Hindawi
Computational Intelligence and Neuroscience
Volume 2023, Article ID 1883597, 11 pages
https://doi.org/10.1155/2023/1883597

Research Article

A Soft Label Method for Medical Image Segmentation with Multirater Annotations

Jichang Zhang, Yuanjie Zheng, and Yunfeng Shi

School of Information Science & Engineering, Shandong Normal University, No. 1 Daxue Road, Changqing District, Jinan 250358, China

Correspondence should be addressed to Yuanjie Zheng; yjzheng@sdnu.edu.cn and Yunfeng Shi; yunfeng@sdnu.edu.cn

Received 22 June 2022; Revised 4 October 2022; Accepted 6 October 2022; Published 18 February 2023

Academic Editor: Changming Sun

In medical image analysis, collecting multiple annotations from different clinical raters is a typical practice to mitigate possible diagnostic errors. For such multirater labels' learning problems, in addition to majority voting, it is a common practice to use soft labels in the form of full-probability distributions obtained by averaging raters as ground truth to train the model, which benefits from uncertainty contained in soft labels. However, the potential information contained in soft labels is rarely studied, which may be the key to improving the performance of medical image segmentation with multirater annotations. In this work, we aim to improve soft label methods by leveraging interpretable information from multiraters. Considering that mis-segmentation occurs in areas with weak supervision of annotations and high difficulty of images, we propose to reduce the reliance on local uncertain soft labels and increase the focus on image features. Therefore, we introduce local self-ensembling learning with consistency regularization, forcing the model to concentrate more on features rather than annotations, especially in regions with high uncertainty measured by the pixelwise interclass variance. Furthermore, we utilize a label smoothing technique to flatten each rater's annotation, alleviating overconfidence of structural edges in annotations. Without introducing additional parameters, our method improves the accuracy of the soft label baseline by 4.2% and 2.7% on a synthetic dataset and a fundus dataset, respectively. In addition, quantitative comparisons show that our method consistently outperforms existing multirater strategies as well as state-of-the-art methods. This work provides a simple yet effective solution for the widespread multirater label segmentation problems in clinical diagnosis.

1. Introduction

Recently, deep learning techniques have made impressive progress on image segmentation tasks and have become a popular choice in the computer vision community [1]. Typically, supervised learning in deep learning is based on the assumption that there is a ground truth (GT). However, the truth is a lie; that is, there is often a lack of human consensus on the category of an object [2–4]. Especially, in medical image segmentation, which is based on knowledge and experience, disagreements between raters are fairly common [5, 6]. Inter-rater variability, as frequently reported by relevant research in the clinical field, usually leads to difficulties in segmenting areas of high uncertainty [7, 8].

To mitigate this inter-rater variability, the most basic yet common approach is the majority voting approach, in which opinions agreed by a majority of raters are taken as true. However, the majority voting approach essentially discards the rich information contained in the multirater labels through one-hot operation (e.g., the probability distribution [0.6, 0.3, and 0.1] is transformed into a hard label [1, 0, and 0]). To combat this issue, soft-label methods that average rater annotations have been intensively investigated [9, 10]. Furthermore, Islam and Glocker [11] introduced a label smoothing method that incorporates fuzzy information about edges into multirater soft labels, called spatially varied label smoothing (SVLS).

However, when we applied the soft labels method to the multirater optic cup (OC) and optic disc (OD) segmentation of the fundus image task, finding that the areas where segmentation errors occur coincides with the highly divergent areas to some extent, see Figure 1. As demonstrated

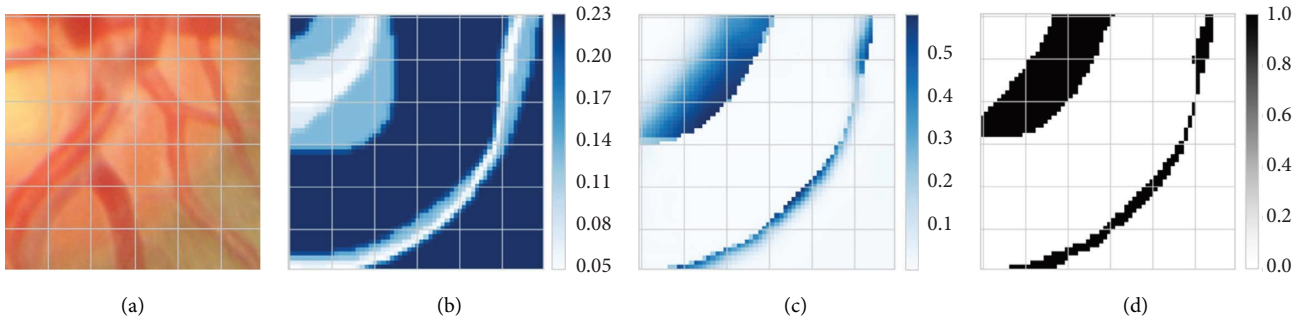

Figure 1: (a) Local visualization of an exemplary fundus image; (b) interclass variance map of annotations from six raters for OC and OD segmentation; (c) corresponding loss map of prediction; (d) corresponding error rates' map of prediction.

in Figure 2, the pixelwise loss and error rates of predictions are statistically positively correlated with the interclass variance which indicates the divergence between raters. We tentatively attempt to explain this phenomenon as follows:

(i) An explanation is that the divergence area is highly uncertain, and the higher the uncertainty of the label, the lower the penalty imposed on the predicted distribution [12]. Highly uncertain annotations make it difficult to impose strong and precise constraints on the model, which is similar to weakly supervised learning that lacks accurate annotations [13]. The dependence on annotations in weakly supervised learning is weakened and replaced by a focus on features [14].

(ii) Furthermore, we provide an intuitive interpretation that is more consistent with the multirater labels segmentation task: uncertainty reflects pixelwise image difficulty, where areas with high difficulty are more challenging for the model to accurately segment. Image difficulty, which is related to the visual characteristics of the images, such as image quality and occlusion of the area of the lesions, is one of the causal factors of inter-rater variability [15]. As demonstrated in Figure 1(a), the blood vessels occluding the edge region of the OD not only make ophthalmologists' judgment difficult but also hinder the accurate prediction of the deep neural network. In contrast to existing methods [16, 17] that treat difficulty as image level, we innovatively consider difficulty to be pixelwise for segmentation tasks.

