applied sciences

MDPI

Article

FGCM: Noisy Label Learning via Fine-Grained Confidence Modeling

Shaotian Yan [1], Xiang Tian [2], Rongxin Jiang [2] and Yaowu Chen [3,*]

[1] College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310007, China
[2] Zhejiang Provincial Key Laboratory for Network Multimedia Technologies, Hangzhou 310007, China
[3] Zhejiang University Embedded System Engineering Research Center, Ministry of Education of China, Hangzhou 310007, China
* Correspondence: cyw@mail.bme.zju.edu.cn

Abstract: A small portion of mislabeled data can easily limit the performance of deep neural networks (DNNs) due to their high capacity for memorizing random labels. Thus, robust learning from noisy labels has become a key challenge for deep learning due to inadequate datasets with high-quality annotations. Most existing methods involve training models on clean sets by dividing clean samples from noisy ones, resulting in large amounts of mislabeled data being unused. To address this problem, we propose categorizing training samples into five fine-grained clusters based on the difficulty experienced by DNN models when learning them and label correctness. A novel fine-grained confidence modeling (FGCM) framework is proposed to cluster samples into these five categories; with each cluster, FGCM decides whether to accept the cluster data as they are, accept them with label correction, or accept them as unlabeled data. By applying different strategies to the fine-grained clusters, FGCM can better exploit training data than previous methods. Extensive experiments on widely used benchmarks CIFAR-10, CIFAR-100, clothing1M, and WebVision with different ratios and types of label noise demonstrate the superiority of our FGCM.

Keywords: noisy labeled data; robust learning; image classification

Citation: Yan, S.; Tian, X.; Jiang, R.; Chen, Y. FGCM: Noisy Label Learning via Fine-Grained Confidence Modeling. *Appl. Sci.* **2022**, *12*, 11406. https://doi.org/10.3390/app122211406

Academic Editors: Wonjoon Kim, Sekyoung Youm and Sungbum Jun

Received: 29 September 2022
Accepted: 3 November 2022
Published: 10 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The double-edged nature of deep neural networks' tremendous capacity to learn and memorize data has been revealed [1–3]. However, although DNNs trained on large-scale datasets achieved massive success in image classification [4], object detection [5], and other visual tasks, they tend to undesirably memorize random false-labeled samples [3], resulting in poor generation ability. Currently, there is growing research interest in noisy label learning [6–13] owing to the scarcity of large-scale datasets with high-quality annotations in most real-world cases [14]. Recent methods for noisy label learning [6,8,12,15–19] are mostly inspired by the experimental observation [3] that DNNs tend to learn patterns existing among samples early and later struggle to memorize hard samples. These methods divide clean samples from noisy ones by treating samples with higher prediction confidence as clean instances while others as noisy instances [1]. For example, MentorNet [16] trains a student network by feeding in small-loss samples selected by a teacher model. Co-teaching [17] and co-teaching+ [18] train two DNNs parallelly, where each network provides small-loss samples to the other. Mcorr [8] estimates the probability of a sample being clean by fitting a two-component beta mixture model on prediction loss. Here, a more robust representation can be learned by excluding samples with low certainty of being clean from further training.

However, the performance of the methods above is limited because it leaves many unused mislabeled and hard samples that contain helpful information [2]. Therefore, we rethink confidence modeling from a more fine-grained perspective of considering the difficulty experienced by DNN models when learning from samples. Most existing methods

overlook the different properties between hard and simple instances. Simple samples have easily identifiable features, while hard samples are rather indistinct. Here, instead of dividing the training data into either clean or noisy, we naturally categorize samples into four kinds: Clean Simple(CS), Noisy Simple(NS), Clean Hard(CH), and Noisy Hard. The Noisy Hard set is further divided into samples with relevant annotations (NHR) and irrelevant labels (NHI). Figure 1 presents the significant differences among the confidence trends of typical examples for each category during the training process, indicating the feasibility of fine-grained categorization. Furthermore, it can be inferred from Figure 1 that the confidence scores of maximum repeating pseudo labels share the same or similar trends with those of ground truth labels during the early training stage for all the five clusters. Thus, we use the former as an alternative when generating sample categories for the latter because the latter is not available in real-world cases.

