterized by θ that generates an estimate of the true segmentation probabilities $\hat{p}_{\theta}(\mathbf{x})$ for the given input image $\mathbf{x}$; (2) Coarse annotation network, parameterized by ϕ, that estimates the confusion matrices (CMs) $\{\mathbf{A}_{\phi}^{(o)}(\mathbf{x}),\ \mathbf{A}_{\phi}^{(c)}(\mathbf{x})\}$ of the noisy coarse annotations.

this would not be possible with the other WSL methods where the parameters of each coarse annotation are estimated on every target image separately.

2 Method

2.1 Problem Set-up

In this work, we consider developing a supervised segmentation model by learning the positive (e.g. object to be segmented) and negative (e.g. object to be not segmented) coarse annotations that are easy and less expensive to be acquired from 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 coarse segmentation labels $\{\{\tilde{\mathbf{y}}_n^{(o)}, \tilde{\mathbf{y}}_n^{(c)}\} \in \mathcal{Y}^{W\times H}\}_{n=1,\ldots,N}^{\{o,c\}\in S(\mathbf{x}_i)}$ from objective and complementary categories where $\tilde{\mathbf{y}}_n^{(o)}$ and $\tilde{\mathbf{y}}_n^{(c)}$ denote the noisy objective and complementary coarse annotations, respectively and $S(\mathbf{x}_n)$ denotes the set of all annotations of image $\mathbf{x}_i$ and $\mathcal{Y} = [1, 2, ..., L]$ denotes the set of segmentation classes.

2.2 Learning from Noisy Coarse Annotations

In this section, we describe how we jointly optimize the parameters of the segmentation network, θ, and the parameters of the coarse annotation network, ϕ.

In short, we minimize the negative log-likelihood of the probabilistic model for both positive and negative coarse annotations via stochastic gradient descent. A detailed description is provided below.

Learning with Positive Coarse Label. Given training input $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ and positive coarse labels $\tilde{\mathbf{Y}}^{\{o\}} = \{\tilde{\mathbf{y}}_n^{\{o\}} : \{o\} \in S(\mathbf{x}_n)\}_{n=1}^N$, we optimize the parameters $\{\theta, \phi\}$ by minimizing the negative log-likelihood (NLL), $-\log p(\tilde{\mathbf{Y}}^{(1)}, ..., \tilde{\mathbf{Y}}^{(O)} | \mathbf{X})$. The optimization objective equates to the sum of cross-entropy (CE) losses between the observed positive coarse segmentation and the predicted label distributions:

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

Minimizing Eq. (1) encourages the positive-specific predictions $\hat{\mathbf{p}}_{\theta,\phi}^{(o)}(\mathbf{x})$ to be as close as possible to the provided positive coarse label distributions $\mathbf{p}^{(o)}(\mathbf{x})$. However, this loss function alone cannot separate the annotation noise from the true label distribution; there are many combinations of pairs $\hat{\mathbf{A}}_\phi(\mathbf{x})$ and segmentation model $\hat{\mathbf{p}}_\theta(\mathbf{x})$ such that $\hat{\mathbf{p}}_{\theta,\phi}(\mathbf{x})$ matches well the provided distribution $\mathbf{p}(\mathbf{x})$ for any input image $\mathbf{x}$ (e.g. permutations of rows in the CMs). Tanno *et al.* [17] addressed an analogous issue for the classification task, and here we add the trace of the estimated CMs to the loss function for positive coarse annotation in Eq.(1) as a regularisation term. We thus optimize the combined loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{obj}}(\theta, \phi) &:= \mathcal{L}_{\mathrm{obj}}(\theta, \phi)(\hat{\mathbf{A}}_\phi^{(o)}(\mathbf{x}_n) \cdot \hat{\mathbf{p}}_\theta(\mathbf{x}_n), \tilde{y}^{(o)}) \\ &:= \sum_{n=1}^{N} \sum_{o=1}^{O} \mathbb{1}(o \in \mathcal{S}(\mathbf{x}_n)) \cdot \Big[\mathrm{CE}\big(\hat{\mathbf{A}}_\phi^{(o)}(\mathbf{x}_n) \cdot \hat{\mathbf{p}}_\theta(\mathbf{x}_n),\ \tilde{\mathbf{y}}_n^{(o)}\big) \\ &+\ \lambda \cdot \mathrm{tr}\big(\hat{\mathbf{A}}_\phi^{(o)}(\mathbf{x}_n)\big)\Big] \end{aligned} \tag{2}$$