In conclusion, the regions with high inter-rater variability have more difficult features but only weaker supervision, which could be a cause of mis-segmentation. In this work, we aim to improve the performance level of the soft label approach on multirater labels' segmentation task based on the previously mentioned explanations. A way to get the best of both sides is to increase the focus on image features while reducing the reliance on highly uncertain annotations. Consequently, we propose a supervised segmentation network that is constrained by consistency regularization. Specifically, consistency regularization exploits the augmentation invariance of images to optimize the feature space while avoiding relying simply on labels and compensating for the disadvantage of unreliable local annotations. The uncertainty as the prior knowledge is formulated as the soft labels' interclass variance, which drives the proposed model's local difference training. In addition, the SVLS approach, which incorporates edge fuzziness into soft labels, is used to soften average expert labels.

Experiments are performed on a synthetic dataset with great disagreement as well as a real-world dataset. In these experiments, our method consistently outperforms existing multirater strategies and state-of-the-art (SOTA) methods. To verify the generalization of the proposed method, we additionally conduct generalization experiments on two other types of datasets.

In summary, the main contributions of this study are as follows:

(1) To embed consistency/inconsistency of multirater into the model, the soft labels obtained by averaging softened annotations of raters are used as GT.

(2) We provide thinking that disagreement among multiple raters, i.e., uncertainty, can be quantified from soft labels and used as prior knowledge to reflect the pixel-level difficulty of an image.

(3) We propose to use consistency regularization to improve the model's attention to features and reduce dependence on GT, especially in regions of high uncertainty. Without introducing additional parameters, the accuracy of our method is improved over that of other methods on synthetic and real-world datasets.

2. Related Works

The problem of multirater labels' segmentation caused by inter-rater variability has started to pique the interest of researchers. There is a study showing that the observed labels depend on three causal factors: the true label, the expertise of the rater, and the image difficulty [16]. For the method of obtaining the true label, it is a common practice to use majority voting [18] and STAPLE [19] or other label fusion strategies to obtain the ground-truth labels [11, 20] so that they can be adapted to the general segmentation model. However, simple label fusion methods neither do take advantage of any image features nor do they carry the inter-

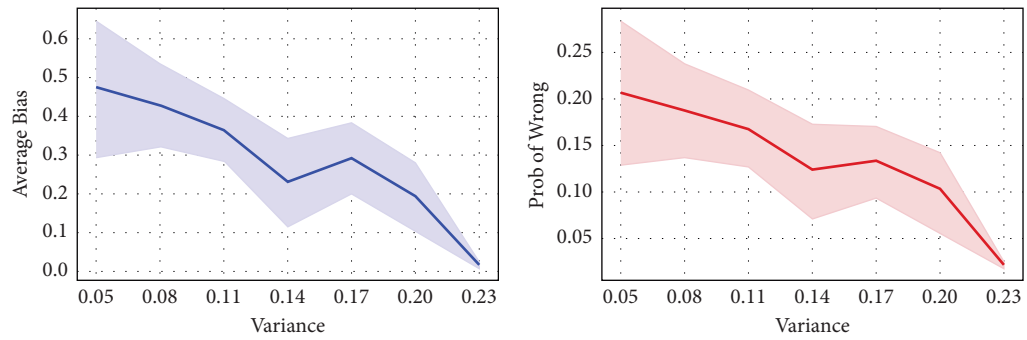

Figure 2: (a) Line graph of pixelwise interclass variances versus pixelwise loss; (b) line graph of pixelwise interclass variances versus pixelwise probability of misprediction. The abovementioned statistics are averaged on the validation set.

rater variability through to the model. Recently, several efforts have started to explore the expertise of the rater using label sampling strategies [21] or rater modeling strategies [22]. For instance, Zhang et al. [23] proposed to use confusion matrices to the model preference of annotators, obtaining segmentation prediction with the least noise by optimizing two coupled convolutional neural networks (CNNs). Yu et al. [15] proposed a multibranch model for the multirater glaucoma classification task, encouraging the specificity branch and the sensitivity branch to generate consistent/opposing predictions for consensus/disagreement samples. Ji et al. [24] proposed MRNet, which embeds the expertise of individual annotators into the model to generate calibrated predictions under different expertise levels for medical image segmentation.

However, there still lacks effective research on the image difficulty represented by image features in the multirater label segmentation task. Furthermore, we consider that multirater labels' segmentation is weakly supervised learning with inaccurate labels, which has not been explored before. Although our approach is uncertainty-driven, unlike works, such as Monte Carlo dropout [25] and ensembles [26, 27], that evaluate uncertainty and produce multiple segmentation hypotheses, our work aims to learn a deterministic single-output deep model.

3. Methodology

The main architecture of our model is illustrated in Figure 3, which is composed of three main parts: (a) segmentation network with consistency regularization for conveying more information about the input; (b) asymmetrical regularization part for generating uncertainty mask to realize local self-ensembling in supervised learning; (c) multirater labels fusion part for obtaining a soft label for each input as the supervised target containing uncertainty. In the test and application phase, just the trained network is required to predict the segmentation of the input image.

3.1. Problem Definition. In this article, we consider the problem of learning a segmentation model from labels annotated by multiple human raters. Given the images $\{X^{W\times H\times L} = x_n\}_{n=1}^{N}$ and the corresponding one-hot labels $\{Y^{W\times H\times C} = y_n^{(r)}\}_{n=1,\cdots,N}^{r=1,\cdots,R}$ (W, H, L, C denote the width, height, channels, and classes), where N is the number of samples and R is the number of raters, each image is independently annotated by raters based on their personal experiences. The objective of the multirater label segmentation task is to learn the projection function $F(\cdot)$, mapping the input image x_n to the estimated prediction $\widehat{y}_n$ which is one-hot form encoded by the full probability distribution $\widehat{p}_n$. In our article, $\widehat{p}_n$ is encouraged to be as similar as p_n, which is the soft label fused by Y_n.

