Gradient and Feature Conformity-Steered Medical Image Classification with Noisy Labels

Xiaohan Xing[1], Zhen Chen[2], Zhifan Gao[3], and Yixuan Yuan[4(✉)]

[1] Department of Radiation Oncology, Stanford University, Stanford, CA, USA
[2] Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, NT, Hong Kong SAR, China
[3] School of Biomedical Engineering, Sun Yat-sen University, Guangdong, China
[4] Department of Electronic Engineering, Chinese University of Hong Kong, NT, Hong Kong SAR, China
yxyuan@ee.cuhk.edu.hk

Abstract. Noisy annotations are inevitable in clinical practice due to the requirement of labeling efforts and expert domain knowledge. Therefore, medical image classification with noisy labels is an important topic. A recently advanced paradigm in learning with noisy labels (LNL) first selects clean data with small-loss criterion, then formulates the LNL problem as semi-supervised learning (SSL) task and employs Mixup to augment the dataset. However, the small-loss criterion is vulnerable to noisy labels and the Mixup operation is prone to accumulate errors in pseudo labels. To tackle these issues, we present a two-stage framework with novel criteria for clean data selection and a more advanced Mixup method for SSL. In the clean data selection stage, based on the observation that gradient space reflects optimization dynamics and feature space is more robust to noisy labels, we propose two novel criteria, i.e., *Gradient Conformity-based Selection* (GCS) and *Feature Conformity-based Selection* (FCS), to select clean samples. Specifically, the GCS and FCS criteria identify clean data that better aligns with the class-wise optimization dynamics in the gradient space and principal eigenvector in the feature space. In the SSL stage, to effectively augment the dataset while mitigating disturbance of unreliable pseudo-labels, we propose a *Sample Reliability-based Mixup* (SRMix) method which selects mixup partners based on their spatial reliability, temporal stability, and prediction confidence. Extensive experiments demonstrate that the proposed framework outperforms state-of-the-art methods on two medical datasets with synthetic and real-world label noise. The code is available at https://github.com/hathawayxxh/FGCS-LNL.

Keywords: Label noise · Gradient conformity · Feature eigenvector conformity · Mixup

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_8.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 75–84, 2023.
https://doi.org/10.1007/978-3-031-43987-2_8

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in medical image classification. However, the great success of DNNs relies on a large amount of training data with high-quality annotations, which is practically infeasible. The annotation of medical images requires expert domain knowledge, and suffers from large intra- and inter-observer variability even among experts, thus noisy annotations are inevitable in clinical practice. Due to the strong memorization ability, DNNs can easily over-fit the corrupted labels and degrade performance [1,2], thus it is crucial to train DNNs that are robust to noisy labels.

An effective paradigm in learning with noisy labels (LNL) first selects clean samples, then formulates the LNL problem as semi-supervised learning (SSL) task by regarding the clean samples as a labeled set and noisy samples as an unlabeled set [3]. However, both the clean data selection and SSL stages in existing methods have some drawbacks. In the clean data selection stage, most existing studies rely on the small-loss [3,4] or high-confidence criteria [5] of individual samples, but neglect the global contextual information and high-order topological correlations among samples, thus unavoidably resulting in confirmation bias [6]. Besides, the above criteria in the output space are directly supervised and easily affected by corrupted labels [7]. Previous studies indicate that optimization dynamics (characterized by sample gradients) can reflect the true class information [8] and feature space is more robust to noisy labels, thus can provide more robust criteria for clean data selection [6,7,9]. Therefore, we aim to achieve more accurate clean data selection by exploring the topological correlation and contextual information in the robust gradient and feature spaces.

In the SSL stage, most existing studies [3,10] estimate pseudo labels for all samples and employ Mixup [11,12] to linearly interpolate the input samples and their pseudo labels for model training. Compared with previous methods that train DNNs by reweighting samples [13] or utilizing clean data only [4,14], the Mixup [12] operation can effectively augment the dataset and regularize the model from over-fitting. However, as the pseudo labels of noisy datasets cannot be always reliable, the traditional Mixup method which randomly chooses the mixup partner for each sample may accumulate errors in pseudo labels. Therefore, it is highly desirable to design a novel Mixup method that can select reliable mixup partners and mitigate the interference of unreliable pseudo labels.