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

Learning with Negative Coarse Label. For some situations, it is easier to provide the negative coarse annotation, e.g., complementary label, to help ML model predict the true label distribution. Thus, we study a readily available substitute, namely complementary labeling. However, if we still use traditional loss functions $\mathcal{L}_{\mathrm{obj}}(\theta, \phi)$ when learning with these complementary labels, similar to Eq. (2), we can only learn a mapping $\mathbb{R} \to \mathcal{Y}$ that tries to predict conditional probabilities $p(\tilde{\mathbf{y}}^{(c)} \mid \mathbf{x})$ and the corresponding complementary pixels classifier that predicts a $\tilde{y}_{wh}^{(c)}$ for a given observation $\mathbf{x}$.

To address the above issue, inspired by Yu [19], which summarizes all the probabilities into a transition matrix $\mathbf{M} \in \mathbb{R}^{L\times L}$, where $\mathbf{m}(\mathbf{x}, w, h)_{ij} := p(\tilde{y}_{wh}^{\{c\}} = i \mid y_{wh} = j, \mathbf{x})$ and $\mathbf{m}(\mathbf{x}, w, h)_{ii} = 0, \forall i, j \in \{1, ..., L\}$. Here, $\mathbf{m}_{ij}$ denotes the entry value in the i-th row and j-th column of $\mathbf{M}$. As shown in Fig. 1, we achieve this simply by adding a linear layer to the complementary label learning channel. This layer outputs $v(\mathbf{x})$ by multiplying the output of the CE function (i.e., $u(\mathbf{x})$) with the transposed transition matrix $\mathbf{M}^\top$. Note that the transition matrix is also widely exploited in Markov chains [5] and has many applications in machine learning, such as learning with label noise [17, 22, 21].

Recall that in transition matrix $\mathbf{M}, \mathbf{m}_{ij} = p(\tilde{y}_{wh}^{\{c\}} = i \mid y_{wh} = j, \mathbf{x})$ and $\mathbf{m}_{ii} = p(\tilde{y}_{wh}^{\{c\}} = i \mid y_{wh} = i, \mathbf{x}) = 0$. We observe that $p(\tilde{\mathbf{y}}^{\{c\}} \mid \mathbf{x})$ can be transferred to $p(\tilde{\mathbf{y}}^{\{c\}} \mid \mathbf{x})$ by using the transition matrix $\mathbf{M}$,

$$\begin{aligned} p(\tilde{y}_{wh}^{(c)} = j \mid \mathbf{x}) &= \sum_{i\neq j} p(\tilde{y}_{wh}^{(c)} = j, \bar{y}_{wh}^{(c)} = i \mid \mathbf{x}) \\ &= \sum_{i\neq j} p(\tilde{y}_{wh}^{(c)} = j \mid \bar{y}_{wh}^{(c)} = i, \mathbf{x}) p(\bar{y}_{wh}^{(c)} = i \mid \mathbf{x}) \\ &= \sum_{i\neq j} p(\tilde{y}_{wh}^{(c)} = j \mid \bar{y}_{wh}^{(c)} = i) p(\bar{y}_{wh}^{(c)} = i \mid \mathbf{x}) \end{aligned}$$

