Knowledge-Based Systems 276 (2023) 110732

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

SSS-Net: A shadowed-sets-based semi-supervised sample selection network for classification on noise labeled images

Kecan Cai [a,b], Hongyun Zhang [a,b,*], Witold Pedrycz [c,d,e], Duoqian Miao [a,b]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China
[b] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, PR China
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada
[d] Systems Research Institute, Polish Academy of Sciences, 00-901 Warsaw, Poland
[e] Istinye University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Sariyer/Istanbul, Turkiye

ARTICLE INFO

Article history:
Received 1 December 2022
Received in revised form 13 June 2023
Accepted 14 June 2023
Available online 22 June 2023

Keywords:
Shadowed sets
Adaptive division threshold
Semi-supervised learning
Images classification
Noisy label

ABSTRACT

Sample selection is a fundamental technique utilized in image classification with noisy labels. A plethora of sample selection approaches published in the literature are based on a small-loss strategy, in which division thresholds are set manually and the correlation between sample losses is ignored. Furthermore, one of the most evident shortcomings of these approaches is that noisy samples with low-quality pseudo-labels can negatively impact the model resulting in poor performance. In this study, a shadowed-sets-based semi-supervised sample selection network called SSS-Net is developed to address these limitations. Our approach leverages a novel technique that combines a loss-similarity-based-clustering method (LSCM) with the shadowed-sets theory to adaptively select clean samples. We then introduce an original high-quality pseudo-label sample reselection (HPSR) strategy, which is designed through the co-training of two networks, to pick the samples with high-quality pseudo-labels. Finally, the selected samples are utilized to further train the network and complete classification. This study presents an automated approach that determines optimal division thresholds to select clean samples adaptively. Furthermore, it improves the current semi-supervised sample selection method by effectively utilizing noisy samples. The suitability and promising performance of the proposed approach are supported through experimental studies using five real-world datasets. Comparative studies involving several state-of-the-art methods are also reported.

1. Introduction

The impressive success of machine learning is largely attributed to the massive amount of labeled data [1–3]. However, it is extremely expensive and time-consuming to obtain high-quality annotations. Many affordable techniques, such as web image crawling or machine labeling [4,5], have been widely used to generate labeled data. Regrettably, these techniques are prone to generate unreliable labels, commonly known as noisy labels [6]. Various methods have been proposed to mitigate the negative impact of noisy labels, which can be classified into three types [1,7]: robust architecture [8–11], robust regularization using loss functions [12,13], and sample selection strategies [14–19].

This study focuses on sample selection strategies, which aim to select samples with reliable labels as training datasets to develop a classification network. It has been observed that deep neural networks tend to learn simple patterns before fitting label noise [20]. Therefore, many sample selection methods adopt a small-loss strategy. They "warm-up" the network for a few epochs by training on all data to obtain losses, after which samples with loss values below a certain threshold are classified as clean (with mostly reliable labels), and the remaining samples are noisy ones (with mostly unreliable labels), as illustrated in Fig. 1(a). This study categorizes small-loss-based sample selection methods into two types: supervised learning (SL) methods [17–19] and semi-supervised learning (SSL) methods [14–16,21–25], as shown in Fig. 1(b) and (c).

The SL methods select the clean samples according to the small-loss strategy and fine-tune the network with the clean samples, which may lead to waste of noisy samples. In contrast, SSL methods generate pseudo-labels on the noisy samples and the network can be obtained based on both the clean samples with reliable labels and the noisy samples with pseudo-labels.

* Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China.
E-mail addresses: caikecan@tongji.edu.cn (K. Cai), zhanghongyun@tongji.edu.cn (H. Zhang), wpedrycz@ualberta.ca (W. Pedrycz), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.knosys.2023.110732
0950-7051/

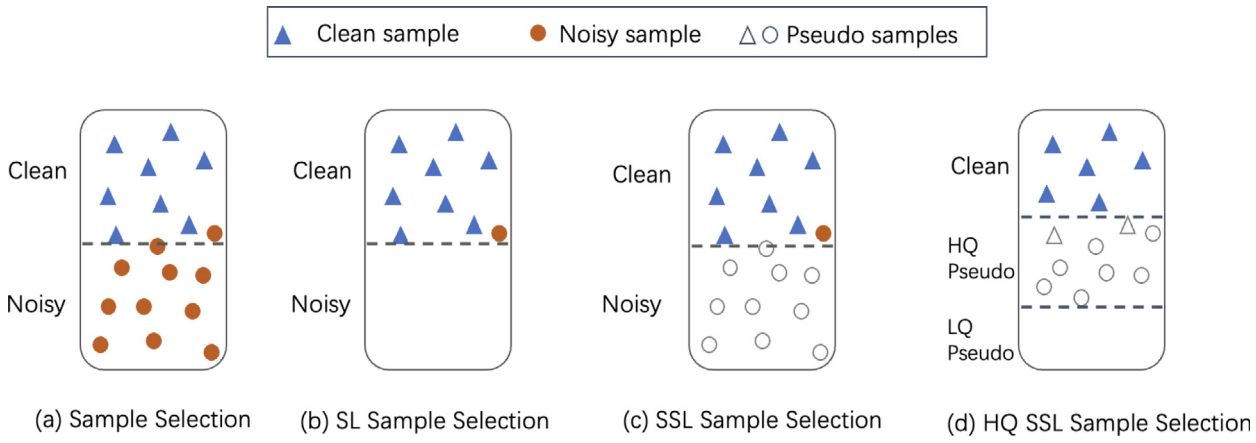

Fig. 1. Sample selection based on the small-loss strategies.

Apparently, the latter outperforms the former by fully utilizing pseudo-labeled samples. However, in existing sample selection methods, the thresholds for selecting clean samples are determined manually based on the sample losses without considering the correlation between sample losses. Furthermore, utilizing noisy samples with low-quality (LQ) pseudo-labels in SL may negatively impact the model training.

To address these challenges, a shadowed-sets-based semi-supervised sample selection network, namely SSS-Net, is formulated in this study. The frame diagram of the model is shown in Fig. 2. First, to perform the operation of adaptive selection of clean samples, the novel technique combined with the loss-similarity-based-clustering method (LSCM) and the shadowed-sets theory is developed. To be more specific, after conducting a warm-up operation and obtaining the sample losses, the LSCM algorithm is designed to preliminarily simulate the distribution of clean and noisy samples and calculate the membership degrees of samples belonging to clean samples. A technique to automatically calculate partition thresholds is proposed based on the shadowed-sets theory and the membership degrees. The thresholds are used to separate clean and noisy samples. As a result, samples can be classified into two types: clean samples and noisy samples with uncertain labels that require further evaluation of their reliability. Then, in order to fully utilize the noisy samples obtained in the previous step while avoiding the negative impact of LQ pseudo-labeled noisy samples, we design a high-quality pseudo-label sample reselection (HPSR) algorithm. This algorithm selects samples with high-quality (HQ) pseudo-labels (as shown in Fig. 1(d)), treating them along with the clean samples as the training set to build the network. It is worth noting that both the clean samples and the HQ pseudo-label samples determined by one network are regarded as training data to build the other network. Furthermore, two networks are trained simultaneously to avoid errors that may occur during single network training. Overall, the main contributions in this study are outlined as follows:

- To better select clean samples, the LSCM algorithm was designed to simulate the distribution of clean and noisy samples, and calculate the membership degree of samples belonging to clean samples based on the distribution.
- To address the issue of manually determined division thresholds in existing sample selection methods, an adaptive division technique is developed based on the shadowed-sets theory. To the best of our knowledge, this is the first time that the shadowed-sets theory has been utilized to improve the classification on noise labeled images.
- To better utilize noisy samples and avoid the negative impact of LQ pseudo-labels, the HPSR algorithm is designed to reselect the HQ pseudo-labeled samples for model training. This not only reduces the number of training samples but also enhances the network's performance.