3.2. Soft Labels. Recently, increasing studies have proposed training a model using soft labels for accounting for the high uncertainty in lesion or structure borders' delineation [11, 28–31]. Averaging multirater labels is an intuitive way to obtain soft labels in multirater annotation tasks as follows:

$$p_n = \frac{1}{R}\sum_{r=1}^{R} y_n^{(r)}. \tag{1}$$

Although the average strategy incorporates uncertainty from inter-rater variability into soft labels, it indulges the overconfidence of each rater. Therefore, we soften each hard label by SVLS and then average them to obtain p^n which contains spatial and inter-rater uncertainty as follows:

$$p_n^{(i,j)} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{\sum w}\sum_{a=1}^{3}\sum_{b=1}^{3} y_n^{r(i-a,j-b)} w^{(a,b)}, \tag{2}$$

where (i, j) is the position of pixel and w is a weight matrix which is obtained by $1/\sqrt{2\pi\sigma^2}e^{-|\vec{x}|^2/2\sigma^2}$ with $\sigma = 1$. SVLS determines the probability of the target pixel based on its neighboring pixels, achieved by a Gaussian-like weight matrix that is applied across the one-hot encoded rater labels $y^{(r)}$ to obtain a soft probability distribution.

The transmission of uncertainty information into the model is inseparable from the appropriate loss function. There is the performance of several common loss functions in Section 4.4, including soft cross-entropy loss, soft dice

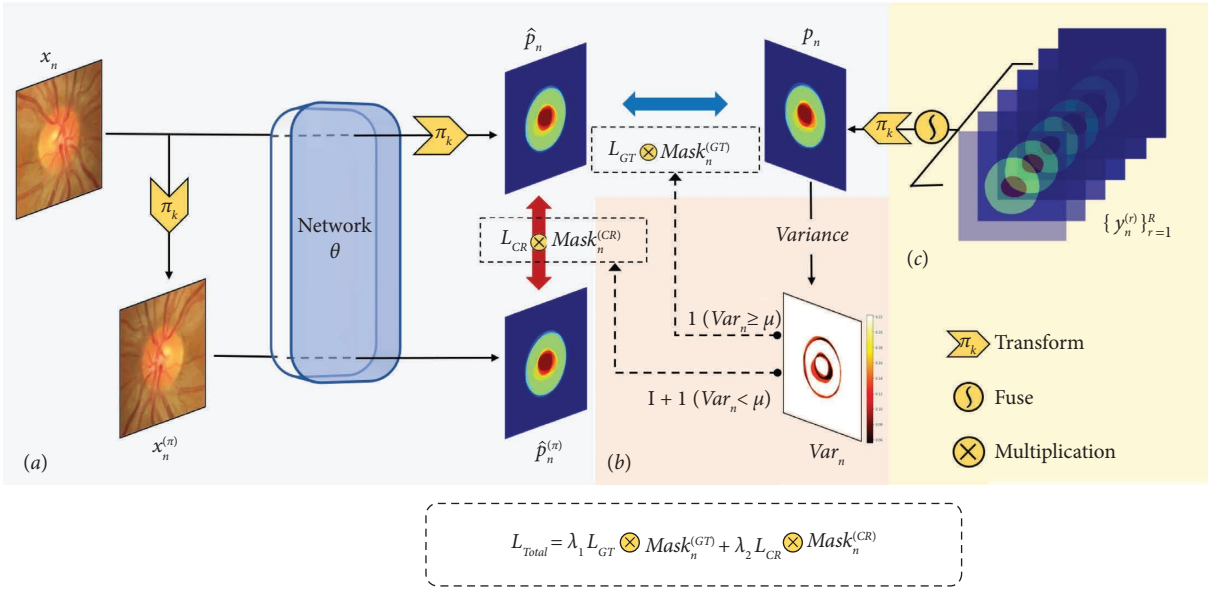

Figure 3: The architecture of our model consists of three parts: (a) segmentation network part; (b) asymmetrical regularization part; (c) multirater labels' fusion part.

loss, and soft focal loss. By comparison, the soft cross-entropy loss is selected as the optimization objective, encouraging the probability distribution of prediction $\hat{p}_n$ to be identical to that of the soft label as follows:

$$\mathscr{L}_{GT} = -\sum_{n=1}^{N} p_n \log\left(\hat{p}_n\right). \tag{3}$$

3.3. Label Uncertainty Measure. It is equally crucial to model uncertainty at the pixel-level as to improve the model's performance, particularly in medical scenarios [25]. Unlike work that uses stochastic networks [32, 33] to model uncertainty, we improve multirater models using uncertainty as a source of prior knowledge. Specifically, we consider the pixelwise interclass variance $\{\mathrm{Var}_n^{W\times H}\}_{n=1,\cdots,N}$ that reflects the uncertainty caused by inter-rater variability and spatial variation. It is inversely proportional to entropy, meaning that the lower the variance, the greater the entropy and, hence, the greater the uncertainty. The appropriate uncertainty can enhance the generalization and calibration of the model. However, the high uncertainty would be detrimental to the model as noise. In the position of the $(i, j)^{th}$ pixel, the variance between classes $\mathrm{Var}\,(p_n^{(i,j)})$ can be formulated by the following:

$$\mathrm{Var}\left(p_n^{(i,j)}\right) = \frac{1}{C}\sum_{c=0}^{C-1}\left(p_n^{(i,j)}(c) - \frac{1}{C}\right)^2, \tag{4}$$

where $\sum_{c=0}^{C-1} p_n^{(i,j)}(c) = 1$. We propose to use uncertainty as a threshold to assign different optimization objectives to different areas of the image. Specifically, the labels of areas with high uncertainty are no longer decisive but are replaced with constraints on the feature space, which will be clarified in the next section. In areas with high rater agreement, soft labels and feature constraints work together to optimize the model. For convenience, we refer to this uncertainty-driven local differential optimization as asymmetrical regularization. The threshold comes into play in the form of a mask of 0-1, acting directly on the loss function. Mask is differentiated into $\mathrm{mask}^{(GT)}$ and $\mathrm{mask}^{(CR)}$ based on the threshold, which correspond to areas of low and high uncertainty, respectively, as follows:

$$\begin{aligned} \mathrm{mask}_n^{(GT)} &= \mathbb{1}\left(\mathrm{Var}_n \geq \mu\right), \\ \mathrm{mask}_n^{(CR)} &= \mathbf{I} + \mathbb{1}\left(\mathrm{Var}_n < \mu\right), \end{aligned} \tag{5}$$

where $\mathbb{1}$ is the indicator function and $\mathbf{I}$ is the identity matrix with the same shape as Var_n.