In this paper, we present a novel two-stage framework to combat noisy labels in medical image classification. In the clean data selection stage, we propose a gradient and feature conformity-based method to identify the samples with clean labels. Specifically, the *Gradient Conformity-based Selection* (GCS) criterion selects clean samples that show higher conformity with the principal gradient of its labeled class. The *Feature Conformity-based Selection* (FCS) criterion identifies clean samples that show better alignment with the feature eigenvector of its labeled class. In the SSL stage, we propose a *Sample Reliability-based Mixup* (SRMix) to augment the training data without aggravating the error accumulation of pseudo labels. Specifically, SRMix interpolates each sample with reliable mixup partners which are selected based on their spatial reliability, temporal stability, and prediction confidence. Our main contributions are as follows:

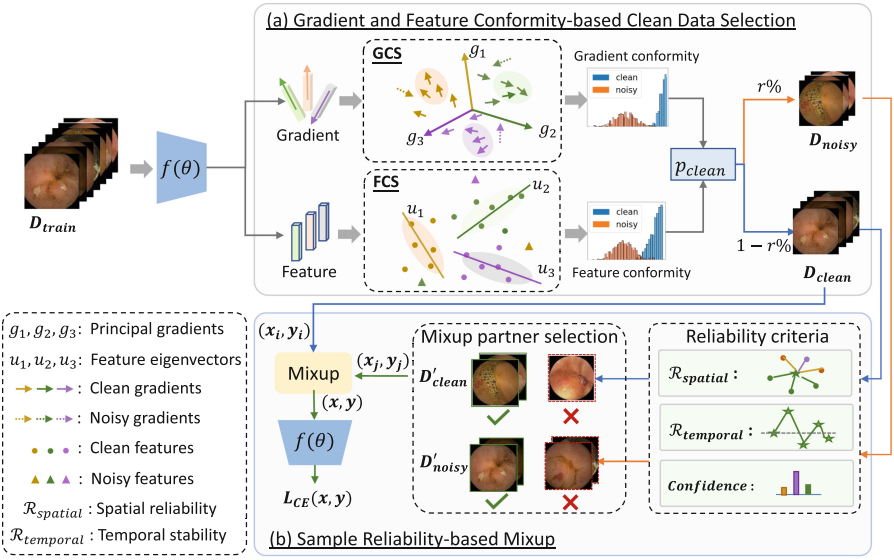

Fig. 1. The framework of our method. (a) Gradient and feature conformity-based clean data selection module, including GCS and FCS criteria, divides training samples into a clean set D_{clean} and a noisy set D_{noisy}. (b) Sample Reliability-based Mixup (SRMix) module interpolates each sample (x_i, y_i) with a reliable mixup partner (x_j, y_j).

- We devise two novel criteria (i.e., GCS and FCS) to improve clean data selection by exploring the topological correlation and contextual information in the gradient and feature spaces.
- We propose a novel SRMix method that selects reliable mixup partners to mitigate the error accumulation of pseudo labels and improve model training.
- Extensive experiments show that our proposed framework is effective in combating label noise and outperforms state-of-the-art methods on two medical datasets with both synthetic and real-world label noise.

2 Method

An overview of our proposed two-stage framework is shown in Fig. 1. The training dataset is denoted as $D_{train} \in \{(x_i, y_i)\}_{i=1}^{N}$, where the given label y_i could be noisy or clean. In the clean data selection stage, we propose a gradient and feature conformity-based method to distinguish clean samples from the noisy dataset. As shown in Fig. 1 (a), the GCS computes the principal gradient of each class (i.e., g_1, g_2, g_3) to represent its optimization dynamics, and measures the label quality of each sample by its gradient conformity with the class-wise principal gradient. The FCS computes the principal feature eigenvector of each class to reflect its contextual information, and measures the label quality of each sample by its feature conformity with the class-wise feature eigenvector. Based

on the integration of these two criteria, the training data is divided into a noisy set D_{noisy} and a clean set D_{clean}. In the SSL stage (see Fig. 1 (b)), our SRMix module interpolates each sample (x_i, y_i) with a reliable mixup partner (x_j, y_j), which is selected based on its spatial reliability, temporal stability, and prediction confidence. The mixed samples are used for model training.