In summary, this study aims to create a deep convolutional neural network for accurately classifying noise-labeled images. To achieve this, a novel clustering technique based on loss similarity is developed to simulate the distribution of clean and noisy samples for the purpose of identify clean samples. The study also incorporates the shadowed-sets theory to adaptively calculate the division threshold for sample selection. Additionally, the study introduces an original strategy for reselecting high-quality pseudo-labeled samples that efficiently utilizes noisy samples. The study is organized as follows: In Section 2, we briefly review the techniques reported in the literature. In Section 3, we introduce and explain a shadowed-sets-based semi-supervised sample selection network called SSS-Net in detail. In Section 4, we reported a number of experiments. Finally, concluding comments are reached in Section 5.

2. Literature review

Sample selection is a commonly used strategy to handle training datasets that contain noisy labeled samples. One approach to sample selection involves training two or more networks simultaneously and selecting samples based on disagreements among these networks. However, this can lead to higher computational costs [1,20]. Another commonly employed heuristic in sample selection is to identify clean samples based on lower loss values, as compared to noisy samples. This approach is utilized in several recent methods and can be classified as either SL or SSL methods.

Several early studies reported in the literature focused on SL methods, with the network updated only based on clean samples. MentorNet [19] serves as a representative SL approach by pre-training an additional network to select clean samples and guide the training process. To prevent confirmation bias due to the selection of a single network, multi-networks are frequently utilized for sample selection training. LongReMix [26] introduced a two-stage learning process. The first stage identifies a small but precise set of clean labels, and the second stage augments this set with new clean samples to improve the robustness of the model. The study in [18] presented a decoupled method to build two networks simultaneously, and then selected samples based on a disagreement between these two networks. Multi-round learning is also used as an effective method in [1] to iteratively refine the selected set of clean samples by repeating the training round. In a nutshell, these studies highlight the importance of clean samples, while little attention has been paid to the noisy samples.

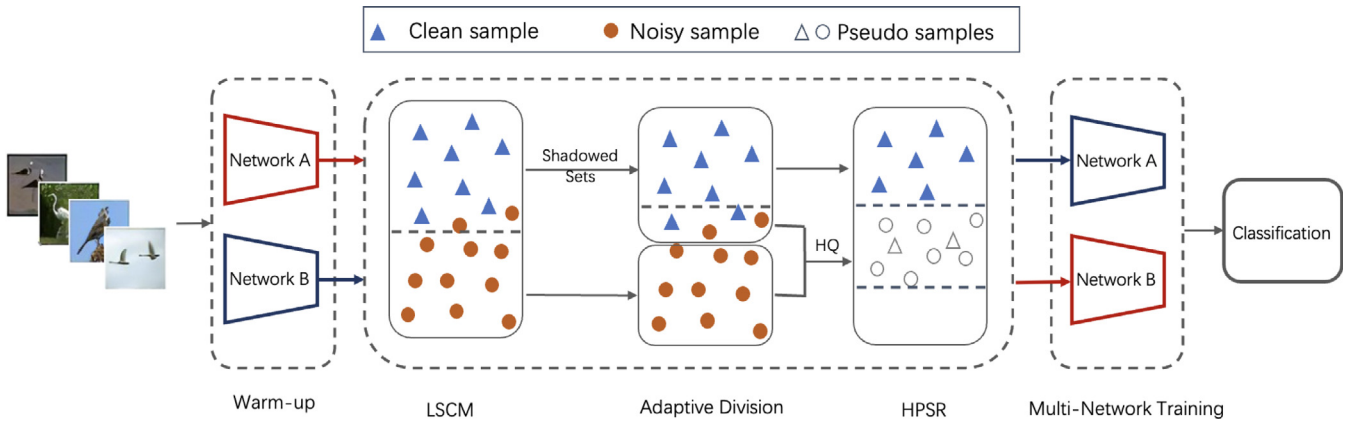

Fig. 2. A general overview of the SSS-Net.

Currently, sample selection strategies are combined with SSL methods to enhance the utilization of noisy samples, which has been demonstrated to be more effective than the use of SL methods [21]. For instance, [27] presents a SSL framework to make better use of pseudo-negative labels data with low prediction confidence, by decreasing the probability of pseudo-negative labels. In [22], the SELFIF method is introduced as a way to refurbish the labels of all noisy samples to increase the number of available training samples. However, using the entire set lacks a mechanism to filter out erroneous labels. To address this, the RoCL method, presented in [23], employs a more precise detection strategy for both clean and noisy samples. It corrects noisy samples with pseudo-labels and uses two-phase learning strategies: supervised training on clean samples and self-supervision on relabeled noisy samples. The study completed in [15] presents the DivideMix method, in which clean samples are treated as labeled data, and the remaining noisy samples are treated as unlabeled. Both transformed labeled and unlabeled data are used to train the SSL model. In [24], the AUGDESC method is formulated by examining multiple augmentation strategies and incorporating them into the DivideMix method to improve classification results on synthetic and real-world noisy label datasets. In [25], the authors explore semantic clustering and the SSL method called Scanmix to deal with noisy labeled samples. This approach uses self-supervised training to pre-train DivideMix, resulting in superior robustness to severe label noise and competitive robustness to non-severe label noise problems.

As mentioned above, it is worth noting that the SSL methods can leverage noisy samples to achieve superior results over SL methods. However, there is still a significant exploratory space for analyzing and utilizing noisy samples effectively. Furthermore, it is evident that the process of separating original noisy labeled datasets into clean and noisy samples is central to the challenges of sample selection. Motivated by these considerations, this study introduces SSS-Net, which aims to enhance the classification performance based on noisy labeled images.

3. The proposed method

In this section, we introduce SSS-Net, a shadowed-sets-based semi-supervised sample selection network for classifying noise-labeled images. We have outlined the research objective and general process of SSS-Net in the Introduction, and have illustrated it using Fig. 2. Furthermore, this section elaborates on the three main phases in detail, which include the LSCM algorithm, an adaptive division technique, and a HPSR algorithm. The first two approaches aim to identify clean samples, while the last one is employed to re-select HQ pseudo-labeled samples.

3.1. The loss-similarity-based-clustering method

To identify the clean samples, this paper introduces the LSCM algorithm that learns the distribution of both clean and noisy samples. Based on this distribution, the algorithm calculates the membership degree of each sample belonging to clean sample set. The obtained membership degree forms the basis for implementing the subsequent automatic division algorithm.

Given a training dataset $D = (X, Y) = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i and $y_i \in [0, 1]^c$ refer to an image and the one-hot label in c classes, respectively. The cross-entropy losses $l(\theta)$ of the deep neural network with parameter vector θ is expressed as follows [15]:

$$l(\theta) = \{l(x_i)\}_{i=1}^{n} = -\left\{\sum_{c=1}^{c} y_i^c log(p_{model}^c(x_i, \theta))\right\}_{i=1}^{n} \quad (1)$$

where p_{model}^c is the network's output softmax probability for class c. To learn the distribution of clean and noisy samples, we have employed a two-class clustering method that fits the similarity between sample losses. Further details about this method will be discussed in the experimental studies section. In this way, one can obtain two clusters, namely the relatively clean samples and the noisy samples, which are denoted as $CLEAN_{rel}$ and $NOISY_{LSCM}$, respectively.