3.4. Consistency Regularization. To improve attention to features and optimize feature space, we propose using consistency regularization as an extra constraint on the model, which has been utilized in semisupervised learning [34] and unsupervised learning [35]. Consistency regularization is a type of self-ensemble learning because it only relies on the images themselves to learn. Inspired by Li et al. [34], we apply rotation consistency to this work. Specifically, there is a problem in the segmentation task using CNN: when the inputs of CNN are rotated, the corresponding network predictions would not be rotated in the same way [36] as follows:

$$\theta\left(\pi_k x_n\right) \neq \pi_k \theta\left(x_n\right), \tag{6}$$

where π_k is a rotation to the image (i.e., horizontal, vertical, or mixed flip) and θ is the parameters of the network. The feature space is automatically optimized when the model is encouraged to make the same judgments about elements before and after rotation. In this article, we use the soft cross-entropy loss function as the optimization target of rotation consistency regularization term:

$$\mathscr{L}_{CR} = -\sum_{n=1}^{N} \theta(\pi_k x_n)\log\left(\pi_k \theta(x_n)\right). \tag{7}$$

When formulas (3) and (7) are combined, the total loss function is as follows:

$$\mathscr{L}_{total} = -\lambda_1 \sum_{n=1}^{N} p_n \log(\widehat{p}_n) - \lambda_2 \sum_{n=1}^{N} \theta(\pi_k x_n)\log\left(\pi_k \widehat{p}_n\right), \tag{8}$$

where $\lambda_1 + \lambda_2 = 1$. By minimizing the loss function, the network is urged to focus more on the image content than on the regression of GT alone [37]. So far, image features can be fully expressed through self-ensembling. In regions of divergence where uncertainty is high, supervision of labels is entirely replaced by unsupervised self-ensembling. Without introducing extra parameters and structures, asymmetrical regularization is achieved by covering the soft label with an uncertainty-based mask (formulas (5) and (7)). Finally, updated formula (8) is shown as follows:

$$\mathscr{L}_{\text{total}} = -\lambda_1 \text{mask}_n^{(GT)} \sum_{n=1}^{N} p_n \log(\widehat{p}_n) + -\lambda_2 \text{mask}_n^{(CR)} \sum_{n=1}^{N} \theta(\pi_k x_n)\log\left(\pi_k \widehat{p}_n\right). \tag{9}$$

4. Experiments

In this section, we introduce the experimental dataset, implementation details, and evaluation metrics. In order to explore the best performance under different combinations of the loss and uncertainty threshold value, we conduct quantitative experiments with different setups on the MNIST and the RIGA validation set in Section 4.4. For comparison with other methods, the common label fusion approach and other SOTA approaches for multirater labels segmentation are used as the benchmark. The results are listed in Section 4.5, showing that our method can exploit the uncertainty of multirater annotations to improve segmentation performance. Additionally, ablation experiments are conducted to evaluate the efficacy of each component of our method.

4.1. Datasets

(i) MNIST is a handwritten digits dataset with 60,000 training and 10,000 test examples. All images are 28×28 grayscale versions of the handwritten numbers 0–9. Zhang et al. [23] synthesized a dedicated dataset of multirater annotation tasks based on MNIST, which simulates raters with different biases to obtain multiple labels by using Morpho-MNIST software [38]. Specifically, the first rater provides good segmentation with approximate GT, the second rater tends to oversegment, the third rater tends to undersegmentation, the fourth rater is prone to the combination of small fractures and oversegmentation, and the fifth rater always annotates everything as the background. We train a model using all five raters' annotations and finally test the model performance on GT.

(ii) RIGA is a publicly available dataset for joint OC and OD segmentation from the University of Michigan [39]. It includes a total of 750 color fundus images from three subsets: 460 images from MESSIDOR, 195 images from BinRushed, and 95 images from Magrabia. Each fundus image has six OC and OD annotations carried out by six ophthalmologists. We select BinRushed and MESSIDOR as the training set, and Magrabia is selected as the test set, where all images are resized to 256×256. In accordance with the experimental design of MRNet [24] and other methods [11, 23], the majority voting of six raters for each test image is used as the silver standard to evaluate the prediction.

(iii) QUBIQ-Kidney and Prostate are subdatasets of Quantification of Uncertainties in Biomedical Image Quantification Challenge (QUBIQ) [40], which are specifically designed to evaluate inter-rater variability. The QUBIQ-Kidney images are 2D CT slices (20 cases for training and 4 cases for testing) in which the kidneys are manually annotated by three raters. The QUBIQ-Prostate images are 2D MRI slices (48 cases for training and 7 cases for testing) in which the prostate is manually annotated by six raters. To match the task objective of the QUBIQ challenge, GT and prediction are binarized at five probability levels (0.1, 0.3, 0.5, 0.7, and 0.9), and evaluation scores for all thresholds will be averaged.

4.2. Implementation Details. For a fair comparison, we employ the same network architecture as the baseline approach. Specifically, for the MNIST experiment, we use the U-Net architecture without pretraining as [23].Moreover, for the RIGA experiment, the main framework utilizes the U-Net architecture with ResNet34 as the backbone. Parameters of the U-Net encoder are initialized with the pretrained model on ImageNet [41]. The abovementioned network is implemented with the PyTorch platform and trained/tested on a Tesla V100 GPU with 32 GB of memory. The proposed network is trained end-to-end using the Adam optimizer [42], and it takes about 4 hours to train our model with a mini-batch size of 4 for 60 epochs. The learning rate is set to 1×10^{-4}.

4.3. Evaluation Metrics. Various evaluation metrics, including the Dice similarity coefficient (DSC) and mean intersection over union (mIoU), were utilized to evaluate the performance of the proposed method for segmenting OC and OD relative to GT. These performance metrics are defined as follows:

$$\begin{aligned} \mathrm{DSC}(\mathscr{D}) &= \frac{2 \times TP}{2 \times TP + FP + FN}, \\ \mathrm{mIoU}(\mathscr{I}) &= \frac{TP}{FP + FN + TP}, \end{aligned} \tag{10}$$

where TP, FP, FN, and TN represent true positives, false positives, false negatives, and true negatives, respectively, in the evaluation confusion matrix. Note that a model with higher metric values can predict more precise segmentation masks. All experimental results are reported as the average of the ten experiments conducted on the test set.