2.1 Gradient and Feature Conformity-Based Clean Data Selection

Inspired by previous studies that optimization dynamics in the gradient space reflects the true class information [8,15] and contextual information in the feature space is more robust to noisy labels [7], we devise the novel GCS and FCS criteria to measure label quality in the gradient and feature spaces.

Gradient Conformity-Based Selection (GCS). The GCS aims to distinguish clean samples from noisy ones by exploring their optimization dynamics in the gradient space. Since training samples from the same class usually exhibit similar optimization dynamics [15], the gradient of a sample should be similar to the principal gradient of its true class, thus we use the gradient conformity as a criterion to evaluate the quality of its given label. Specifically, for each sample x_i, its gradient $g(x_i)$ is computed as:

$$g(x_i) = \frac{\partial(-log p^{'}(x_i))}{\partial f(x_i)}, \quad p^{'}(x_i) = max \sum\nolimits_{x_j} p(x_j), \forall x_j \in KNN\{x_i\}, \tag{1}$$

where $f(x_i)$ is the feature vector of the sample x_i, and x_j denotes the K-Nearest Neighbors (KNN) of x_i. $p^{'}(x_i)$ is the probability of the most likely true class predicted by its KNN neighbors. Therefore, the gradient $g(x_i)$ is very likely to reflect the true class information and optimization dynamics of x_i. For each class, we select $\alpha\%$ samples with the smallest loss as an anchor set $\mathcal{A}_c$, which is depicted in the shaded areas of the GCS in Fig. 1 (a). Then, the principal gradient of the c-th class is computed as:

$$g_c = \frac{1}{N_c \cdot \alpha\%} \sum\nolimits_{x_i} g(x_i), \quad x_i \in \mathcal{A}_c, \tag{2}$$

which is the average gradient of all samples in the anchor set $\mathcal{A}_c$ of the c-th class. Then, we can measure the similarity between the gradient of the sample x_i and the principal gradient of class y_i with the cosine similarity $s_g(x_i) = cos < g(x_i), g_{y_i} >$. For the sample x_i, if y_i is a noisy label, $g(x_i)$ should be consistent with the principal gradient of its true class and diverge from g_{y_i}, thus yielding small $s_g(x_i)$. By fitting Gaussian mixture models (GMM) on the similarity score $s_g(x_i)$, we can get $c_g(x_i) = GMM(s_g(x_i))$, which represents the clean probability of the sample x_i decided by the GCS criterion. To the best of our knowledge, this is the first work that explores gradient conformity for clean data selection.

Feature Conformity-Based Selection (FCS). Since feature space is more robust to noisy labels than the output space [7], our FCS criterion explores high-order topological information in the feature space and utilizes the feature conformity with class-wise principal eigenvectors as a criterion to select clean samples. Specifically, for each class, we compute the gram matrix as:

$$M_c = \sum\nolimits_{x_i} f(x_i) \cdot f(x_i)^T, \quad x_i \in \mathcal{A}_c, \tag{3}$$

where $f(x_i)$ denotes the feature vector of the sample x_i in the anchor set $\mathcal{A}_c$ of the c-th class. Then, we perform eigen-decomposition on the gram matrix: $M_c = U_c \cdot \Sigma_c \cdot U_c^T$, where U_c is the eigenvector matrix and Σ_c is a diagonal matrix composed of eigenvalues. The principal eigenvector u_c of U_c is utilized to represent the distribution and contextual information of the c-th class. Then, for each sample x_i, we measure its label quality based on the conformity of its feature $f(x_i)$ with the principal eigenvector u_{y_i} of its given label: $s_f(x_i) = cos < f(x_i), u_{y_i} >$. Samples that better align with the principal eigenvectors of their labeled class are more likely to be clean. According to the FCS criterion, the clean probability of the sample x_i is obtained by $c_f(x_i) = GMM(s_f(x_i))$. Compared with existing methods that utilize class-wise average features to represent contextual information [7], the eigenvectors in our method can better explore the high-order topological information among samples and are less affected by noisy features.