Since noisy samples may still exist in $CLEAN_{rel}$, it is indispensable to further capture the clean samples from $CLEAN_{rel}$. To quantify this, we consider a fuzzy set representation. The membership degree of $CLEAN_{rel}$ considered in this study is defined over the space $U = \{x_i\}_{i=1}^{n}$,

$$\begin{aligned} &\mu : U \rightarrow [0, 1] \\ &\quad x_i \rightarrow \mu(x_i) \end{aligned} \quad (2)$$

where μ refers to a membership function of U. The membership function is defined in the following form:

$$\mu(x_i) = \begin{cases} 1, & l(x_i) \leq a \\ (b - l(x_i))/(b - a), & a < l(x_i) < b \\ 0, & b \leq l(x_i) \end{cases} \quad (3)$$

where $\mu(x_i) \in [0, 1]$ is the membership degree, indicating the degree of x_i belonging to U, and U refers to $CLEAN_{rel}$ in this study. a and b are two parameters with $l_{min} \leq a < b \leq l_{max}$, where l_{max} and l_{min} refer to the maximum and minimum loss values of $CLEAN_{rel}$. We set $a = l_{min} + 1/2(l_{max} - l_{min})$ and $b = l_{min} + 3/4(l_{max} - l_{min})$. The clean samples can be further selected from $CLEAN_{rel}$ on the basis of the membership degree and a division threshold, which is described in the following subsection.

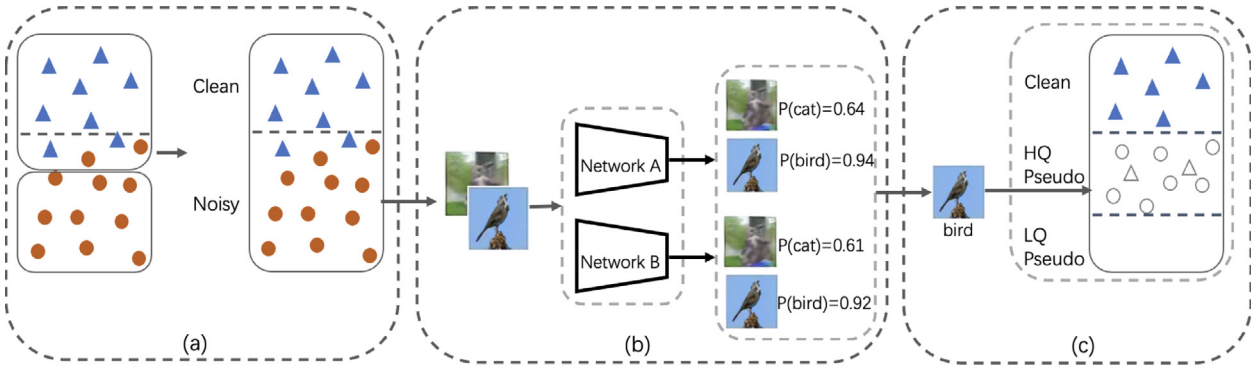

Fig. 3. High-quality pseudo-label sample reselection method.

3.2. The adaptive selection technique

The division thresholds for sample selection are crucial in identifying clean samples. However, in existing methods, the thresholds are usually set manually based on experience or parameter tuning to find an optimal value [28,29]. Therefore, in this study, an adaptive method to systematically determine the division thresholds is developed based on the shadowed-sets theory. We briefly recall ideas of shadowed sets used in the proposed method.

Definition 1 (*Shadowed Sets [30,31]*)**.** A shadowed set S on universe $U = \{x_i\}_{i=1}^n$ is defined as a mapping from U to the set $\{0, [0, 1], 1\}$, viz., $S : U \to \{0, [0, 1], 1\}$, where mapping S is given as follows:

$$S_{\mu(x_i)} = \begin{cases} 1, & \alpha \le \mu(x_i) \le 1 \\ [0, 1], & \beta < \mu(x_i) < \alpha \\ 0, & 0 \le \mu(x_i) \le \beta \end{cases} \tag{4}$$

where α and β are division thresholds with $0 \le \beta < \alpha \le 1$. The optimal values of α and β are obtained by solving the following minimization problem [30,31],

$$\begin{aligned} &\arg\min_{(\alpha,\beta)} V_{(\alpha,\beta)}\left(\mu(x_i)\right) \\ &= \min \Big| \sum_{\alpha \le \mu(x_i) \le 1} (1 - \mu(x_i)) + \sum_{0 \le \mu(x_i) \le \beta} \mu(x_i) \\ &- card(x_i \mid \beta < \mu(x_i) < \alpha) \Big| \end{aligned} \tag{5}$$

where $card(\cdot)$ denotes the cardinality. For the sake of computational complexity, it is assumed that the $\alpha + \beta = 1$ [30]. In this way, the division thresholds (α, β) can transform the data U into three disjoint regions, viz., elevated region 1, shadow region $[0, 1]$ and reduced region 0, and it has a one-to-one correspondence in Three-way decisions including positive region (POS), boundary region (BND) and negative region (NEG).

Definition 2 (*Three-way Decisions [31,32]*)**.** Suppose $U = \{x_i\}_{i=1}^n$ is a nonempty finite set of elements, and a nonempty finite set of conditions is given. U can be partitioned into three pairwise disjoint regions using a mapping function f.

$$f : U \to \{POS, BND, NEG\} \tag{6}$$

where POS, BND and NEG denote the regions of acceptance, non-commitment and rejection, respectively. Compared to the conventional two-way decisions of accepting or rejecting, the three-way decisions involving a "no-commitment" option requires further analysis to determine whether it should be accepted or rejected. Specifically, POS, BND, and NEG can be further expressed based on the division thresholds α and β obtained through shadowed-sets theory, as shown below:

$$\begin{aligned} POS &= \{x_i \in U \mid \alpha \le \mu(x_i) \le 1\} \\ BND &= \{x_i \in U \mid \beta < \mu(x_i) < \alpha\} \\ NEG &= \{x_i \in U \mid 0 \le \mu(x_i) \le \beta\} \end{aligned} \tag{7}$$

Similarly, in the scenario of noise labeled sample selection, the division thresholds can be calculated using the membership degree on the basis of the shadowed-sets theory, which separates the samples into three disjoint regions, namely clean regions, boundary regions, and noisy regions. These three regions denote clean samples with reliable labels, uncertain samples with uncertainty labels and noisy samples with unreliable labels, respectively, which are represented below:

$$\begin{aligned} CLEAN &= \{x_i \in U \mid \alpha \le \mu(x_i) \le 1\} \\ BND &= \{x_i \in U \mid \beta < \mu(x_i) < \alpha\} \\ NOISY &= \{x_i \in U \mid 0 \le \mu(x_i) \le \beta\} \end{aligned} \tag{8}$$

Due to the uncertainty of the samples located in the boundary regions, we consider them to be potentially unreliable. Therefore, we classify all samples into two types: clean samples with reliable labels, and noisy samples with potentially unreliable labels. This classification is presented in the following format:

$$\begin{aligned} CLEAN &= \{x_i \in U \mid \alpha \le \mu(x_i) \le 1\} \\ NOISY_{adaptive} &= \{x_i \in U \mid 0 \le \mu(x_i) < \alpha\} \end{aligned} \tag{9}$$

based on the description above, which includes the LSCM algorithm and the adaptive division technique, the total noisy samples are composed with two parts, as shown below:

$$NOISY_{total} = NOISY_{LSCM} + NOISY_{adaptive} \tag{10}$$

3.3. High-quality pseudo-label sample reselection strategy

In existing SSL methods, all clean and noisy samples with pseudo-labels are considered as training sets. However, it is important to note that noisy samples with LQ pseudo-labels can have a negative effect on the training model. To overcome this, a HPSR algorithm is proposed to utilize the noisy samples more effectively while avoiding the negative impact of LQ noisy samples. The algorithm comprises two main parts: generating pseudo-labels for all noisy samples as shown in Fig. 3(b) and selecting only the HQ pseudo-label samples in Fig. 3(c).