4.4. Performance of Our Methods. Here, we provide a quantitative comparison among different loss functions including soft cross-entropy loss (CE), soft dice loss (DL), and soft focal loss [43] (FL). Table 1 displays the top five combinations of loss functions with the highest $\mathscr{D}$ under the corresponding optimal hyperparameter settings, including the uncertainty threshold μ and the unsupervised loss weight λ_2, where experiments are performed on the MNIST validation set. The proposed method exhibits the best performance when both the supervision loss and the consistency regularization loss are CE. In Figure 4, we further present the comparison of model accuracy at different μ and λ_2 settings under this loss function combination on the MNIST and RIGA validation set. Moreover, the average unsupervised proportion corresponding to different μ in the two datasets is performed in Figure 4. By comparison, the optimal (μ, λ_2) combinations on the MNIST and RIGA datasets are (0.5, 0.005) and (0.5, 0.002), respectively, with corresponding unsupervised proportions of 8% and 4%.

4.5. Comparisons with Other Methods. To demonstrate the advantage of the proposed method, we compare our method to the SOTA methods on the MNIST and RIGA datasets. We use the publicly released code with default parameters to retrain the SOTA methods with the same training/test set as ours for a fair comparison.

Table 2 quantitatively compares our framework to three hard label methods, five soft label methods, and other SOTA multirater labels' segmentation methods, including (a) Mode-UNet: UNet trained using a single label randomly selected; (b) MV-UNet: UNet trained using one-hot labels obtained by majority voting; (c) STAPLE-UNet: UNet trained using one-hot labels obtained by STAPLE [19]; (d) Average-UNet: UNet trained using soft labels obtained by average raters [31]; (e) GLS-UNet: UNet trained using soft labels smoothed by general label smoothing [44]; (f) Sharpen-UNet: UNet trained using soft labels smoothed by label sharpen [45] under temperature $(T) = 0.5$ or 1.5; (g) Mixup-UNet: UNet trained using soft labels smoothed by Mixup; (h) SVLS-UNet: UNet trained using soft labels smoothed by SVLS [11]; (i) LNL [23]; (j) MRNet [24] on the MNIST and RIGA test sets.

Table 1: The segmentation performance (mean ± standard deviation) of different combinations of loss functions: supervised loss + consistency regularization loss.

Loss	μ	λ_2	Performance $\mathscr{D} \pm$ std (%)
CE and CE	0.005	0.5	**94.09 ± 0.51**
CE and DL	0.005	0.5	93.55 ± 0.78
DL and DL	0.005	0.1	92.39 ± 1.09
DL and CE	0.005	0.1	91.15 ± 1.16
FL and FL	0.005	0.5	89.91 ± 0.59

As shown in Table 2, our proposed method consistently achieves superior performance compared with other methods. For value $\mathscr{D}$, our method outperforms the SOTA method by 1.13% on the synthetic MNIST dataset. Figure 5 shows the visualization results, wherein our method recovers the most realistic result from several annotations containing obvious human errors. Additionally, compared to the suboptimal MRNet method, the segmentation results predicted by the proposed method are smoother and more structured at the edge. For the real-world dataset RIGA, the performance improvement is especially prominent for the retinal OC segmentation, where the inter-rater variability is more significant, with a 2.29% increase in $\mathscr{D}$ value over the current best method (listed in Table 2).

Figure 6(a) visualizes five examples of the silver standard and the corresponding segmentation results predicted by six different methods. As shown in Figure 6(b), the edge of OC occluded by blood vessels in the area with high inter-rater divergence (indicated by arrows) and, similarly, the high-incidence area of misprediction (red areas) by other methods. Compared to other methods, the proposed method shows lower prediction errors in the aforementioned area, demonstrating the robustness of our method to difficult features.

4.6. Ablation Studies. In this section, ablation studies are performed on the RIGA dataset over each component of the proposed method, including label smoothing (LS), consistency regularization (CR), and asymmetrical regularization (AR), as listed in Table 3. Meanwhile, the effect of different label smoothing techniques including GLS and Sharpen and SVLS on the performance of our method is also explored. The baseline model is the UNet trained using soft labels of average raters. All experiments are performed with the same network structure and training hyperparameters, Sections 4.2 and 4.4. $\mathscr{D}_{\text{ave}}$ represents the average value of $\mathscr{D}_{OC}$ and $\mathscr{D}_{OD}$.

As shown in Table 3, the segmentation performance of the model reaches SOTA when all components are activated. As we sequentially remove the proposed components from the U-Net Baseline, the model performance degrades gradually. In particular, the inclusion of CR improved the

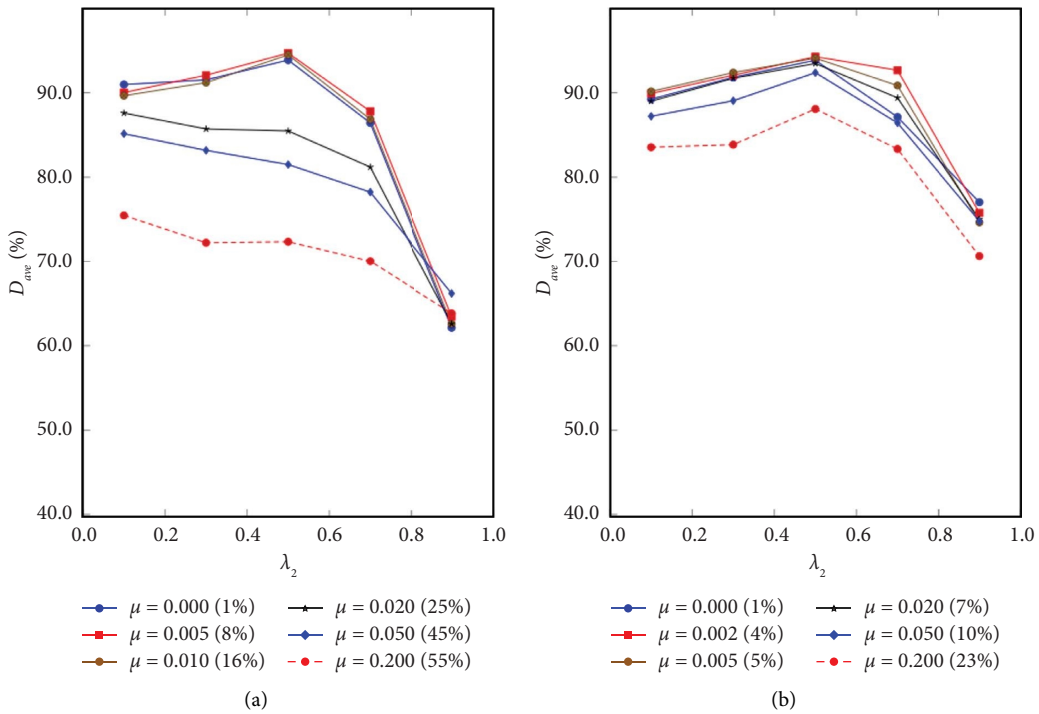

Figure 4: Segmentation accuracy under different μ and λ_2 on the (a) MNIST and (b) RIGA.