Integration of GCS and FCS. Finally, we average the clean probabilities estimated by the GCS and FCS criteria to identify clean data. As shown in Fig. 1 (a), for a dataset with the noise rate of $r\%$, we divide all samples into a clean set D_{clean} (i.e., $(1 - r\%)$ samples with higher clean probabilities) and a noisy set D_{noisy} (i.e., $r\%$ samples with lower clean probabilities).

2.2 Sample Reliability-Based Mixup (SRMix)

By regarding D_{clean} as a labeled set and D_{noisy} as an unlabeled set, we can formulate the LNL task into an SSL problem and employ Mixup [12] to generate mixed samples for model training [3]. In the traditional Mixup [12], each sample (x_i, y_i) is linearly interpolated with another sample (x_j, y_j) randomly chosen from the mini-batch. However, the pseudo labels of noisy datasets cannot be always reliable and the Mixup operation will aggravate the error accumulation of pseudo labels. To mitigate the error accumulation of pseudo labels, we propose a *Sample Reliability-based Mixup* (SRMix) method, which selects mixup partners based on their spatial reliability, temporal stability, and prediction confidence.

Intuitively, samples with reliable pseudo labels should have consistent predictions with their neighboring samples, stable predictions along sequential training epochs, and high prediction confidence. As shown in Fig. 1 (b), we select reliable mixup partners for each sample based on the triple criteria. First, for each

sample x_j, we define the spatial reliability as:

$$\mathcal{R}_{spatial}(x_j) = 1 - Normalize(||p(x_j) - \frac{1}{K}\sum\nolimits_{x_k} p(x_k)||_2^2), \forall x_k \in KNN\{x_j\} \quad (4)$$

where $p(x_j)$ and $p(x_k)$ are the pseudo labels of sample x_j and its neighbor x_k. $Normalize()$ denotes the min-max normalization over all samples in each batch. If the pseudo label of a sample is more consistent with its neighbors, a higher $\mathcal{R}_{spatial}(x_j)$ will be assigned, and vice versa. Second, for each sample x_j, we keep the historical sequence of its predictions in the past T epochs, e.g., the prediction sequence at the t-th epoch is defined as $P_t(x_j) = [p_{t-T+1}(x_j), ..., p_{t-1}(x_j), p_t(x_j)]$. The temporal stability of x_j can be defined as:

$$\mathcal{R}_{temporal}(x_j) = 1 - Normalize\left(\sqrt{\frac{1}{T}\sum\nolimits_{n=0}^{T-1}(p_{t-n}(x_j) - \bar{p}(x_j))^2}\right), \quad (5)$$

where $\bar{p}(x_j)$ is the average prediction of the historical sequence. According to Eq. (5), a sample with smaller variance or fluctuation over time will be assigned with a larger $\mathcal{R}_{temporal}(x_j)$, and vice versa. Finally, the overall sample reliability is defined as: $\mathcal{R}(x_j) = \mathcal{R}_{spatial}(x_j) \cdot \mathcal{R}_{temporal}(x_j) \cdot max(p(x_j))$, where $max(p(x_j))$ denotes the prediction confidence of the pseudo label of the sample x_j. The possibility of x_j being chosen as a mixup partner is set as

$$p_m(x_j) = \begin{cases} \mathcal{R}(x_j), & \mathcal{R}(x_j) \geq \tau_R \\ 0, & \text{else}, \end{cases} \quad (6)$$

where τ_R is a predefined threshold to filter out unreliable mixup partners. For each sample x_i, we select a mixup partner x_j with the probability $p_m(x_j)$ defined in Eq. (6), and linearly interpolate their inputs and pseudo labels to generate a mixed sample (x, y). The mixed sample is fed into the network and trained with cross-entropy loss $\mathcal{L}_{CE}(x, y)$. Considering the estimation of sample reliability might be inaccurate in the initial training stage, we employ the traditional Mixup in the first 5 epochs and utilize our proposed SRMix for the rest training epochs. Compared with the traditional Mixup, the SRMix can effectively mitigate error accumulation of the pseudo labels and promote model training.