In the process of generating pseudo-labels, in order to avoid errors that may occur during a single network training, we employed two networks using a co-training approach to generate pseudo-labels. Specifically, for each sample x_i, we obtained the

output softmax probability P over C classes in the following manner,

$$\begin{aligned} P_i^1(c_m) &= net\ (x_i, \theta^1) \\ P_i^2(c_m) &= net\ (x_i, \theta^2) \end{aligned} \tag{11}$$

where $net(\theta^1)$ is a deep conventional neural network with parameters vector θ^1, and $net(\theta^2)$ is another deep neural network with parameters vector θ^2. $P_i^1(c_m)$ is the softmax probability that sample x_i belongs to class $c_m \in C$ using $net(\theta^1)$. $P_i^2(c_m)$ is the softmax probability that sample x_i belongs to class c_m using $net(\theta^2)$. Finally, the prediction probability of sample x_i belongs to class c_m can be calculated by averaging the outputs of these two networks,

$$Pred_i(c_m) = mean\ (P_i^1(c_m), P_i^2(c_m)) \tag{12}$$

in other words, c_m is the pseudo-label of sample x_i. Similarly, for another sample $x_j (x_i \neq x_j)$, the predicted probability of sample x_j belonging to a certain category $c_n \in C$ can be calculated as follows,

$$\begin{aligned} P_j^1(c_n) &= net\ (x_j, \theta^1) \\ P_j^2(c_n) &= net\ (x_j, \theta^2) \\ Pred_j(c_n) &= mean\ (P_j^1(c_n), P_j^2(c_n)) \end{aligned} \tag{13}$$

where $P_j^1(c_n)$ is the output softmax probability that sample x_j belongs to class c_n using $net(\theta^1)$. $P_j^2(c_n)$ is the softmax probability that sample x_j belongs to class c_n using $net(\theta^2)$. $Pred_j(c_n)$ is the prediction probability calculated by averaging the outputs of $net(\theta^1)$ and $net(\theta^2)$, namely, c_n is the pseudo-label of sample x_j.

Usually, we take the maximum value of the softmax layer as the predicted probability, and the corresponding class with the highest probability is considered as the predicted label, which is also known as the pseudo-label. The higher the probability is, the more reliable the pseudo-label becomes, indicating its higher quality. Therefore, we compare the predicted probabilities of samples x_i and x_j, and select the sample with higher prediction probability as HQ sample for the following model training. The specific process is as follows:

$$HQ = \begin{cases} x_i, & Pred_j(c_n) < Pred_i(c_m) \\ x_j, & Pred_i(c_m) \leq Pred_j(c_n) \end{cases} \tag{14}$$

where $Pred_i(c_m)$ and $Pred_j(c_n)$ are the prediction probabilities of samples x_i and x_j.

3.4. Loss calculation

The loss of the SSS-Net is defined as follows:

$$\begin{aligned} L_{CL} &= -\frac{1}{|CL|} \sum_{x,y \in CL} \sum_{c \in C} y^c log(p_{model}^c(x, \theta)) \\ L_{HQ} &= -\frac{1}{|HQ|} \sum_{x,y \in HQ} \|y - p_{model}(x, \theta)\|_2^2 \end{aligned} \tag{15}$$

where the loss on clean samples L_{CL} is the cross-entropy loss and the loss on HQ pseudo-label samples L_{HQ} is the mean squared. x is an image and y is the one-hot label over C classes. p_{model}^c is the network's output softmax probability for class c. According to the previous work in [33], the following regularization term is utilized to regularize the model's average output:

$$L_{reg} = \frac{1}{C} \sum_{c \in C} log \left(\frac{1}{C} \Big/ \frac{1}{|CL| + |HQ|} \sum_{x \in (CL+HQ)} (x, \theta) \right) \tag{16}$$

where c is the label over C classes, the final loss is given as follows:

$$L = L_{CL} + \lambda_u L_{HQ} + \lambda_r L_{reg} \tag{17}$$

where the parameters λ_u and λ_r are the same as in the previous work [15] to regularize the loss. The SSS-Net model for classification on noise labeled images is implemented in the form of Algorithm 1.

4. Experimental studies

4.1. Datasets and experimental setup

Various datasets were utilized to support and quantify performance of the developed SSS-Net. These datasets included synthetic datasets, namely CIFAR-10 and CIFAR-100, as well as real-world datasets such as Clothing1M, WebVision, and ILSVRC2012. For the synthetic datasets, we generated noisy labels by randomly replacing a percentage of the training data with all possible labels. In this study, we experimented with 20%, 50%, and 80% label noise levels. More information about the generation mechanism of noisy labels can be found in [15]. Regarding the real-world datasets, these datasets already contain a certain fraction of label noise, thus, we did not introduce any additional noise. Below, we provide further details on these datasets.

CIFAR-10 and CIFAR-100: The CIFAR-10 and CIFAR-100 datasets have 10 categories of images and 100 categories, respectively, where each contains 60k color images of size 32×32 for classification, and 50k for training and 10k for testing.

Clothing1M: The Clothing1M is a large-scale real-world dataset with 14 categories of noisy labels, involving 1 million training images captured from online shopping websites.

WebVision and ILSVRC2012: Webvision is another real-world dataset, with over 2.4 million images captured from the web in ImageNet ILSVRC2012. Following previous work [15], the SSS-Net is studied on the mini-WebVision 1.0 dataset and ImageNet ILSVRC2012 validation set, where the mini-WebVision has the first 50 classes of the Google image subset.

Experimental Setup of Backbone Networks. We follow the co-training strategy in [34] and use MixMatch [15] as the semi-supervised method. For the CIFAR-10 and CIFAR-100 dataset, we follow the reported work in [34,35] using a 34-layer ResNet as backbone, which can be implemented using the SGD (*momentum* $= 0.9$) with an initial learning rate of 0.02 (reduced by 0.002 after 150 epochs), weight decay of 0.0005 and a batch size of 128. The number of epochs of the warm-up periods were set as 10 and 30 for the CIFAR-10 and CIFAR-100, respectively, and the model was trained for 300 epochs. As for the Clothing1M dataset, the ResNet-50 with ImageNet pre-trained weights is employed with an initial learning rate of 0.002 (reduced by 0.0002 after 40 epochs). The Inception-ResNet is utilized on the mini-WebVision and ILSVRC2012 datasets with an initial learning rate of 0.01 (reduced by 0.001 after 50 epochs). The network is trained using SGD (*momentum* $= 0.9$), a weight decay of 0.001 and a batch size of 32. The warm-up period is 1 epoch and the network is trained for 100 epochs. The above parameters are set on the basis of Refs. [15,34].

4.2. Results of combining different clustering methods

In light of the previous work in [15], we have combined an existing no-clustering-based method with various clustering techniques, such as K-means, MiniBatchKMeans, and Hierarchical, as outlined in Table 1. Our results demonstrate that the hierarchical clustering incorporated model outperforms other incorporated models in terms of both the best test accuracy and

Algorithm 1 The SSS-Net model.

Input: train dataset (X, Y), test dataset (X', Y'), randomly initialization parameters θ_1, θ_2.
Output: test accuracy.