Table 2: Quantitative results with different strategies on the MNIST and RIGA test set.

	Methods	MNIST		RIGA			
		$\mathscr{D}$ (%)	$\mathscr{J}$ (%)	$\mathscr{D}_{OD}$ (%)	$\mathscr{D}_{OC}$ (%)	$\mathscr{J}_{OD}$ (%)	$\mathscr{J}_{OC}$ (%)
Hard	Mode-UNet	62.89	57.30	96.90	82.41	94.62	75.09
	MV-UNet	89.14	80.59	97.03	84.92	94.35	73.47
	STAPLE-UNet	82.26	74.51	96.28	85.37	92.84	75.68
Soft	Average-UNet	90.54	82.85	97.04	85.40	94.52	76.58
	GLS-UNet	87.32	78.29	96.14	86.83	93.71	75.95
	Sharpen$^{(T=0.5)}$-UNet	90.50	81.03	96.85	84.71	94.33	77.18
	Sharpen$^{(T=1.5)}$-UNet	87.67	80.13	96.77	86.13	93.82	77.90
	Mixup-UNet	86.61	78.58	96.83	84.72	94.02	75.18
	SVLS-UNet	90.32	82.05	97.40	86.09	94.95	76.87
SOTA	LNL	84.52	76.33	*97.67*	*87.56*	95.46	*78.76*
	MRNet	*93.63*	*88.09*	97.60	86.54	*95.78*	78.19
	Ours	**94.76**	**90.82**	**97.98**	**89.85**	**96.04**	**81.97**

The best results are highlighted, and the second best results are italic.

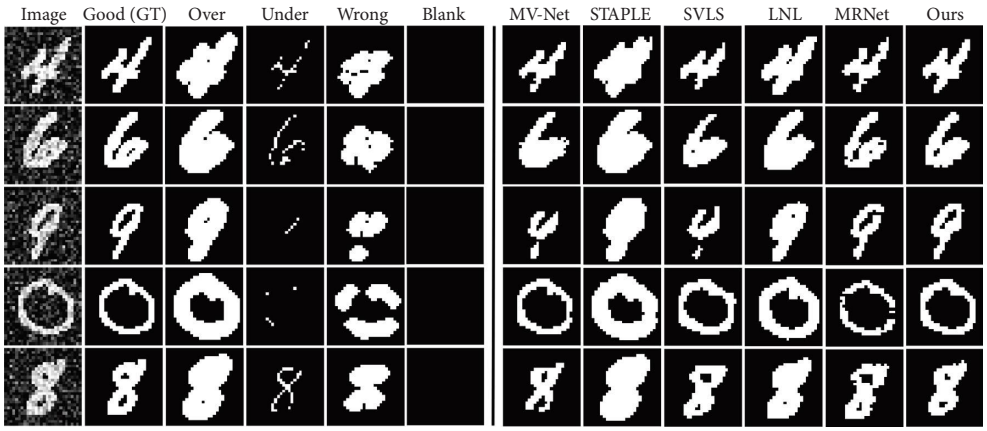

Figure 5: Visualization of five raters' annotations and predictions of six methods on the synthetic MNIST test set.

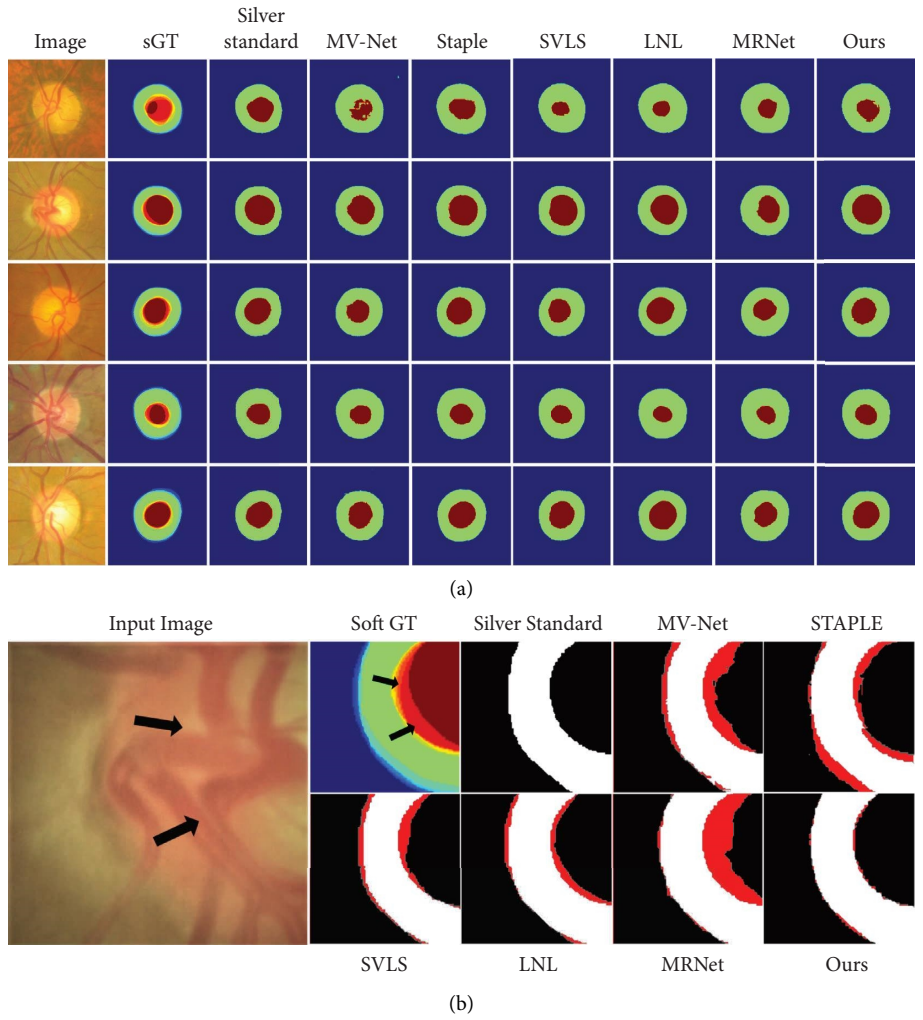

Figure 6: (a) Visualization of segmentation predictions on RIGA test set. (b) An example of a visualized partial map of predicted errors.