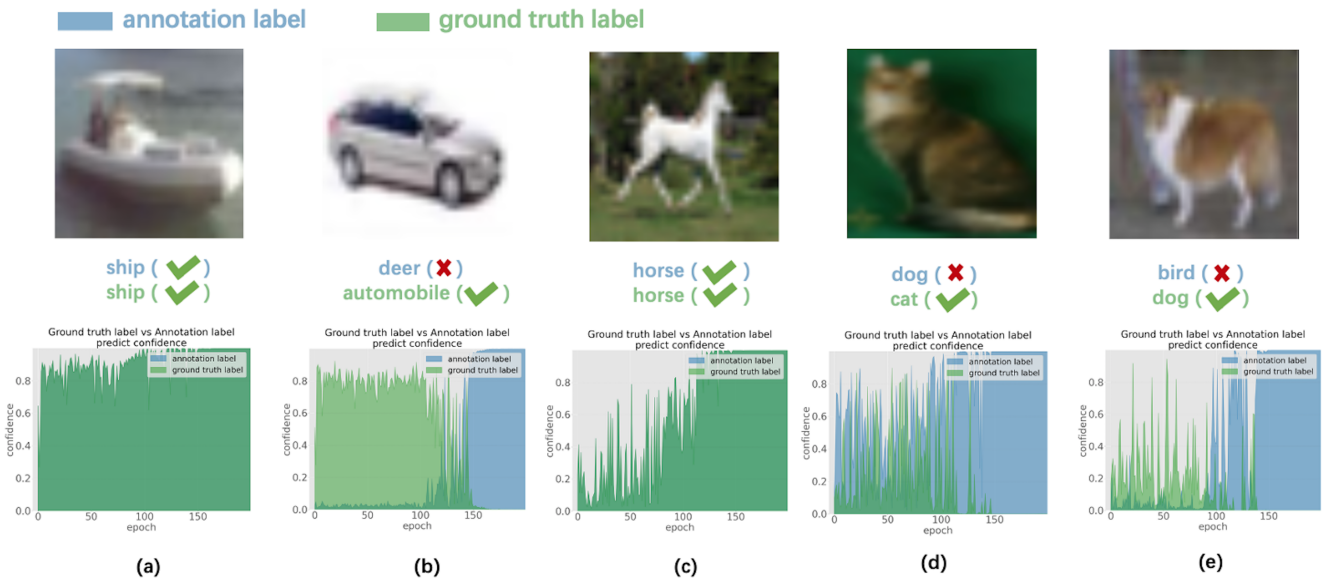

Figure 1. Typical examples of images in the training set and corresponding confidence score trends for different kinds of samples on CIFAR-10 with 20% symmetrical noise. The words in blue below each image are their annotation labels which might be wrong, while the green words are the ground truth labels. The crosses or ticks in brackets indicate the rightness of corresponding labels. We train a PreAct-18 resnet using SGD with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 128. The initial learning rate is set to 0.06 and reduced to 0.006 in the 80th epoch. The blue and green curves represent the trends of confidence scores for annotation labels and ground truth labels during the training process, respectively. (**a**) Clean Simple samples (CS) (i.e., highly distinguishable instances with correct annotations). (**b**) Noisy Simple samples (NS) (i.e., mislabeled samples with easily identifiable features). (**c**) Clean Hard samples (CH). (**d**) Noisy Hard samples with Relevant annotations (NHR). (**e**) Noisy Hard samples with Irrelevant labels (NHI).

In this paper, we aim to robustly train deep neural networks on noisily labeled datasets by recalling as much reusable training samples as possible. As mentioned above, existing sample selection methods treat samples with higher prediction confidence as clean instances. In this way, only a small portion of data can be selected. Moreover, it is hard to distinguish hard samples from noisy samples, both having low prediction confidence. Therefore, inspired by the findings induced from Figure 1, we propose a simple yet effective Fine-Grained Confidence Modeling (FGCM) framework. First, FGCM clusters samples into the five categories mentioned above by modeling the prediction scores for both annotations and the maximum repeating pseudo labels of samples. Second, to include as many clean and reusable samples in training, FGCM applies different strategies to different categories. In particular, we take data from CS and CH as it is, correct labels from NS to the maximum

repeating pseudo labels, since there is a high possibility for the maximum repeating pseudo labels of NS instances to be correct according to Figure 1, and drop the labels of NHI together with NHR which are ambiguous. By iteratively repeating the process, FGCM can effectively prevent models from over-fitting to false-labeled samples and gradually recall more reusable samples from dropped data to further strengthen the performance of the trained model. Finally, semi-training strategy is employed to realize the use of NHI and NHR sets, enhancing the performance of trained model. With the fine-grained categorization of samples, FGCM can maximize the use of the training data.Extensive experiments on various datasets show that FGCM outperforms SOTA methods.

Our main contributions are summarized as follows:

- We propose fine-grained training sample categories based on the difficulty of learning them and their label correctness.
- We propose a simple yet effective FGCM framework for noisy label learning, which can effectively maximize the useation of the training data and prevent models from over-fitting to noisy labels.
- Extensive experiments show the effectiveness and robustness of FGCM under different ratios and types of noise, outperforming previous methods.

2. Related Work

In this section, we briefly review the existing literature on noisy label learning. As listed in Table 1, previous methods can be grouped into several categories.

Noise transition matrix. Many early studies tackle the task of noisy label learning by estimating a noise transition matrix which represents the probability of one class being mislabeled into another. Some methods try to improve the performance by adjusting the prediction results of DNNs using the estimated label transition probability. Other methods, such as Fcorr [20], correct the loss by multiplying the estimated transition probability with the softmax outputs. The key to these methods is the reliability of the estimated noise transition matrix. Webly learning [21] estimates a noise transition matrix by training an additional noise adaptation layer. Probabilistic noise modeling [22] trains two networks which are used to predict the noise type and label transition probability, respectively.

The performance of these methods is often limited in realistic cases due to the complexity of estimating noise transition matrix on real-world datasets.

Table 1. Summary of literature reviewed according to the category of methods.

Category	Methods
Noise transition matrix	Fcorr [20], Webly learning [21], Probabilistic noise modeling [22]
Robust Loss Function	Robust MAE [23], Generalized Cross Entropy [24], Symmetric Cross Entropy [25]
Sample Selection	MentorNet [16], Co-teaching [17], Co-teaching+ [18], Iterative-CV [26]
Hybrid Methods	SELF [27], Mcorr [8], DivideMix [19], MOIT+ [28], Sel-CL+ [12]

Robust Loss Function. Some studies try to design noise tolerant loss functions. The robust MAE [23] showed that the mean absolute error (MAE) loss is more robust to label noise than cross entropy (CE) loss, yet it is not applicable on complicated real world datasets and suffered from increased difficulty in training compared with cross entropy loss. Thus, the generalized cross entropy (GCE) [24] is introduced, trying to combine the advantages of both MAE and CE loss. Symmetric cross entropy (SCE) [25] proposes a noise robust term called reverse cross entropy loss, which can improve the learning on hard classes. SCE archives significantly better performance than CE and GCE. However, the accuracy of these

loss function methods is still not promising and their performance will be greatly degraded in hard cases.