Intuitively, if $\mathbf{v}_i(\mathbf{x})$ tries to predict the probability $p(\tilde{y}^{(c)} = i \mid \mathbf{x}), \forall i \in [1, ..., L]$, then $\mathbf{M}^{-\top}\mathbf{v}$ can predict the probability $p(\tilde{y}^{(o)} \mid \mathbf{x})$, which is the positive prediction of the corresponding complementary coarse label. To enable end-to-end learning rather than transferring after training, we let

$$\mathbf{v}(\mathbf{x}) = \mathbf{M}^\top \mathbf{u}(\mathbf{x})$$

where $\mathbf{u}(\mathbf{x})$ is now an intermediate output of the complementary coarse annotation, and $\mathcal{L}_{\text{comp}}(\theta, \phi) = \arg\max_{i\in[L]} \mathbf{v}_i(\mathbf{x})$. Then, the modified loss function $\bar{\mathcal{L}}_{\text{obj}}(\theta, \phi)$ is

$$\begin{aligned} &\bar{\mathcal{L}}_{\text{obj}}(\theta, \phi)(\mathbf{u}(\mathbf{x}), \tilde{y}^{(c)}) := \mathcal{L}_{\text{comp}}(\theta, \phi)(\mathbf{v}(\mathbf{x}),\ \tilde{y}^{(c)}) \\ &:= \mathcal{L}_{\text{comp}}(\theta, \phi)(\mathbf{M}^\top \cdot \{\hat{\mathbf{A}}_\phi^{(c)}(\mathbf{x}_n) \cdot \hat{\mathbf{p}}_\theta(\mathbf{x}_n)\},\ \tilde{y}^{(c)}) \\ &:= \sum_{n=1}^{N}\sum_{c=1}^{C} \mathbb{1}(c \in \mathcal{S}(\mathbf{x}_n)) \cdot \Big[\text{CE}\big(\mathbf{M}^\top \cdot \{\hat{\mathbf{A}}_\phi^{(c)}(\mathbf{x}_n) \cdot \hat{\mathbf{p}}_\theta(\mathbf{x}_n)\},\ \tilde{\mathbf{y}}_n^{(c)}\big) \\ &+\ \lambda \cdot \text{tr}\big(\hat{\mathbf{A}}_\phi^{(c)}(\mathbf{x}_n)\big)\Big] \end{aligned} \tag{3}$$

In this way, if we can learn an optimal $\mathbf{v}$ such that $\mathbf{v}_i(\mathbf{x}) = p(\tilde{y}^{(c)} = i \mid \mathbf{x}), \forall i \in [L]$, meanwhile, we can also find the optimal $\mathbf{u}$ and the loss function $\mathcal{L}_{\text{comp}}(\theta, \phi)$, which can be easily applied to deep learning. With sufficient training examples with complementary coarse labels, this DNN often simultaneously learns good classifiers for both $(\mathbf{x}, \tilde{y}^{(c)})$ and $(\mathbf{x}, \tilde{y}^{(o)})$.

Finally, we combine the positive annotation loss $\mathcal{L}_{\text{obj}}$ and the negative annotation loss $\mathcal{L}_{\text{comp}}$ as our objective and optimize:

$$\mathcal{L}_{\text{final}}(\theta, \phi) := \mathcal{L}_{\text{obj}}(\theta, \phi) + \mathcal{L}_{\text{comp}}(\theta, \phi). \tag{4}$$

3 Experiments

In this section, we first describe our dataset and then show the coarse annotation refinement scenarios in simulated and real-world settings, separately.

Datasets. MNIST dataset [7] 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 Cityscapes [3] dataset contains 5000 high-resolution (2048×1024 pixels) urban scene images collected across 27 European Cities. The dataset comprises 5,000 fine annotations (2,975 for training, 500 for validation, and 1,525 for testing) and 20,000 coarse annotations where 11,900 samples for training and 2,000 for validation (i.e., coarse polygons covering individual objects). The LES-AV [10] is a dataset for retinal vessel segmentation on color fundus images. It comprises 22 fundus photographs with available manual annotations of the retinal vessels including annotations of arteries and veins. The 22 images/patients are acquired with resolutions of 30° field-of-view (FOV) and 1444×1620 pixels (21 images), and 45° FOV and 1958×2196 pixels (one image), with each pixel $= 6\mu m$. We divide them into 18 images for training and 4 images for testing.