Table 3: Ablation experiment results on the RIGA dataset.

Module			Performance
LS	CR	AR	$\mathscr{D}_{ave} \pm$ std (%)
SVLS	√	√	**93.92 ± 0.56**
Sharpen$^{(1.5)}$	√	√	93.55 ± 0.73
GLS	√	√	93.49 ± 0.63
SVLS	√		93.36 ± 0.71
Average	√	√	92.87 ± 0.65
Average	√		92.32 ± 0.82
SVLS			91.74 ± 0.78
Average			91.22 ± 1.06

baseline by 1.10%. Then, the combination of CR and AR yielded an additional 0.55% improvement. This means that, in areas with higher rater inconsistency in annotations, the potential representation of image features is more reliable than in uncertain annotations. It is proved experimentally that features are also one of the important causes of observer variability rather than just the rater knowledge. In addition, adding SVLS alone improves $\mathscr{D}_{\text{ave}}$ of baseline by 0.52% while utilizing it with CR and AR jointly improves $\mathscr{D}_{\text{ave}}$ of the proposed method without SVLS by 1.05%. It demonstrates that the positive effects of CR and AR are further strengthened under the threshold of uncertainty with SVLS.

4.7. Generalization Capability. To further verify the generalization capability of the proposed method, we additionally perform experiments on the kidney segmentation task of the QUBIQ multirater segmentation challenge. We use the same

Table 4: Quantitative results with different strategies on the QUBIQ-kidney and prostate test sets.

Methods	Kidney		Prostrate		#Parameters
	$\mathscr{D}^{(\text{soft})}$ (%)	$\mathscr{J}^{(\text{soft})}$ (%)	$\mathscr{D}^{(\text{soft})}$ (%)	$\mathscr{J}^{(\text{soft})}$ (%)	
MV-UNet	66.59	57.83	83.50	73.71	22.0M
STAPLE-UNet	65.01	56.31	83.36	73.69	22.0M
Average-UNet	69.33	58.21	85.82	77.02	22.0M
SVLS-UNet	70.04	58.65	86.11	77.38	22.0M
LNL	68.40	58.59	85.44	76.91	22.2M
MRNet	**71.36**	**60.43**	*87.39*	*78.14*	81.1M
Ours	*70.25*	*59.08*	**87.67**	**78.55**	22.0M

The best results are highlighted, and the second best results are italic.

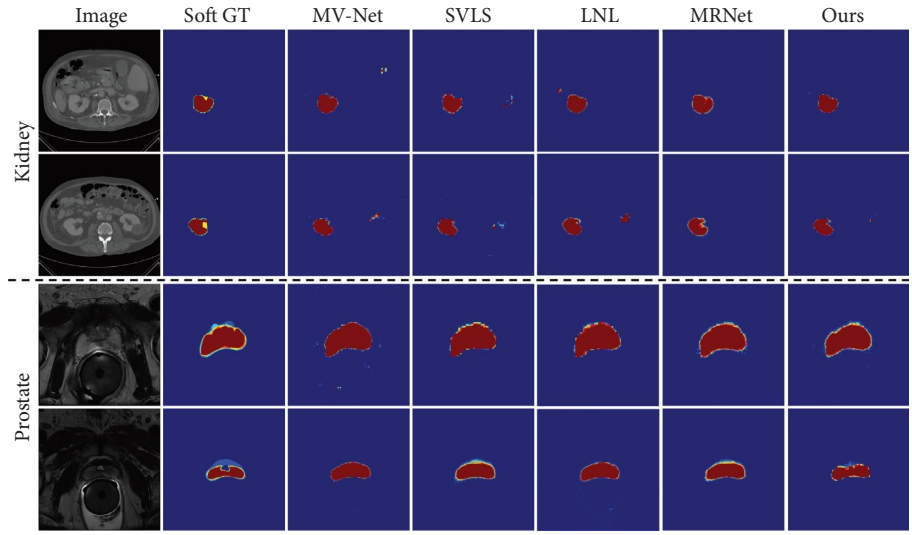

Figure 7: Visualization of segmentation predictions on the QUBIQ-kidney and prostate test sets.

multithreshold scores $\mathscr{D}^{(\text{soft})}$ and $\mathscr{J}^{(\text{soft})}$ as QUBIQ challenge, which can better evaluate the ability of the model to reflect potential inter-rater agreement/disagreement. Specifically, after the GT and prediction are binarized at multiple threshold levels (0.1, 0.3, 0.5, 0.7, and 0.9), the $\mathscr{D}$ and $\mathscr{J}$ metrics averaged across five thresholds are $\mathscr{D}^{(\text{soft})}$ and $\mathscr{J}^{(\text{soft})}$. As listed in Table 4, compared to the comparative methods, the proposed method achieves optimal performance on the QUBIQ-Prostate dataset and achieves suboptimal performance on the QUBIQ-Kidney dataset. Furthermore, the advantage of a low number of parameters facilitates the application of our method to other multirater datasets. Several representative examples of the comparison methods for such two datasets are visualized in Figure 7.

5. Conclusion

In this article, we focus on the utilization of rich annotation information from multiple clinical raters, which are relatively less explored but widely presented in medical image segmentation. Based on the deep learning method using soft labels, we proposed a local self-ensembling learning model related to pixelwise variance with the intention of reducing the reliance upon uncertain local labels and optimizing the feature space. Our method achieves performance improvement over the soft labels' learning method without requiring the introduction of extra parameters and structures. In addition, we incorporate structural uncertainty into soft labels via the label smoothing technique to further improve segmentation performance level. Empirical experiments demonstrated the overall superior performance of our method on a synthetic dataset and a real-world dataset. Our method provides a solution for automatically learning a reliable clinical-aided diagnosis system using multirater annotations.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61773246 and 81871508), the Taishan Scholar Project of Shandong Province (TSHW201502038), and the Natural Science Foundation of Shandong Province (ZR2018ZB0419).