Sample Selection. Many recent studies adopted sample selection strategies which select a certain number of possibly clean labeled samples from the training dataset and exclude other samples from the update to prevent the model from over-fitting to noisy labels. The small-loss trick is widely used to distinguish true-labeled samples from noisy samples because the losses of clean samples tend to be smaller than false-labeled samples due to the memorization effect of DNNs. MentorNet [16] trains a student network by feeding in small-loss samples selected by a teacher model. Co-teaching [17] and Co-teaching+ [18] train two networks simultaneously and each network selects small-loss samples for the training of the other model. Iterative-CV [26] randomly divides training samples into subsets and iteratively removes samples with large losses using a cross validation strategy. Despite the significant improvements these methods obtain, samples acquired by them are mostly clean simple instances. Hard samples containing rich information which can greatly help the generalization of DNNs are discarded. Moreover, in extreme circumstances where the noise rate is high, too few samples can be selected to train the model thoroughly. Last but not least, the ground truth noise rate or a clean validation dataset, which are not available in most cases, need to be provided to decide how many samples should be selected.

Hybrid Methods. Recent studies attempted to combine sample selection with other techniques. SELF [27] uses the mean-teacher model to progressively filter out noisy samples. Mcorr [8] proposes to fit a two-component mixture model on prediction loss to estimate the probability of a sample being clean or noisy. Following Mcorr, DivideMix [19] divides training samples into a clean or noisy set using the obtained probabilities. Then semi-supervised learning is performed with instances in the clean set treated as labeled data while those in noisy set as unlabeled data.

There are some methods try to alleviate the effect of noisy labels using supervised contrastive learning techniques. MOIT+ [28] pretrains models with supervised contrastive learning. Then the learned features are used to divide samples into clean or noisy. Then a classifier is trained in a semi-supervised manner on the divided datasets. Sel-CL+ [12] achieves approximately 10% higher accuracy than MOIT+ on severely noisy datasets by pretraining models with unsupervised contrastive learning to completely remove the impact of noisy labels. Then, Sel-CL+ uses the learned low-dimensional representations to select confident pairs of samples for the supervised contrastive training of models instead of selecting confident samples as MOIT+ does. In this way, Sel-CL+ is able to leverage supervised information from not only the pairs with correct annotations, but also the pairs which are false-labeled from the same class.

Although these methods achieved promising performance, the improvements mostly are mostly owed to the using of semi-supervised or contrastive learning techniques. As with traditional sample selection methods, they still leave a large part of helpful supervisory signal unused, particularly from large loss training samples which consist of noisy labeled samples and hard samples.

3. Proposed Method

In this section, in detail, we introduce FGCM, our proposed training framework for learning from noisy labels. In summary, FGCM consists of three steps. (a) Generating fine-grained sample clusters based on their confidence trends. (b) Performing cluster refinement to promote the accuracy of sample clusters and constructing a cleaned training set with samples from CS, CH, and NS sets. Step (a) and (b) are iteratively repeated in the training process to expand the cleaned training set. (c) Performing mixed semi-training with instances from NHI and NHR sets treated as unlabeled data and instances in the cleaned training set as labeled data to maximum the use of training samples. The three steps will be introduced in detail in the following subsections.

First, to avoid confusion, we define the notations of the three kinds of labels which will be frequently mentioned for the rest of the paper. Annotation labels, denoted as $\mathcal{AY} = \{ay_1, ay_2, \ldots, ay_n\}$, are the assigned labels of samples, part of which are false-labeled while others are correct. $\mathcal{GY} = \{gy_1, gy_2, \ldots, gy_n\}$ refers to ground truth labels which are the correct labels of samples. Maximum repeating pseudo labels $\mathcal{MY} = \{my_1, my_2, \ldots, my_n\}$ are the most frequent pseudo labels for each sample predicted by DNNs from the beginning of training to the current epoch.

3.1. Generalization of Sample Clusters

The process of generating sample clusters and performing cluster refinements is presented in Algorithm 1 which is located in page 6. Let $\mathcal{S} = (\mathcal{X}, \mathcal{AY}) = \{(x_i, ay_i)\}_{i=1}^{N}$ denote the training set consists of N samples, where x_i is an input image with ay_i being the corresponding annotation label. To obtain the maximum repeating pseudo label my_i for sample x_i, we need to record the history prediction confidence for every sample as C^{e_k}. Given a model with parameter θ, the confidence score $c_i^{e_k}(\tilde{y})$ represents the probability of sample i to be classified into class $\tilde{y}$ after kth epoch of training:

$$c_i^{e_k}(\tilde{y}) = \mathcal{P}_{model}^{\tilde{y}}(x_i; \theta_{e_k}) \tag{1}$$

where $\mathcal{P}_{model}^{\tilde{y}}$ refers to the model's output softmax probability for class $\tilde{y}$. We also maintain a history prediction list $\mathcal{H}^{e_k}$ which contains the pseudo labels produced by the model for each sample from the beginning of the training process to the current epoch k:

$$\mathcal{H}^{e_k} = \{\{\arg\max \mathcal{P}_{model}(x_i; \theta_{e_j})\}_{i=1}^{N}\}_{j=0}^{k} \tag{2}$$

where $\arg\max \mathcal{P}_{model}(x_i; \theta_{e_j})$ refers to the predicted label with maximum confidence, as known as the pseudo label, for sample i in epoch j. We use the label with maximum number of occurrences in $\mathcal{H}_i^{e_k}$ as my_i for sample i. The model is trained on the whole training set in the early stage when the impact of noisy labels is minor. After the warmup phase ends, a joint confidence score $jc_i^{e_{k-1}}$ is constructed for sample i in every epoch k:

$$jc_i^{e_{k-1}} = \left(\frac{\sum_{j=0}^{k-1} c_i^{e_j}(ay_i)}{k}, \frac{\sum_{j=0}^{k-1} c_i^{e_j}(my_i)}{k}\right) \tag{3}$$