Synthetic Noisy Coarse Annotations. We generate synthetic coarse noisy annotations from an assumed expert consensus label on MNIST, Cityscapes and retinal fundus image dataset, to demonstrate the efficacy of the approach in an idealized situation where the expert consensus label is known. We simulate the positive and negative coarse noisy annotations by performing morphological transformations (e.g., thinning, thickening, fractures, etc) on the expert consensus label and background (complementary label), using MorphoMNIST software [2]. In particular, *positive coarse noisy annotation* is prone to be poor segmentation, which is simulated by combining small fractures and over-segmentation; *negative coarse noisy annotation* always annotates on the background or complementary label using the same approach. To create synthetic coarse noisy labels in the multi-class scenario, we use a similar simulation to create coarse labels on the Cityscapes dataset. We first choose a target class and then apply morphological operations on the provided coarse mask to create the two synthetic coarse labels at different patterns, namely, objective coarse and complementary coarse annotations. We create training data by deriving labels from the simulated annotations.

Comparison Methods and Evaluation Metrics. Our experiments are based on the assumption that no expert consensus label is available a priori, hence, we compare our method against multiple weakly-supervised and semi-supervised methods. In particular, we explore our method with ablation studies, e.g., our method without negative coarse annotation; we also consider two pop-

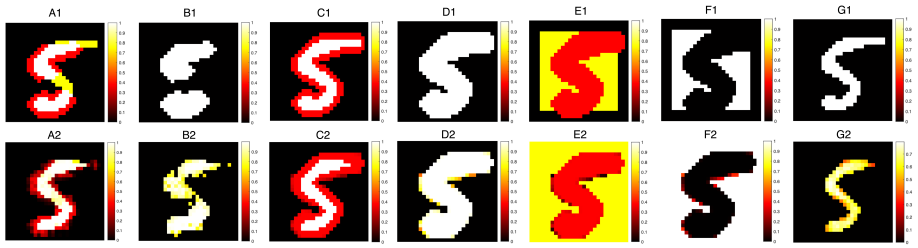

Fig. 2. Visualisation of the estimated labels, the estimated pixel-wise CMs, and the estimated TMs on MNIST datasets along with their targets (best viewed in color). White is the true positive, yellow is the false negative, red is the false positive, and black is the true negative. A1-A2: the target and estimated CM: $\mathbf{A}_{\phi}^{(o)}(\mathbf{x})$ for positive coarse annotation; B1-B2: the given and estimated positive coarse annotation $\tilde{\mathbf{y}}^{(o)}$; C1-C2: the target and estimated intermediate CM: $\mathbf{A}_{\phi}^{(c)}(\mathbf{x})$ for negative coarse annotation; D1-D2: the target and estimated intermediate negative coarse annotation $\mathbf{u}(\mathbf{x})$; E1-E2: the target and estimated TM: $\mathbf{M}^{\top}$ for negative coarse annotation; F1-F2: the provided and estimated negative coarse annotation ($\tilde{\mathbf{y}}^{(c)}$ and $\mathbf{v}(\mathbf{x})$); G1-G2: target label and our estimation.

ular interactive image segmentation algorithms for generating masks from scribbles: GrabCut [14] and LazySnapping [8], then training FCNs using the masks generated by these methods. Meanwhile, we compare our weakly-supervised results based on the noisy coarse annotations and strongly-supervised results based on the expert consensus annotations. For evaluation metrics, we use mIoU between estimated segmentation $\hat{\mathbf{p}}_{\theta}(\mathbf{x})$ and expert consensus label $\mathbf{y}_{GT}$.