3 Experiments

3.1 Datasets and Implementation Details

WCE Dataset with Synthetic Label Noise. The Wireless Capsule Endoscopy (WCE) dataset [16] contains 1,812 images, including 600 normal images, 605 vascular lesions, and 607 inflammatory frames. We perform 5-fold cross-validation to evaluate our method. Following the common practice in the LNL community [3,4,10], we employ symmetric and pairflip label noise with diverse settings on the training set to simulate errors in the annotation process. The symmetric noise rate is set as $20\%, 40\%, 50\%$, and the pairflip noise rate is set as 40%. The model performance is measured by the average Accuracy (ACC) and Area Under the Curve (AUC) on the 5-fold test data.

Histopathology Dataset with Real-World Label Noise. The histopathology image dataset is collected from Chaoyang Hospital [10] and is annotated by 3 professional pathologists. There are 1,816 normal, 1,163 serrated, 2,244 adenocarcinoma, and 937 adenoma samples of colon slides in total. The samples with the consensus of 3 pathologists are selected as the test set, including 705 normal, 321 serrated, 840 adenocarcinoma, and 273 adenoma samples. The rest samples are utilized to construct the training set, with randomly selected opinions from one of the three doctors used as the noisy labels. The model performance is measured by the average Accuracy (ACC), F1 Score (F1), Precision, and Recall on 3 independent runs.

Table 1. Comparison with state-of-the-art LNL methods on the WCE dataset under diverse noise settings. Best and second-best results are **highlighted** and underlined.

Method	20% Sym.		40% Sym.	
	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
CE (Standard)	86.98 ± 1.32	96.36 ± 0.77	65.84 ± 1.31	83.03 ± 0.98
Co-teaching (NeurIPS 2018) [4]	91.97 ± 1.48	98.45 ± 0.48	84.27 ± 2.51	94.72 ± 0.93
Coteaching+ (ICML2019) [14]	91.86 ± 0.47	98.16 ± 0.26	73.92 ± 1.37	88.00 ± 0.89
DivideMix (ICLR 2020) [3]	93.98 ± 2.27	98.52 ± 0.73	89.79 ± 1.57	96.79 ± 0.92
EHN-NSHE (TMI 2022) [10]	92.71 ± 0.87	98.56 ± 0.35	82.40 ± 2.12	94.22 ± 1.17
SFT (ECCV 2022) [19]	93.32 ± 1.09	**98.66** ± 0.50	84.33 ± 4.64	94.59 ± 1.61
TSCSI (ECCV 2022) [7]	90.07 ± 2.21	97.63 ± 1.02	85.65 ± 4.94	95.61 ± 2.18
Our method	**94.65** ± 2.08	98.56 ± 0.69	**92.44** ± 2.38	**97.70** ± 0.76
Method	50% Sym.		40% Pairflip	
	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
CE (Standard)	52.59 ± 0.76	70.36 ± 0.44	62.36 ± 2.84	81.05 ± 1.49
Co-teaching (NeurIPS 2018) [4]	77.51 ± 2.23	92.14 ± 1.25	82.01 ± 1.40	93.99 ± 0.71
Coteaching+ (ICML2019) [14]	58.00 ± 1.33	75.23 ± 1.24	61.26 ± 1.54	79.12 ± 0.99
DivideMix (ICLR 2020) [3]	82.40 ± 1.93	92.67 ± 1.05	88.30 ± 3.33	96.28 ± 1.60
EHN-NSHE (TMI 2022) [10]	70.14 ± 3.41	85.01 ± 2.70	80.74 ± 4.84	92.15 ± 2.63
SFT (ECCV 2022) [19]	66.94 ± 2.54	83.55 ± 1.94	80.52 ± 6.55	90.65 ± 4.32
TSCSI (ECCV 2022) [7]	76.50 ± 5.04	79.75 ± 6.49	75.28 ± 2.29	96.37 ± 3.03
Our method	**86.59** ± 2.26	**95.25** ± 1.38	**92.38** ± 2.74	**97.80** ± 1.03