The two elements in $jc_i^{e_{k-1}}$ represent the average of historical prediction confidence for annotation and maximum repeating pseudo label, respectively, reflecting the trends of their confidence scores during the training process. Gaussian Mixture Model (GMM) [29] is a widely used unsupervised modeling technique. A Gaussian Mixture Model is a weighted sum of M component Gaussian densities, where M is the number of clusters in the data. GMM assumes that the input data points are generated from a mixture of M Gaussian distributions. With parameter estimation using the Expectation-Maximization algorithm [30], GMM outputs the probabilities of samples belonging to each cluster. In this paper, we categorize samples into the cluster with maximum possibility. Owing to the different trends showed in Figure 1, five clusters of sample points can be generated by fitting a five-component Gaussian Mixture Model(GMM) [29] to $jc^{e_{k-1}}$. The clustering result, presented in Figure 2b, is highly consistent with that in Figure 2a, which is produced by replacing my_i with gy_i, indicating the effectiveness of using $\mathcal{MY}$ as the alternative for $\mathcal{GY}$. The cluster with the highest average confidence score of ay_i is CS, and the second-highest cluster is CH. The cluster with the maximum difference between the average scores of my_i and ay_i is NS. The other two clusters are NHR and NHI, whose labels are ambiguous and less valuable to training models. Moreover, the clusters can be further refined with priors as described in the next subsection. Consequently, we can obtain a new training set $\mathcal{T}$ with the cleanest labels by screening out instances from NHR and NHI and rectifying NS samples' labels with their maximum repeating pseudo labels. It is worth noting that though NHR and NHI are not used in training, their existences improve the clustering performance. Training on a

cleaned dataset effectively prevents the model from over-fitting to false-labeled samples. Moreover, the cleaned training set $\mathcal{T}$ can be expanded with more reusable samples gradually recalled from NHR and NHI by iteratively repeating the process to strengthen the performance of the trained model further.

Algorithm 1: Generalization of Sample Clusters and Cluster Refinements

Input: θ, training dataset $\mathcal{S} = (\mathcal{X}, \mathcal{AY}) = \{(x_i, ay_i)\}_{i=1}^{N}$

for $k = 0$ *to MaxEpoch* **do**
  if $k <$ *WarmUpEpoch* **then**
    /*training on the whole $\mathcal{S}$*/
    $\theta = WarmUp(\mathcal{S}, \theta)$
  else
    /*find the maximum repeating pseudo labels*/
    $\mathcal{MY} = \{my_i\}_{i=1}^{N} \leftarrow max(\mathcal{H})$
    /*construct a joint confidence score*/
    $jc_i^{e_{k-1}} = (\frac{\sum_{j=0}^{k-1} c_i^{e_j}(ay_i)}{k}, \frac{\sum_{j=0}^{k-1} c_i^{e_j}(my_i)}{k})$
    /*generate clusters with GMM*/
    $\mathcal{NS}, \mathcal{CS}, \mathcal{CH}, \mathcal{NHR}, \mathcal{NHI} \leftarrow GMM(jc^{e_{k-1}})$
    /* performing cluster refinements */
    $\mathcal{X}_{\widetilde{\mathcal{NS}}} = \{x_{ay_i \neq my_i}\}; x_i \in \mathcal{X}_{\mathcal{NS}}$
    $\mathcal{X}_{\widetilde{\mathcal{CS}}} = \{x_{ay_i \equiv my_i}\}; x_i \in \mathcal{X}_{\mathcal{CS}}$
    $\mathcal{X}_{\widetilde{\mathcal{CH}}} = \{x_{ay_i \equiv my_i}\}; x_i \in \mathcal{X}_{\mathcal{CH}}$
    /* constructing the new training set $\mathcal{T}$ */
    if $k ==$ *WarmUpEpoch* **then**
      $\mathcal{T} = (\mathcal{X}_{\widetilde{\mathcal{NS}}}, \mathcal{MY}_{\widetilde{\mathcal{NS}}}) \cup (\mathcal{X}_{\widetilde{\mathcal{CS}}}, \mathcal{AY}_{\widetilde{\mathcal{CS}}}) \cup (\mathcal{X}_{\widetilde{\mathcal{CH}}}, \mathcal{AY}_{\widetilde{\mathcal{CH}}})$
    else
      $\mathcal{T} = \mathcal{T} \cup (\mathcal{X}_{\widetilde{\mathcal{NS}}}, \mathcal{MY}_{\widetilde{\mathcal{NS}}}) \cup (\mathcal{X}_{\widetilde{\mathcal{CS}}}, \mathcal{AY}_{\widetilde{\mathcal{CS}}}) \cup (\mathcal{X}_{\widetilde{\mathcal{CH}}}, \mathcal{AY}_{\widetilde{\mathcal{CH}}})$
    end
    $\theta = Train(\mathcal{T}, \theta)$
  end
  /*record history confidence and prediction*/
  $\mathcal{C}^{e_k} = \{\mathcal{P}_{model}(x_i; \theta)\}_{i=1}^{N}$
  $\mathcal{H}^{e_k} = \{\arg\max \mathcal{C}_i^{e_k}\}_{i=1}^{N}$
end