Strategy performance of utilizing coarse annotation. Our method jointly propagates information into unmarked pixels and learns network parameters. An easy way is to first use any existing interactive image segmentation methods to generate masks based on coarse annotation, and then use these masks to train FCNs. In Table 1, we compare our methods with these two-step solutions. We investigate two popular interactive image segmentation algorithms for generating masks from coarse annotation: GrabCut [14] and LazySnapping [8]. Given the coarse annotations, GrabCut generates the mask only for the target pixels while LazySnapping produces the masks not only for the objective but also for the non-target pixels. Training FCNs using the masks generated by these methods shows inferior semantic segmentation accuracy. This is because these traditional methods [14, 8] only focus on the low-level color/spatial information and are unaware of semantic content. The generated masks cannot be the reliable "GT" for training the supervised networks. On the contrary, our coarse-based supervised method achieves a score of 82.5% on MNIST and 68.3% on Cityscapes dataset, about 10% higher than the two-step solutions. This is because our model can capture the patterns of mistakes for each noisy coarse annotation, and the high-level information can help with the coarse-to-fine propagation of the true label estimation. This behavior is shown in Fig. 2.

Methods	MNIST	Cityscapes
GrabCut + FCN	75.2 ± 0.3	53.6 ± 0.4
LazySnapping+FCN	78.5 ± 0.2	59.4 ± 0.4
Ours (w/o negative annotation)	77.2 ± 0.2	62.3 ± 0.2
Ours (Full)	82.5 ± 0.1	68.3 ± 0.2

Table 1. Segmentation results (mIoU (%)) on the MNIST and Cityscapes validation set via different strategies of utilizing coarse annotations.

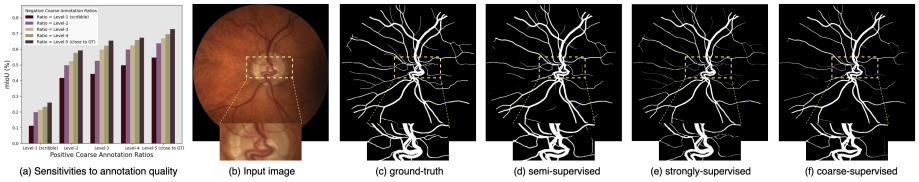

Fig. 3. Sensitivities to quality of positive and negative coarse annotations, and the visualization of the estimated labels with different supervision approaches.

Performance on Retinal Vessel Segmentation. We illustrate the results of our approach on a more challenging dataset with real coarse and noisy labels from the medical domain. This dataset, called LES-AV, consists of images of the retinal fundus acquired from different patients. The task is to segment the vessel into a binary mask (see Fig. 3). The process of segmenting the blood vessel in the retinal image is crucial for the early detection of eye diseases.

An experienced annotator was tasked with providing the practical positive and negative coarse annotations for each sample on LES-AV dataset. We generate such a real-world dataset to show the segmentation results and evaluate the performance of different supervision approaches. Meanwhile, we also created 5 different ratio levels for the positive and negative coarse annotations from *level-1* (tend to scribble) to *level-5* (tend to GT) with increasing ratios compared to the given expert consensus labels. We use such a dataset to evaluate the sensitivity to annotation quality of our model on medical image data.

We show the results of sensitivities to annotation quality in Fig. 3(a). Our model performs robustly and gradually improved when the ratio of coarse annotations is increased. Especially when the ratio is increased from *level-1* to *level-2*, our model's performance is increased significantly and comparable to the mask-level results. By applying our practical annotations, we conduct a group of experiments under different supervision. The results in Table. 2 indicate that our WSL approach achieves comparable results to the strongly-supervised method. Meanwhile, by including some extra coarse annotations, the result is improved 3%. Finally, we present the segmentation visualization in Fig. 3(d∼f).

supervision	w/ masks	w/ coarse	total	mIoU (%)
weakly	–	16	16	65.8 ± 0.3
strongly	16	–	16	69.2 ± 0.2
semi	16	4	20	71.6 ± 0.3

Table 2. Comparisons of our method using different annotations on the LES-AV retinal image validation set. The term "w/ masks" shows the number of training images with mask-level annotations, and "w/ coarse" shows the number of training images with coarse annotations.