Implementation Details. Our method follows the baseline framework of DivideMix [3] and adopts the pre-trained ResNet-50 [17] for feature extraction. We implement our method and all comparison methods on NVIDIA RTX 2080ti GPU using PyTorch [18]. For the WCE dataset, our method is trained for 40 epochs with an initial learning rate set to 0.0001 and divided by 10 after 20 epochs. For the histopathology dataset, the network is trained for 20 epochs with the learning rate set to 0.0001. For both datasets, the network is trained by Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and batch size of 16. The

number of neighbors K is set as 10 in Eq. (1) and Eq. (4). Length T of the historical sequence is set as 3. The reliability threshold τ_R is set as 0.2 for the WCE dataset and 0.05 for the histopathology dataset.

Table 2. Comparison with state-of-the-art methods on the histopathology dataset with real-world label noise. Best and second-best results are **highlighted** and underlined.

Method	ACC (%)	F1 (%)	Precision (%)	Recall (%)
CE (Standard)	80.36 ± 1.29	73.00 ± 0.84	76.47 ± 3.32	72.13 ± 0.29
Co-teaching (NeurIPS 2018) [4]	80.57 ± 0.55	72.39 ± 1.05	76.58 ± 1.40	71.33 ± 1.76
Coteaching+ (ICML2019) [14]	82.15 ± 0.34	74.63 ± 0.30	77.04 ± 0.56	73.94 ± 0.37
DivideMix (ICLR 2020) [3]	82.89 ± 0.80	77.36 ± 1.08	78.31 ± 1.74	76.77 ± 0.76
EHN-NSHE (TMI 2022) [10]	83.06 ± 0.28	76.68 ± 0.39	78.53 ± 0.41	75.00 ± 0.42
SFT (ECCV 2022) [19]	82.68 ± 0.97	76.65 ± 0.89	78.53 ± 1.45	75.64 ± 0.73
TSCSI (ECCV 2022) [7]	82.03 ± 2.12	75.54 ± 2.05	78.51 ± 3.60	74.57 ± 1.78
Our method	$\mathbf{84.29} \pm 0.70$	$\mathbf{78.98} \pm 0.81$	$\mathbf{80.34} \pm 0.85$	$\mathbf{78.19} \pm 0.99$

Table 3. Ablation study on the WCE dataset under 40% symmetric and pairflip noises.

	GCS	FCS	SRMix	40% Sym.		40% Pairflip	
				ACC (%)	AUC (%)	ACC (%)	AUC (%)
1				89.79 ± 1.57	96.79 ± 0.92	88.30 ± 3.33	96.28 ± 1.60
2	✓			91.17 ± 1.36	97.43 ± 0.78	91.12 ± 2.99	97.36 ± 1.13
3		✓		90.56 ± 2.03	96.89 ± 1.03	91.01 ± 2.23	97.22 ± 1.06
4			✓	90.84 ± 1.66	96.74 ± 1.25	89.40 ± 2.70	96.27 ± 1.63
5	✓	✓		92.11 ± 2.31	97.44 ± 0.92	91.72 ± 3.01	97.60 ± 1.06
6	✓	✓	✓	92.44 ± 2.38	97.70 ± 0.76	92.38 ± 2.74	97.80 ± 1.03

3.2 Experimental Results

Comparison with State-of-the-Art Methods. We first evaluate our method on the WCE dataset under diverse synthetic noise settings and show the results in Table 1. We compare with three well-known LNL methods (i.e., Co-teaching [4], Coteaching+ [14], and DivideMix [3]) and three state-of-the-art LNL methods (i.e., EHN-NSHE [10], SFT [19], and TSCSI [7]). As shown in Table 1, our method outperforms existing methods under all noise settings, and the performance gain is more significant under severe noise settings (e.g., noise rate $\geq 40\%$). Under the four settings, our method outperforms the second-best model by 0.67%, 2.65%, 4.19%, and 4.08% in accuracy. These results indicate the effectiveness of our method.

We then evaluate our method on the histopathology dataset with real-world label noise. As shown in Table 2, our method outperforms existing state-of-the-art methods, indicating the capability of our method in dealing with complex real-world label noise.