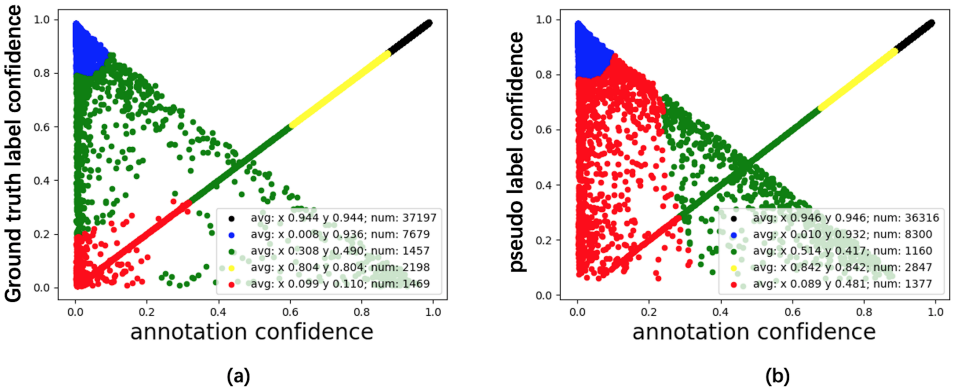

Figure 2. Sample clusters generated by FGCM. (**a**) is generated using confidence score of ground truth labels and annotation labels. (**b**) is generated using confidence score of maximum repeating pseudo labels and annotation labels.

3.2. Cluster Refinement

Despite possessing a reasonably high number of accurately labeled samples, the cleaned training set still inevitably contains a small part of wrongly clustered samples, especially when the noise rate of the original dataset $\mathcal{S}$ is high. In NS instances, my_i and ay_i should differ with each other. However, my_i and ay_i should be the same in CS and CH instances. After acquiring the fine-grained clustering results, we use this prior knowledge to perform cluster refinements and obtain cleaned NS, CS as well as CH, filtering out outlier samples:

$$\mathcal{X}_{\widetilde{\mathcal{NS}}} = \{x_{ay_i \neq my_i}\}; x_i \in \mathcal{X}_{\mathcal{NS}}; \mathcal{X}_{\widetilde{\mathcal{CS}}} = \{x_{ay_i \equiv my_i}\}; x_i \in \mathcal{X}_{\mathcal{CS}}; \mathcal{X}_{\widetilde{\mathcal{CH}}} = \{x_{ay_i \equiv my_i}\}; x_i \in \mathcal{X}_{\mathcal{CH}}$$

where $\mathcal{X}_{\mathcal{NS}}, \mathcal{X}_{\mathcal{CS}}, \mathcal{X}_{\mathcal{CH}}$ are the $\mathcal{NS}$ set, $\mathcal{CS}$ set and $\mathcal{CH}$ set generated by fine-grained clustering. Furthermore, $\mathcal{X}_{\widetilde{\mathcal{NS}}}, \mathcal{X}_{\widetilde{\mathcal{CS}}}, \mathcal{X}_{\widetilde{\mathcal{CH}}}$ are the refined ones.

3.3. Mixed Semi-Training

With the sample selection steps introduced above, we already constructed a cleaned labeled set $\mathcal{X}_l$ which consists of N_l samples since the pseudo labels of CS, NS and CH clusters have a high probability of being correct:

$$\mathcal{X}_l = \{(x_i, l_i) : i \in (1, \ldots, N_l)\} = (\mathcal{X}_{\widetilde{\mathcal{NS}}}, \mathcal{MY}_{\widetilde{\mathcal{NS}}}) \cup (\mathcal{X}_{\widetilde{\mathcal{CS}}}, \mathcal{AY}_{\widetilde{\mathcal{CS}}}) \cup (\mathcal{X}_{\widetilde{\mathcal{CH}}}, \mathcal{AY}_{\widetilde{\mathcal{CH}}}).$$

where l_i is the corresponding cleaned label for sample x_i. To maximize the use of training samples, instead of dropping all the samples in NHI and NHR clusters, we drop their labels to construct an unlabeled set $\mathcal{X}_u$ consists of N_u samples:

$$\mathcal{X}_u = \{u_i : i \in (1, \ldots, N_u)\} = (\mathcal{X}_{\mathcal{NHR}}) \cup (\mathcal{X}_{\mathcal{NHI}})$$

where u_b refers to an unlabeled sample. We reinitialize the model to prevent over-fitting, and make use of the labeled as well as the unlabeled dataset to train it in a mixed semi-supervised way inspired by [31]. Cross entropy loss for labeled batches $\mathcal{L}_x$ is defined as:

$$\mathcal{L}_x = \frac{1}{B} CE(\mathcal{P}_{model}(x_i; \theta), l_i)_{i=1}^{B} \tag{4}$$

where B is the batch size of labeled batches. Strong augmentations have been proved to be beneficial to the training of model [31]. Sohn et al. take the whole training set as unlabeled data and applies strong augmentations to unlabeled data [31]. To speed up the semi-training process while improving the performance, we propose mixed batches $Batch_m = \{m_b : b \in (1, \ldots, (\mu + 1)B)\} \in \mathcal{X}_l \cup \mathcal{X}_u$ randomly sampled from both labeled data and unlabeled data. Every $Batch_m$ samples B samples from labeled dataset and μB samples from unlabeled dataset. Therefore, the batch size for mix batches is $(\mu + 1)B$.