$\theta_1, \theta_2 = Warm-up\ (X, Y, \theta_1, \theta_2)$ // standard training for a specific number of epochs
$loss_1 = Net\ (X, Y, \theta_1)$
$loss_2 = Net\ (X, Y, \theta_2)$
while $i < MaxEpoch$ **do**
  $cluster = argmin_{c_1,c_2} = LSCM(loss_1)$ // c_1 and c_2 refer to the clustering results.
  $\mu_A = T(cluster)$ // T is the membership function.
  $\alpha, \beta = shadowedsets\ (\mu_A)$
  $CL = AdaptiveSelection\ (X, \alpha, \beta)$ // *CL* refers to *CLEAN*.
  $HQ = HPSR\ (NOISY_{total})$
  for *iter* $= 1$ **to** *num_iters* **do**
    from *CL* draw a mini-batch $\{x_{cl}, y\}$
    from *HQ* draw a mini-batch $\{x_{hq}\}$
    $L = L_{cl} + \lambda_u L_{hq} + \lambda_r L_{reg}$
    $\theta_1 = SGD(L, \theta_1)$
  end for
  $loss_2 = Net\ (X, Y, \theta_1)$ // model per-sample $loss_2$ with θ_1
  use $loss_2$ and steps 5-15 to obtain θ_2
  $losses_1 = Net\ (X, Y, \theta_2)$ // model per-sample $loss_1$ with θ_2
  $accuracy = test\ (X', Y', \theta_1, \theta_2)$
end while
return test accuracy.

Table 1
Clustering analysis on CIFAR-10 and CIFAR-100.

Dataset		CIFAR-10			CIFAR-100		
Clustering/Noise ratio		20%	50%	80%	20%	50%	80%
Backbone w/No-Clustering	Best	96.00	90.28	77.82	78.23	67.95	50.96
	Avg	93.49	83.76	56.27	76.26	61.33	30.56
Backbone w/K-Means	Best	96.14	94.43	**91.52**	78.72	**77.48**	58.23
	Avg	95.61	94.19	90.92	77.75	**77.30**	58.18
Backbone w/MiniBatchKMeans	Best	96.66	95.26	88.60	79.89	76.39	55.91
	Avg	96.28	95.05	86.01	79.28	75.97	55.22
Backbone w/Hierarchical	Best	**96.75**	**95.51**	91.44	**80.67**	77.42	**59.89**
	Avg	**96.53**	**95.24**	**91.25**	**80.12**	77.28	**59.41**

the average accuracy of the last 20 epochs for noise ratios of 0.2, 0.5, and 0.8. These findings suggest that hierarchical clustering is a promising approach for improving model performance.

Fig. 4 shows the test accuracy achieved by different clustering methods during 300 epochs of the training cycle on the CIFAR-10 and CIFAR-100. It can be observed that all methods exhibit similar performance during the network's warm-up phase, and the accuracy significantly increases after 150 epochs due to the update of model parameters, as described in the experimental setup subsection. Overall, one can conclude that combining clustering methods within the existing model can effectively improve accuracy. It is worth noting that the hierarchical clustering method outperforms the other mentioned methods in terms of test accuracy and stability, across various noise ratios. Therefore, we employed the hierarchical clustering method in the following experimental studies.

4.3. Comparison with state-of-the-art methods

In the compared methods, for the Clothing1M dataset, ResNet-50 is used as backbone network, and for the mini-WebVision and ILSVRC2012 datasets, the backbone framework used is Inception-ResNet. For the CIFAR-10 and CIFAR-100 datasets, Resnet-18, ResNet-32, ResNet-34, and ResNet-101 are used in different state-of-the-art methods. The compared methods mostly focus on improving accuracy, which increases the complexity of the model and requires a significant amount of time spent on parameter tuning to determine the division thresholds manually.

However, the proposed method not only focuses on improving accuracy but also emphasizes adaptability and complexity. According to Tables 2–4, our experimental results obtained on five datasets demonstrate consistent superiority over the compared methods, particularly showing significant improvements on the CIFAR-100 dataset. For example, compared to the latest method, our method achieves a maximum increase of 3.06% in accuracy and an average improvement of 2.75% when the CIFAR-100 dataset has a 20% noise rate. Furthermore, the proposed method enhances the adaptability of the model by automatically obtain the optimal division threshold. Additionally, the complexity of the proposed method is reduced by selecting and retaining only high-quality pseudo-labeled samples for model training (as shown in Table 6). More specifically, compared with existing methods, the proposed SSS-Net offers the following advantages: Firstly, existing methods have not considered the correlation between sample losses. In this study, we designed the LSCM, which can aid in selecting clean samples. Secondly, by utilizing the shadowedsets theory, we can select clean samples adaptively, which avoids the problem of artificially determining the division thresholds in existing methods. Moreover, existing SSL methods use all pseudo-labeled samples without considering that noisy samples with low-quality pseudo-labels can have a negative impact on the model's performance. To address this, we propose the HPSR strategy, which selects samples with high-quality pseudo-labels, and

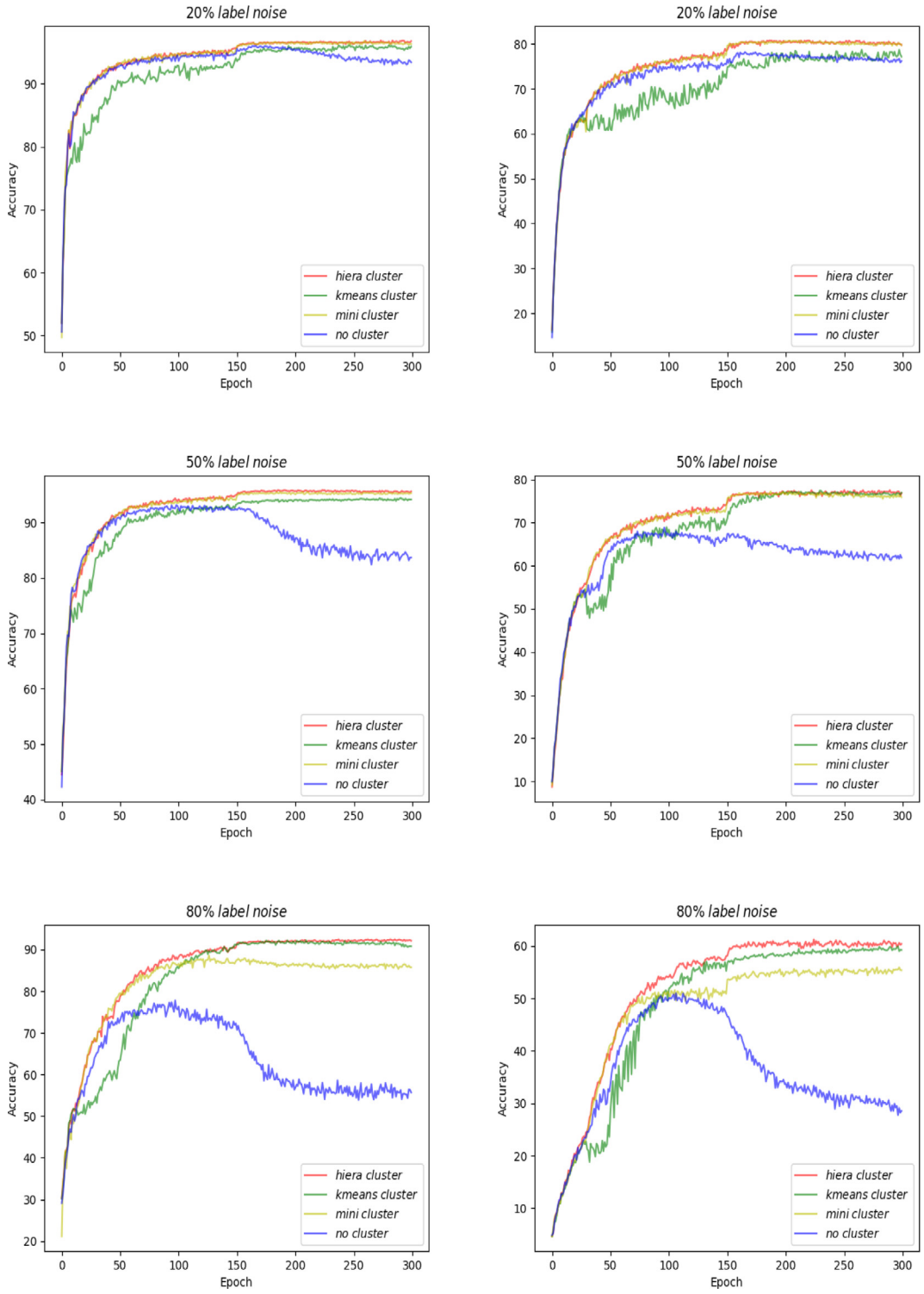

Fig. 4. Comparison with different clustering methods with various label noise on CIFAR-10 (left) and CIFAR-100 (right).