References

[1] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, pp. 2063–2079, 2018.

[2] G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, Chicago, IL, USA, 2008.

[3] L. Aroyo and C. Welty, "Truth is a lie: crowd truth and the seven myths of human annotation," *AI Magazine*, vol. 36, no. 1, pp. 15–24, 2015.

[4] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2020.

[5] P. Brennan and A. Silman, "Statistical methods for assessing observer variability in clinical measures," *BMJ British Medical Journal*, vol. 304, no. 6840, pp. 1491–1494, 1992.

[6] P. Bridge, A. Fielding, P. Rowntree, and A. Pullar, "Intra-observer variability: should we worry?" *Journal of Medical Imaging and Radiation Sciences*, vol. 47, no. 3, pp. 217–220, 2016.

[7] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in CT," *European Radiology*, vol. 29, no. 3, pp. 1391–1399, 2019.

[8] A. S. Becker, K. Chaitanya, K. Schawkat et al., "Variability of manual segmentation of the prostate in axial T2-weighted MRI: a multi-reader study," *European Journal of Radiology*, vol. 121, Article ID 108716, 2019.

[9] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9617–9626, Long Beach, CA, USA, August 2019.

[10] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, and M. Poesio, "Beyond black & white: leveraging annotator disagreement via soft-label multi-task learning," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 2591–2597, Mexico City, Mexico, June 2021.

[11] M. Islam and B. Glocker, "Spatially varying label smoothing: capturing uncertainty from expert annotations," in *Proceedings of the Information Processing in Medical Imaging*, pp. 677–688, Springer, Berlin, Germany, June 2021.

[12] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," 2017, https://arxiv.org/abs/1705.10694.

[13] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[14] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 648–660, 2018.

[15] S. Yu, H.-Y. Zhou, K. Ma et al., "Difficulty-aware glaucoma classification with multi-rater consensus modeling," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 741–750, Springer, Berlin, Germany, October 2020.

[16] J. Whitehill, T.-f. Wu, J. Bergsma, J. Movellan, and P. Ruvolo, "Whose vote should count more: optimal integration of labels from labelers of unknown expertise," *Advances in Neural Information Processing Systems*, vol. 22, pp. 2035–2043, 2009.

[17] R. Tudor Ionescu, B. Alexe, M. Leordeanu, M. Popescu, D. P. Papadopoulos, and V. Ferrari, "How hard can it be? Estimating the difficulty of visual search in an image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2157–2166, Las Vegas, NV, USA, June 2016.

[18] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[19] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[20] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.

[21] M. H. Jensen, D. R. Jørgensen, R. Jalaboi, M. E. Hansen, and M. A. Olsen, "Improving uncertainty estimation in convolutional neural networks using inter-rater agreement," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 540–548, Springer, Berlin, Germany, June 2019.

[22] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, "Learning from noisy labels by regularized estimation of annotator confusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11244–11253, Long Beach, CA, USA, June 2019.

[23] L. Zhang, R. Tanno, M.-C. Xu et al., "Disentangling human error from ground truth in segmentation of medical images," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15750–15762, 2020.

[24] W. Ji, S. Yu, J. Wu et al., "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12341–12351, Nashville, TN, USA, June 2021.

[25] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] S. Lee, S. Purushwalkam Shiva Prakash, M. Cogswell, V. Ranjan, D. Crandall, and D. Batra, "Stochastic multiple choice learning for training diverse deep ensembles," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[27] C. Rupprecht, I. Laina, R. DiPietro et al., "Learning in an uncertain world: representing ambiguity through multiple hypotheses," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3591–3600, Venice, Italy, October 2017.

[28] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation," in *Proceedings of the International Symposium on Biomedical Imaging*, pp. 1563–1566, IEEE, Venice, Italy, June 2019.

[29] H. Li, D. Wei, S. Cao, K. Ma, L. Wang, and Y. Zheng, "Superpixel-guided label softening for medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 227–237, Springer, Berlin, Germany, October 2020.

[30] C. Gros, A. Lemay, and J. Cohen-Adad, "SoftSeg: advantages of soft versus binary training for image segmentation," *Medical Image Analysis*, vol. 71, Article ID 102038, 2021.

[31] J. Lourenço-Silva and A. L. Oliveira, "Using soft labels to model uncertainty in medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 585–596, Springer, Berlin, Germany, June 2022.

[32] S. Hu, D. Worrall, S. Knegt, B. Veeling, H. Huisman, and M. Welling, "Supervised uncertainty quantification for segmentation with multiple annotations," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 137–145, Springer, Berlin, Germany, June 2019.

[33] M. Monteiro, L. Le Folgoc, D. Coelho de Castro et al., "Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12756–12767, 2020.

[34] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2021.

[35] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[36] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: deep translation and rotation equivariance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5028–5037, Honolulu, HI, USA, July 2017.

[37] S. Wang, C. Li, R. Wang et al., "Annotation-efficient deep learning for automatic medical image segmentation," *Nature Communications*, vol. 12, pp. 5915–6013, 2021.

[38] D. C. Castro, J. Tan, B. Kainz, E. Konukoglu, and B. Glocker, "Morpho-MNIST: quantitative assessment and diagnostics for representation learning," *Journal of Machine Learning Research*, vol. 20, pp. 1–29, 2019.

[39] A. Almazroa, S. Alodhayb, E. Osman et al., "Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images," *International Ophthalmology*, vol. 37, no. 3, pp. 701–717, 2017.

[40] B. Menze, Joskowicz, S. Bakas, A. Jakab, E. Konukoglu, and A. Becker, "Quantification of uncertainties in biomedical image quantification challenge at miccai," 2020, https://qubiq.grand-challenge.org/.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, https://arxiv.org/abs/1711.05101.

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

[44] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in Neural Information Processing Systems*, vol. 32, pp. 4694–4703, 2019.

[45] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: a holistic approach to semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.