For mixed batches, first we record the confidence scores for every sample m_b as $\mathcal{C}^m$.

$$\mathcal{C}^m = \mathcal{P}_{model}(m_b; \theta)_{b=1}^{(\mu+1)B} \tag{5}$$

Then, labels with maximum confidence for unlabeled data are employed as their pseudo labels Pse_b^m while labels for labeled data are kept:

$$Pse_b^m = \begin{cases} argmax(\mathcal{C}_b^m) & \text{if } m_b \in \mathcal{X}_u \\ l_b & \text{if } m_b \in \mathcal{X}_l \end{cases} \tag{6}$$

By performing various augmentations [31], the cross entropy loss for mixed batches is defined as:

$$\mathcal{L}_m = \frac{1}{(\mu + 1)B} \Sigma_{b=1}^{(\mu+1)B} \mathbb{1}_{max(\mathcal{C}_b^m) > \tau} CE(\mathcal{P}_{model}(\mathcal{A}(m_b); \theta), Pse_b^m) \tag{7}$$

where only predicted labels with confidence higher than τ are counted in the loss and $\mathcal{A}$ refers to augmentations. Finally, a coefficient factor λ_m is used to balance the loss of mixed batches and forms the overall loss.

$$loss = \mathcal{L}_x + \lambda_m \mathcal{L}_m \tag{8}$$

4. Experiments

4.1. Experiment Settings

4.1.1. Datasets

We validate our method on four commonly used benchmark datasets, including CIFAR-10, CIFAR-100 [32] with noisy synthetic labels and clothing1M [22], WebVision 1.0 [33] with real-world noise. CIFAR-10 and CIFAR-100 are accurately labeled with 50K training images and 10K test images. Clothing1M is a large-scale real-world dataset with a noise rate of approximately 38.5% [34]. It consists of one million training images collected from the web and 10,526 accurately labeled test images. WebVision contains 2.4 million training images with an estimated noise level of 20%. Following previous methods [15,19], we use a mini version of WebVision which consists of the first 50 classes from Google subset for training. The WebVision validation set and the ImageNet ILSVRC12 validation set are used for evaluating. To be consistent with previous works, we evaluated two kinds of noisy synthetic labels on CIFAR-10 and CIFAR-100. In symmetric noise (Sym) settings, every category has the same noise rate, and noisy labels are evenly distributed among all the classes. As for asymmetric noise (Asym), a certain percent of labels of each class are relabeled into a similar category (e.g., dog ← cat).

4.1.2. Implementation Details

We employed an 18-layer PreAct Resnet [35] as the backbone network for all the experiments on CIFAR-10 and CIFAR-100. The net was trained using SGD with a momentum of 0.9, a weight decay of 0.0005, and a batch size of 128. The network was trained for 350 epochs. We set the initial learning rate as 0.06 and reduced it by a factor of 10 after 80,150 epochs. The warmup period was 80 epochs for CIFAR-10 and CIFAR-100. We used the same hyperparameters for all the experiment settings, demonstrating the robustness of FGCM. For clothing1M, we randomly sampled 18,000 images for every class from the noisy dataset as our training set and used the clean test set to evaluate performance. To be consistent with previous methods [15], we used imagenet pretrained resnet-50 [36] as the backbone. We set the initial learning rate as 0.01 and reduced it by a factor of 10 after every 10 epochs. The warmup period for clothing1M was 10 epochs. An inception-resnet v2 [37] is employed for webvision. We set the initial learning rate as 0.01, and reduce it by a factor of 10 after 40,60 epochs. The warmup period is 60 epochs for webvision.

For all the experiments, the unlabeled threshold τ, unlabeled data ratio μ, loss weight of mixed batches λ_m are set to 0.95, 7 and 1, respectively.

4.2. Experimental Results and Analysis

Compared Methods

In this section, we present a thorough comparison between the performance of FGCM and several state-of-the-art methods on the four commonly used benchmark datasets. Table 2 presents the evaluation results of FGCM and other methods on CIFAR-10 and CIFAR-100 under different ratios of symmetric noise based on PreAct-18 resnet. Meanwhile, Table 3 shows the results with different ratios of asymmetric noise. Here we briefly introduce some of the SOTA methods we compared:

- **DivideMix.** DivideMix trains two networks simultaneously and iteratively divides samples into either clean or noisy sets via a two-component mixture model. Furthermore, semi-supervised learning is performed with the clean and noisy set treated as labeled and unlabeled data, respectively.

- **ELR.** Using semi-supervised learning techniques, ELR first estimates target probabilities base on the outputs of model in the early training stage. Then, ELR hinders the memorization of noisy labels by employing a regularization term to maximize the inner product between the targets and model outputs.
- **MOIT+.** Networks are first pretrained with supervised contrastive learning. Samples are divided based on the learned features. Then a classifier is trained in a semi-supervised manner on the divided datasets.
- **Sel-CL+.** The training of Sel-CL+ is warmed up with unsupervised contrastive learning. With the obtained low-dimensional representations, Sel-CL+ selects confident sample pairs and trains the models with supervised contrastive learning technique.

Table 2. Performance comparison with state-of-the-art methods on CIFAR-10 and CIFAR-100 with different ratios of symmetric noise as PreAct-18 resnet being the backbone.