discarding low-quality pseudo-labeled samples to improve the efficiency of the proposed model.

In summary, the SSS-Net demonstrates significant superiority and originality. We have conducted experiments using comparative methods under the same experimental conditions and hyperparameter settings to verify and support this conclusion. For specific experimental details, such as the network backbone, parameter setup, and dataset partitioning, refer to the Experimental Setup section. Here, we provide a more detailed analysis of our conclusions based on the specific experimental results. Table 2 shows the best test accuracy (%) across all epochs and the average test accuracy (%) during the last 20 epochs on the CIFAR-10 and CIFAR-100 datasets with different levels of label noise ranging from 20% to 80%. The comparison results show that our proposed method outperforms other methods for most noise ratios, particularly for low noise-ratio datasets. Additionally, the t-SNE method [46] was used to visualize the features of the testing images on CIFAR-10 with 20%, 50%, and 80% label noise.

Table 2
Comparison results with state-of-the-art methods on CIFAR-10 and CIFAR-100.

Dataset		CIFAR-10			CIFAR-100		
Method/Noise ratio		20%	50%	80%	20%	50%	80%
Cross-Entropy	Best	86.8	79.4	62.9	62	46.7	19.9
	Avg	82.7	57.9	26.1	61.8	37.3	8.8
O2U-Net (2019) [36]	Best	91.6	89.59	43.41	74.12	69.21	39.39
Co-teaching+ (2019) [15]	Best	89.5	85.7	67.4	65.6	51.8	27.9
	Avg	88.2	84.1	45.5	64.1	45.3	15.5
INCV (2019) [34]	Best	89.89	85.11	55.77	–	–	–
	Avg	89.71	84.78	52.27	–	–	–
MentorMix (2020) [37]	Best	95.6	–	81.0	78.6	–	41.2
DivideMix (2020) [15]	Best	96.1	94.6	93.2	77.3	74.6	60.2
	Avg	95.7	94.4	92.9	76.9	74.2	59.6
RoCL (2021) [23]	Best	–	–	86.02	–	–	54.51
Augmentation (2021) [24]	Best	96.3	95.4	93.8	79.5	77.2	**66.4**
	Avg	96.2	95.1	93.6	79.2	77.0	**66.1**
Sel-CL+ (2022) [38]	Best	95.5	93.9	89.2	76.5	72.4	59.6
UNICON (2022) [39]	Best	96.0	95.6	93.9	78.9	77.6	63.9
ScanMix (2023) [25]	Best	96.0	94.5	93.5	77.0	75.7	66.0
	Avg	95.7	93.9	92.6	76.0	75.4	65.0
LongMix (2023) [26]	Best	96.18	95.19	**94.09**	78.03	75.84	62.24
	Avg	95.98	94.79	**93.73**	77.56	74.87	61.60
Ours	Best	**97.00**	**95.79**	92.02	**81.09**	**77.92**	61.51
	Avg	**96.61**	**95.60**	91.69	**80.31**	**77.10**	61.12

Table 3
Comparison results with state-of-the-art methods on the WebVision and ImageNet.

Dataset	WebVision		ILSVRC2012	
Method	top1	top5	top1	top5
INCV (2019) [34]	65.24	85.34	61.60	84.98
MentorMix (2020) [37]	76.00	90.20	72.90	91.10
DivideMix (2020) [15]	77.32	91.64	75.20	90.84
RoCL (2021) [23]	80.04	92.68	75.81	92.28
DSOS (2022) [40]	78.76	92.32	75.88	92.36
Sel-CL+ (2022) [38]	79.96	92.64	**76.84**	93.04
LongMix (2023) [26]	78.92	92.32	–	–
ScanMix (2023) [25]	**80.04**	93.04	75.76	92.60
Ours	79.32	**93.32**	75.96	**93.42**

Table 4
Comparison results with state-of-the-art methods on Clothing1M.

Dataset	Clothing1M
Meta-Learning (2019) [41]	73.47
P-correction (2019) [42]	73.49
PTD-R-V (2020) [43]	71.67
BARE (2021) [44]	72.28
CTRR (2022) [45]	72.90
DSOS (2022) [40]	73.63
Ours	**73.99**

Fig. 6 illustrates these clusters, where each cluster represents a category and different colors correspond to different labels. The three figures on the bottom layer show the testing images with varying ratios of noisy labels, while the upper layer indicates that the testing images are accurately categorized into 10 classes with true labels. This demonstrates that our model performs well in terms of learning the true class distribution and is robust to noisy labels.

Table 3 summarizes the baseline accuracies reported in the corresponding papers. It is evident that SSS-Net outperforms other methods in terms of top-1 and top-5 accuracy on the validation datasets of mini-WebVision and ImageNet ILSVRC12. Table 4 reports the comparison results of various methods, including Meta-Learning and DSOS, on the Clothing1M dataset. The proposed method demonstrates a significant advantage in test accuracy over the other three methods. Additionally, the proposed method also exhibits a narrow advantage over the P-correction and Meta-Learning methods.

Table 5
Results of ablation study results on CIFAR-10 and CIFAR-100.

Dataset		CIFAR-10			CIFAR-100		
Methods/Noise ratio		20%	50%	80%	20%	50%	80%
Proposed methods	Best	**97.00**	**95.79**	**92.02**	**81.09**	**77.92**	**61.51**
	Avg	**96.61**	**95.60**	**91.69**	**80.31**	**77.10**	**61.12**
w/o LSCM	Best	96.09	92.69	79.26	76.81	64.75	51.16
	Avg	95.12	91.71	78.83	76.46	64.28	50.78
w/o Shadowed sets	Best	96.83	95.75	91.75	80.86	77.37	60.10
	Avg	96.54	95.51	91.34	79.99	76.57	59.41
w/o HPSR	Best	96.85	95.69	91.23	80.84	77.31	61.17
	Avg	96.58	95.41	90.90	80.03	76.64	60.70

4.4. Ablation study

In this section, we validate the effectiveness of each module in the proposed method by conducting ablation experiments. The ablation studies, by removing different components from the proposed method, can be implemented on the CIFAR-10 and CIFAR-100 datasets to provide insights into what makes the proposed model successful. The results in Table 5 are the best test accuracy and the average accuracy of the last 20 epochs when the accuracy of model is stable, where the decrease in accuracy suggests the indispensable of each component. Specifically, we can see that LSCM significantly improves the accuracy of the model, demonstrating the effectiveness of sample losses and clustering used in this paper for selecting clean samples. Compared with the method with manually setting thresholds, the method of automatically calculating thresholds based on shadowed-sets theory can improve the accuracy by an average of 0.85% and a maximum of 1.71% on the CIFAR-100 dataset with different noise rates. This indicates that the automatically obtained thresholds are more reasonable, and, as shown in Table 6, this method of automatically calculating thresholds has little effect on the