4 Conclusion

We introduced a new theoretically grounded algorithm for recovering the expert consensus label distribution from noisy coarse annotations. Our method enjoys implementation simplicity, requiring only adding a complementary label learning term to the loss function. Experiments on both synthetic and real data sets have shown superior performance over the common WSL and SSL methods in terms of both segmentation accuracy and robustness to the quality of coarse annotations and label noise. Furthermore, the method is capable of estimating coarse annotations even when scribble is given per image.

Our work was primarily motivated by medical imaging applications that require pixel-level annotation for a large number of images. These include segmentation of retinal vessels on fundus photography, organs on computed tomography, and cells on histopathology slides. However, future work shall consider imposing structures on the CMs and TMs to broaden the applicability to scribble or spot annotations, which could contribute to saving the labor of the labeling process. Another promising direction is to explore the use of promptable foundation models [16] for image segmentation, which have recently shown strong potential in handling a variety of input forms and tasks. Incorporating such models into our framework could further enhance flexibility and reduce annotation burdens.

References

1. Bronik, K., Zhang, L.: Conditional advancement of machine learning algorithm via fuzzy neural network. Pattern Recognition **155**, 110732 (2024)
2. Castro, D.C., Tan, J., Kainz, B., Konukoglu, E., Glocker, B.: Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. Journal of Machine Learning Research **20** (2019)
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
4. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1635–1643 (2015)

5. Gagniuc, P.A.: Markov chains: from theory to implementation and experimentation. John Wiley & Sons (2017)
6. Jiang, P.T., Yang, Y., Hou, Q., Wei, Y.: L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16886–16896 (June 2022)
7. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
8. Li, Y., Sun, J., Tang, C.K., Shum, H.Y.: Lazy snapping. ACM Transactions on Graphics (ToG) **23**(3), 303–308 (2004)
9. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3159–3167 (2016)
10. Orlando, J.I., Barbosa Breda, J., Keer, K.v., Blaschko, M.B., Blanco, P.J., Bulant, C.A.: Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 65–73. Springer (2018)
11. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1742–1750 (2015)
12. Pathak, D., Shelhamer, E., Long, J., Darrell, T.: Fully convolutional multi-class multiple instance learning. arXiv preprint arXiv:1412.7144 (2014)
13. Peterson, J.C., Battleday, R.M., Griffiths, T.L., Russakovsky, O.: Human uncertainty makes classification more robust. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9617–9626 (2019)
14. Rother, C., Kolmogorov, V., Blake, A.: " grabcut" interactive foreground extraction using iterated graph cuts. ACM transactions on graphics (TOG) **23**(3), 309–314 (2004)
15. Saha, O., Cheng, Z., Maji, S.: Improving few-shot part segmentation using coarse supervision. arXiv preprint arXiv:2204.05393 (2022)
16. Simons, S.J., Papież, B.W.: Spinefm: Leveraging foundation models for automatic spine x-ray segmentation. arXiv preprint arXiv:2411.00326 (2024)
17. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11244–11253 (2019)
18. Wang, Z., Popel, A.S., Sulam, J.: Label cleaning multiple instance learning: Refining coarse annotations on single whole-slide images. arXiv preprint arXiv:2109.10778 (2021)
19. Yu, X., Liu, T., Gong, M., Tao, D.: Learning with biased complementary labels. In: Proceedings of the European conference on computer vision (ECCV). pp. 68–83 (2018)
20. Zhang, K., Zhuang, X.: Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11656–11665 (June 2022)
21. Zhang, L., Tanno, R., Bronik, K., Jin, C., Nachev, P., Barkhof, F., Ciccarelli, O., Alexander, D.C.: Learning to segment when experts disagree. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 179–190. Springer (2020)

22. Zhang, L., Tanno, R., Xu, M.C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D.: Disentangling human error from ground truth in segmentation of medical images. Advances in Neural Information Processing Systems **33**, 15750–15762 (2020)
23. Zhang, L., Wu, F., Bronik, K., Papiez, B.W.: Diffuseg: domain-driven diffusion for medical image segmentation. IEEE Journal of Biomedical and Health Informatics (2025)