Dataset	**CIFAR-10**			**CIFAR-100**		
Noise type	Sym					
Method/Noise ratio	20%	50%	80%	20%	50%	80%
Cross-Entropy	86.8	79.4	62.9	61.8	37.3	8.8
Fcorr [20]	86.8	79.8	63.3	61.5	46.6	19.9
Co-teaching+ [18]	89.5	85.7	67.4	65.6	51.8	27.9
Pcorr [38]	92.4	89.1	77.5	69.4	57.5	31.1
FINE [39]	91.0	87.3	69.4	70.3	64.2	25.6
Meta-Learning [40]	92.9	89.3	77.4	68.5	59.2	42.4
Mcorr [8]	94.0	92.0	86.8	73.9	66.1	48.2
DivideMix [19]	95.2	94.2	93.0	75.2	72.8	58.3
ELR [15]	93.8	92.6	88.0	74.5	70.2	45.2
MSLC [41]	93.5	90.5	69.9	72.5	68.9	24.3
MOIT+ [28]	94.1	91.8	81.1	75.9	70.6	47.6
Sel-CL+ [12]	95.5	93.9	89.2	76.5	72.4	59.6
FGCM	95.6	94.9	94.1	77.1	74.9	61.1

As stated in Table 2, Co-teaching+ gains a significant improvement from Cross-Entropy by excluding big-loss samples from training. However, the performance is limited compared with other methods due to discarding too many training samples. MOIT+ outperforms Co-teaching by a large margin. Leveraging supervised contrastive learning technique, MOIT+ learns a more robust representation and significantly alleviates the effect of noisy labels. Sel-CL+ gains further improvements by selecting confident sample pairs instead of clean samples. In this way, the learned representations benefit from not only the pairs with correct annotations, but also the pairs which are false-labeled from the same class. Therefore, the performance of Sel-CL+ is remarkably better than MOIT+ on high noise ratio settings (i.e., 80% symmetric noise on CIFAR-10 and CIFAR-100). DivideMix achieves comparable performance with Sel-CL+ on CIFAR-10 and CIFAR-100 with symmetric noise, owing to treating detected noisy samples as unlabeled data for semi-supervised training. However, the performance is not optimal under asymmetric noise as presented in Table 3, indicating that the small loss strategy does not work well for asymmetric noise. FGCM, our proposed methods, sets a new SOTA performance across all noise types and ratios. With the confidence of maximum repeating pseudo labels, FGCM can effectively perceive the difficulty experienced by DNN models when learning the samples, while with the confidence trends of annotation labels FGCM is able to decide whether a sample is clean. Combining

the two together, the proposed FGCM is able to mine more reusable supervision information, compared with DivideMix and Sel-CL+, from noisy datasets by categorizing training samples into fine-grained clusters. The improvement is significant especially for high noise ratios, which is more challenging. Sel-CL+ achieves competitive performance for low noise ratios (i.e., asymmetric noise and 20% symmetric noise) yet FGCM outperforms Sel-CL+ by a large margin for high noise ratios (e.g., 5.5% for 80% symmetric noise on CIFAR-10). This is owed to the ability of FGCM to accurately correct labels for noisy samples. While the performance of many sample selection methods degrades on the asymmetric noise, FGCM outperforms DivideMix by 2.4% for 40% asymmetric noise on CIFAR-10. For asymmetric noise, the training difficulties of different classes vary greatly due to the different degree of similarities between classes, causing the losses of samples in some annotated classes are generally higher than the other classes. Therefore, it is not convincing to realize the division of samples only based on the loss of their annotation labels. FGCM, however, is able to discriminate between hard samples and noisy samples by the fine-grained categorization, thus still achieving SOTA performance for asymmetric noise.

Notably, the results of original DivideMix and ELR are acquired by using the average predictions of two trained models while other methods use only one. For a fair comparison, we report their results without model ensembles.

Table 3. Performance comparison with state-of-the-art methods on CIFAR-10 with different ratios of asymmetric noise as PreAct-18 resnet being the backbone.

Dataset	**CIFAR-10**			
Noise type	Asym			
Method/Noise ratio	10%	20%	30%	40%
Cross-Entropy	88.8	86.1	81.7	76.0
GCE [24]	89.5	85.6	80.6	76.0
Pcorr [38]	93.1	92.9	92.6	91.6
Mcorr [8]	89.6	91.8	92.2	91.2
DivideMix [19]	93.8	93.2	92.5	91.4
ELR [15]	94.4	93.3	91.5	85.3
MOIT+ [28]	94.2	94.3	94.3	93.3
Sel-CL+ [12]	95.6	95.2	94.5	93.4
FGCM	95.6	95.4	94.7	93.8

Test results on clothing1M and WebVision are listed in Tables 4 and 5.

Table 4. Performance comparison with state-of-the-art methods on clothing1M.

Method	**TCNet [13]**	**Meta-Cleaner [42]**	**ELR**	**FINE**	**MSLC**	**Divide-Mix+**	**ELR+**	**FGCM**
Accuracy	71.15	72.50	72.87	72.91	74.02	74.76	74.81	74.91

The reported results of DivideMix+ and ELR+ are with model ensembles however FGCM still archives comparable or even better performance. The results demonstrate the effectiveness of our proposed FGCM.

Figure 3 illustrates some of mislabeled images in CLothing1M and WebVision filtered and rectified by FGCM. Images are randomly chosen from the NS cluster generated by FGCM. The red labels above each image are the annotation labels from corresponding datasets while the green label below each image is the pseudo label produced by FGCM.

Figure 3 demonstrates the effectiveness of FGCM in filtering and rectified noisy labels by modeling the confidence of samples in a fine-grained manner.