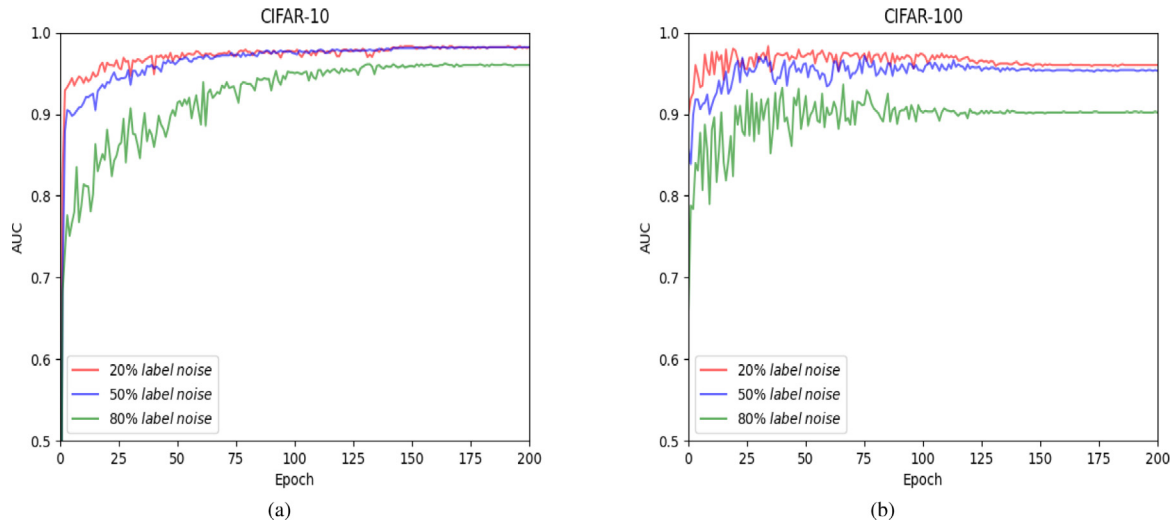

Fig. 5. Results of AUC on CIFAR-10 and CIFAR-100.

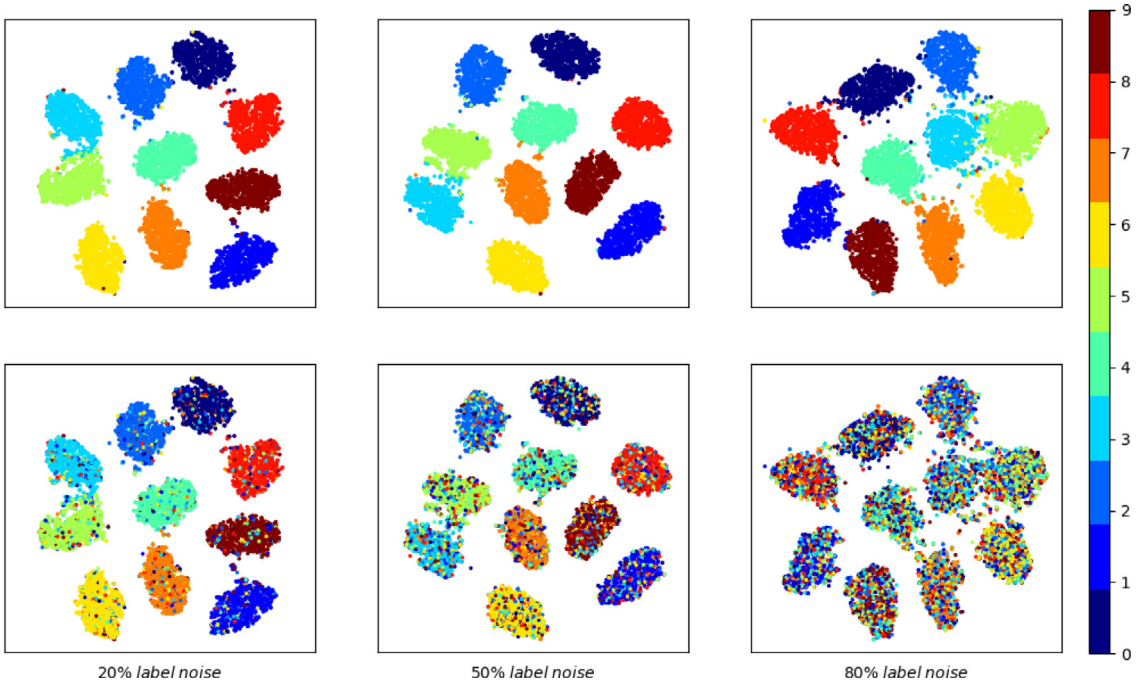

Fig. 6. Visualized results of the SSS-Net on the CIFAR-10 with 20%, 50% and 80% label noise.

complexity of model training. The HPSR preserves high-quality pseudo-labeled samples and discards low-quality ones, which decreased computation time by around 12%–25% across various noise rates on CIFAR-10, as shown in Table 6. Moreover, on the CIFAR-10 dataset with different noise rates, the accuracy can be improved by an average of 0.36% and a maximum of 0.79%.

4.5. Model evaluation

Fig. 5 shows the AUC (Area Under Curve) values of the SSS-Net model on CIFAR-10 and CIFAR-100 datasets during the first 200 epochs. One can see that our method has high reliability and stability both at low label noise and higher label noise. Furthermore, it is also apparent that the proposed method has significant advantages in terms of 20% label noise and 50% label noise over the 80% label noise, which can also be supported and verified as illustrated in Table 2.

To evaluate the efficiency of the proposed model, we conducted an experiment comparing the computing time of the backbone and the backbone combine with three modules (LSCM, shadow-sets-based adaptive selection, and HPSR) over 200 epochs on the CIFAR-10 dataset. We tested the models under varying label noise rates of 20%, 50%, and 80%. Our results, presented in Table 6, show that combining LSCM increased computation time by approximately 10%–18% across different noise rates. The shadowed-sets module for automatic threshold calculation slightly increased computation time by about 2%–5% compared to the backbone. However, HPSR, which consists of two parts, reselection of high-quality samples and discarding low-quality noisy samples, reduced computation time by approximately 12%–25% across various noise rates.

Table 6
Time spent of each module on CIFAR-10.

Dataset	CIFAR-10		
Noise ratio	20%	50%	80%
Backbone	723.0 m	638.0 m	519.0 m
Backbone w/LSCM	855.0 m	703.0 m	605.0 m
Backbone w/Shadowed sets	760.0 m	674.0 m	529.0 m
Backbone w/HPSR	537.0 m	493.0 m	458.0 m

5. Conclusions

A shadowed-sets-based semi-supervised sample selection network, namely SSS-Net, is developed in this study for classification on noise labeled images. Apparently, each component involved in the proposed model is indispensable, which is supported in the ablation studies subsection. The superiority and effectiveness of the proposed LSCM, shadowed-sets-based adaptive sample selection method and the HPSR approach are supported and verified by a great deal of experiments using synthetic and real-world datasets that are compared with state-of-the-art methods for classification of noise labeled images. And we demonstrated the stability of the model by calculating the AUC value. In the future, it could be worth studying the combination of learning from noisy labels with other aspects of learning, such as label-free sample learning.