Ablation Study. To quantitatively analyze the contribution of the proposed components (i.e., GCS, FCS, and SRMix) in combating label noise, we perform an ablation study on the WCE dataset under 40% symmetric and pairflip noise. As shown in Table 3, compared with the DivideMix baseline (line 1) [3], replacing the small-loss criterion by GCS or FCS both improve the model performance significantly (lines 2–3), and their combination leads to further performance gains (line 5). Furthermore, better performance can be achieved by replacing the traditional Mixup with our proposed SRMix method (line 1 *vs.* line 4, line 5 *vs.* line 6). These results indicate that filtering out unreliable mixup partners can effectively improve the model's capacity in combating label noise.

More comprehensive analysis of the GCS and FCS criteria is provided in the supplementary material. Figure S1 demonstrates that compared with the normalized loss [3], the GCS and FCS criteria are more distinguishable between the clean and noisy data. This is consistent with the improvement of clean data selection accuracy in Fig. S2. As shown in Fig. S3, both the feature and gradient of each sample are aligned with the center of its true class, further validating the rationality of using gradient and feature conformity for clean data selection.

4 Conclusion

In this paper, we present a two-stage framework to combat label noise in medical image classification tasks. In the first stage, we propose two novel criteria (i.e., GCS and FCS) that select clean data based on their conformity with the class-wise principal gradients and feature eigenvectors. By exploring contextual information and high-order topological correlations in the gradient space and feature space, our GCS and FCS criteria enable more accurate clean data selection and benefit LNL tasks. In the second stage, to mitigate the error accumulation of pseudo labels, we propose an SRMix method that interpolates input samples with reliable mixup partners which are selected based on their spatial reliability, temporal stability, and prediction confidence. Extensive experiments on two datasets with both diverse synthetic and real-world label noise indicate the effectiveness of our method.

Acknowledgements. This work was supported by National Natural Science Foundation of China 62001410 and Innovation and Technology Commission-Innovation and Technology Fund ITS/100/20.

References

1. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Commun. ACM **64**(3), 107–115 (2021)
2. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. NeurIPS **33**, 20331–20342 (2020)
3. Li, J., Socher, R., Hoi, S.C.H.: DivideMix: learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020)
4. Han, B., et al.: Co-teaching: robust training of deep neural networks with extremely noisy labels. In: NeurIPS, vol. 31 (2018)
5. Bai, Y., Liu, T.: Me-momentum: extracting hard confident examples from noisily labeled data. In: ICCV, pp. 9312–9321 (2021)
6. Li, J., Li, G., Liu, F., Yu, Y.: Neighborhood collective estimation for noisy label identification and correction. arXiv preprint arXiv:2208.03207 (2022)
7. Zhao, G., Li, G., Qin, Y., Liu, F., Yu, Y.: Centrality and consistency: two-stage clean samples identification for learning with instance-dependent noisy labels. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022. ECCV 2022. LNCS, vol. 13685, pp 21–37. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19806-9_2
8. Tang, H., Jia, K.: Towards discovering the effectiveness of moderately confident samples for semi-supervised learning. In: CVPR, pp. 14658–14667 (2022)
9. Iscen, A., Valmadre, J., Arnab, A., Schmid, C.: Learning with neighbor consistency for noisy labels. In: CVPR, pp. 4672–4681 (2022)
10. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard sample aware noise robust learning for histopathology image classification. IEEE Trans. Med. Imaging **41**(4), 881–894 (2021)
11. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.:. MixMatch: a holistic approach to semi-supervised learning. In: NeurIPS, vol. 32 (2019)
12. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
13. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: ICML, pp. 2304–2313. PMLR (2018)
14. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: ICML, pp. 7164–7173. PMLR (2019)
15. Arpit, D., et al.: A closer look at memorization in deep networks. In: ICML, pp. 233–242. PMLR (2017)
16. Dray, X., et al.: Cad-cap: UNE base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle. Endoscopy **50**(03), 000441 (2018)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
18. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: NeurIPS, vol. 32 (2019)
19. Wei, Q., Sun, H., Lu, X., Yin, Y.: Self-filtering: a noise-aware sample selection for label noise with confidence penalization. arXiv preprint arXiv:2208.11351 (2022)