CRediT authorship contribution statement

Kecan Cai: Writing – review & editing, Writing – original draft, Validation, Software, Investigation. **Hongyun Zhang:** Writing – review & editing, Supervision, Methodology. **Witold Pedrycz:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Duoqian Miao:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that there is no conflict of interest, such as personal or professional relationships, affiliations, knowledge or beliefs, that may affect the subject matter or materials discussed in the manuscript.

Data availability

Data will be made available on request.

Acknowledgments

We sincerely thank the editors and the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China under Grants 62076182, 61976158, 61976160 and 62076184, and the Natural Science Foundation of Shanghai, China under Grant 22ZR1466700.

References

[1] G. Algan, I. Ulusoy, Image classification with deep learning in the presence of noisy labels: A survey, Knowl.-Based Syst. 215 (2021) 106771.
[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[3] Q. Jin, M. Yuan, H. Wang, M. Wang, et al., Deep active learning models for imbalanced image classification, Knowl.-Based Syst. 257 (2022) 109817.
[4] A. Kuznetsova, H. Rom, N. Alldrin, et al., The open images dataset v4, Int. J. Comput. Vis. 128 (7) (2020) 1956–1981.
[5] R. Tanno, A. Saeedi, S. Sankaranarayanan, et al., Learning from noisy labels by regularized estimation of annotator confusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11244–11253.
[6] Y. Li, J. Yang, Y. Song, et al., Learning from noisy labels with distillation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1910–1918.
[7] H. Song, M. Kim, D. Park, et al., Learning from noisy labels with deep neural networks: A survey, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[8] K. Lee, S. Yun, K. Lee, et al., Robust inference via generative classifiers for handling noisy labels, in: International Conference on Machine Learning, 2019, pp. 3763–3772.
[9] Z. Zeng, X. Wang, W. Li, et al., Two-stage natural scene image classification with noise discovering and label-correlation mining, Knowl.-Based Syst. 260 (2023) 110137.
[10] Y. Wu, J. Li, Y. Yuan, et al., Commonality autoencoder: Learning common features for change detection from heterogeneous images, IEEE Trans. Neural Netw. Learn. Syst. 33 (2021) 4257–4270.
[11] Y. Wu, D. Hangqi, M. Gong, et al., Evolutionary multiform optimization with two-stage bidirectional knowledge transfer strategy for point cloud registration, IEEE Trans. Evol. Comput. PP (2022).
[12] J. Yao, J. Wang, I.W. Tsang, et al., Deep learning from noisy image labels with quality embedding, IEEE Trans. Image Process. 28 (4) (2018) 1909–1922.
[13] G. Zheng, A.H. Awadallah, S. Dumais, Meta label correction for noisy label learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 11053–11061.
[14] P. Wu, S. Zheng, M. Goswami, et al., A topological filter for learning with label noise, Adv. Neural Inf. Process. Syst. 33 (2020) 21382–21393.
[15] L. Junnan, S.C. Hoi, DivideMix: Learning with noisy labels as semi-supervised learning, in: ICLR. International Conference on Learning Representations, ICLR, 2020.
[16] T. Zhou, S. Wang, J. Bilmes, Robust curriculum learning: from clean label detection to noisy label self-correction, in: International Conference on Learning Representations, 2020.
[17] E. Arazo, D. Ortego, P. Albert, et al., Unsupervised label noise modeling and loss correction, in: International Conference on Machine Learning, 2019, pp. 312–321.
[18] E. Malach, S. Shalev-Shwartz, Decoupling" when to update" from" how to update", in: Advances in Neural Information Processing Systems, 2017.
[19] L. Jiang, Z. Zhou, T. Leung, et al., Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels, in: International Conference on Machine Learning, 2018, pp. 2304–2313.
[20] D. Arpit, S. Jastrzbski, N. Ballas, et al., A closer look at memorization in deep networks, in: International Conference on Machine Learning, 2017, pp. 233–242.
[21] A. Ghosh, A. Lan, Contrastive learning improves model robustness under label noise, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2703–2708.
[22] H. Song, M. Kim, J. Lee, SELFIE: Refurbishing unclean samples for robust deep learning, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 5907–5915.
[23] T. Zhou, S. Wang, J. Bilmes, Robust curriculum learning: from clean label detection to noisy label self-correction, in: International Conference on Learning Representations, 2021.
[24] K. Nishi, Y. Ding, A. Rich, et al., Augmentation strategies for learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8022–8031.
[25] R. Sachdeva, F.R. Cordeiro, V. Belagiannis, et al., ScanMix: learning from severe label noise via semantic clustering and semi-supervised learning, Pattern Recognit. 134 (2023) 109121.
[26] F.R. Cordeiro, R. Sachdeva, V. Belagiannis, et al., LongReMix: Robust learning with high confidence samples in a noisy label environment, Pattern Recognit. 133 (2023) 109013.
[27] H. Xu, H. Xiao, H. Hao, L. Dong, X. Qiu, C. Peng, Semi-supervised learning with pseudo-negative labels for image classification, Knowl.-Based Syst. 260 (2023) 110166.
[28] P. Wu, S. Zheng, M. Goswami, et al., A topological filter for learning with label noise, in: Advances in Neural Information Processing Systems, 2020, pp. 21382–21393.
[29] Y. Wu, Y. Zhang, X. Fan, et al., Inenet: Inliers estimation network with similarity learning for partial overlapping registration, IEEE Trans. Circuits Syst. Video Technol. 33 (2023) 1413–1426.
[30] W. Pedrycz, Shadowed sets: representing and processing fuzzy sets, IEEE Trans. Syst. Man Cybern. B 28 (1) (1998) 103–109.
[31] X. Deng, Y. Yao, Decision-theoretic three-way approximations of fuzzy sets, Inform. Sci. 279 (2014) 702–715.
[32] Y. Yao, Three-way decisions with probabilistic rough sets, Inform. Sci. 180 (3) (2010) 341–353.

[33] E. Arazo, D. Ortego, P. Albert, et al., Unsupervised label noise modeling and loss correction, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 312–321.
[34] P. Chen, B.B. Liao, G. Chen, et al., Understanding and utilizing deep neural networks trained with noisy labels, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 1062–1070.
[35] K. He, X. Zhang, S. Ren, et al., Identity mappings in deep residual networks, in: European Conference on Computer Vision, 2016, pp. 630–645.
[36] J. Huang, L. Qu, R. Jia, et al., O2U-net: A simple noisy label detection approach for deep neural networks, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 3325–3333.
[37] L. Jiang, D. Huang, M. Liu, et al., Beyond synthetic noise: Deep learning on controlled noisy labels, in: International Conference on Machine Learning, 2020, pp. 4804–4815.
[38] S. Li, X. Xia, S. Ge, et al., Selective-supervised contrastive learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 316–325.
[39] N. Karim, M.N. Rizve, N. Rahnavard, et al., UNICON: Combating label noise through uniform selection and contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9676–9686.
[40] P. Albert, D. Ortego, E. Arazo, et al., Addressing out-of-distribution label noise in webly-labelled data, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 392–401.
[41] J. Li, M. Khodak, S. Caldas, et al., Differentially private meta-learning, 2019, arXiv preprint arXiv:1909.05830.
[42] K. Yi, J. Wu, Probabilistic end-to-end noise correction for learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7017–7025.
[43] X. Xia, T. Liu, B. Han, et al., Part-dependent label noise: Towards instance-dependent label noise, Adv. Neural Inf. Process. Syst. 33 (2020) 7597–7610.
[44] D. Patel, P. Sastry, Adaptive sample selection for robust learning under label noise, 2021, arXiv preprint arXiv:2106.15292.
[45] L. Yi, S. Liu, Q. She, et al., On learning contrastive representations for learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16682–16691.
[46] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) (2008).