Table 5. Performance comparison with state-of-the-art methods on WebVision validation set and the ImageNet ILSVRC12 validation set. Results for baseline methods are copied from [12,15,19].

Method	WebVision		ILSVRC12	
	Top1	Top5	Top1	Top5
Fcorr [20]	61.15	82.68	57.36	82.36
Decoupling [43]	62.54	84.74	58.26	82.26
D2L [44]	62.68	84.00	57.80	81.36
MentorNet [16]	63.00	81.40	57.80	79.92
Co-teaching [17]	63.58	85.20	61.48	84.70
Iterative-CV [26]	65.24	85.34	61.60	84.98
DivideMix+ [19]	77.32	91.64	75.20	90.84
ELR [15]	76.26	91.26	68.71	87.84
ELR+ [15]	77.78	91.68	70.29	89.76
ProtoMix [45]	76.3	91.5	73.3	91.2
FGCM	77.84	91.76	74.56	90.24

The outstanding performance on both synthetic and real-world datasets demonstrates that FGCM can effectively alleviate the bad memorization effect of DNNs and still maintain good performance for the trained model even when there is a high number of false-labeled training samples.

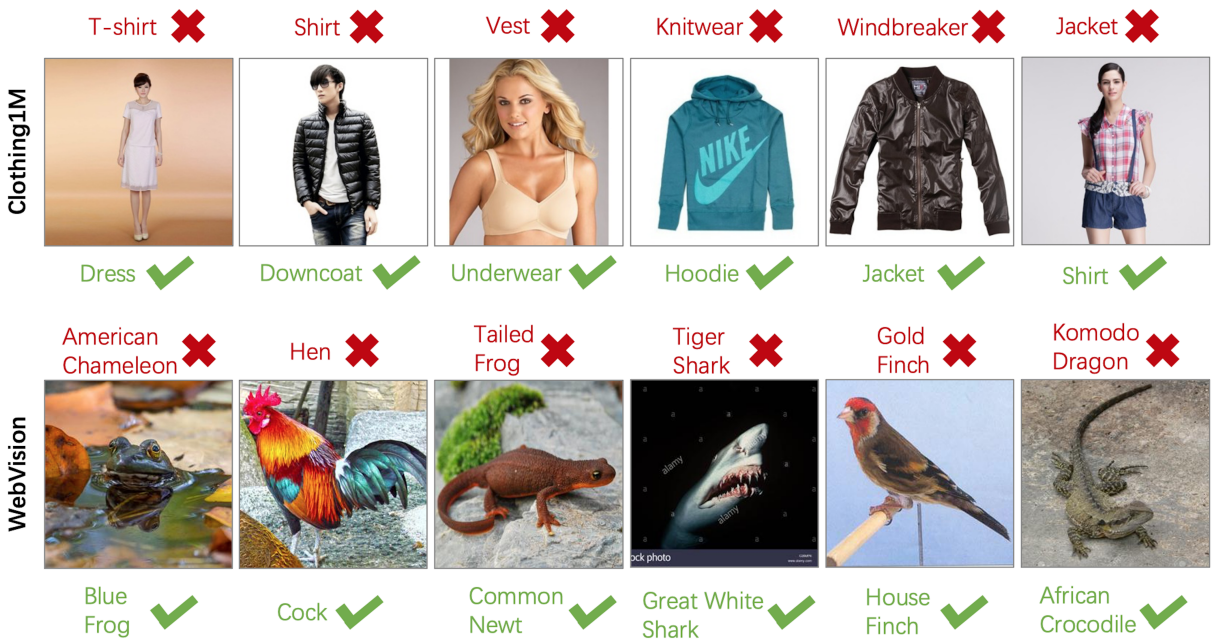

Figure 3. Examples of rectified labels by FGCM in Clothing1M and WebVision datasets. Red labels denote original annotations which are obviously wrong. Green labels denote rectified labels produced by FGCM which are obviously correct.

4.3. Ablations

To better evaluate our method, we perform multiple ablation studies in this section. As is shown in Table 6, to evaluate the effectiveness of cluster refinement, we train a model without using cluster refinement, the decrease in accuracy indicates that cluster refinement is beneficial to the performance of model. To study the impact of the noisy simple set, we

train a model without using rectified noisy simple set as labeled data. The performance further drops, which suggests that using samples from noisy simple set can improve the generation ability of the trained model. We train a model only using samples from the intersection of CS, NS and CH to validate the effect of mixed semi-training. The results prove the efficacy of mixed semi-training, especially when under high level of noise rate. We randomly sample 20,000 labels from the dataset to build labeled batches and use the rest labels as unlabeled data. There is a severe drop in the performance of the model trained without fine-grained sample categorization. In particular, the performance is totally ruined, under 80% noise rate. The results provide strong evidence to demonstrate the effectiveness of fine-grained sample categorization.

Table 6. Ablation Study results on CIFAR-10 with two different noise rates.

Dataset	**CIFAR-10**	
Method Noise Rates	**Sym 20%**	**Sym 80%**
FGCM (CS + NS + CH + NHI + NHR + cluster refinement)	95.6	94.1
FGCM (CS + NS + CH + NHI + NHR	95.2	93.6
FGCM (CS + CH + NHI + NHR)	94.5	93.0
FGCM (CS + NS + CH)	93.8	76.9
FGCM w/o fine-grained sample categorization	88.25	50.0
Cross Entropy	86.8	62.9

4.3.1. Ablation Study on Warm-Up Epoch

The performance of loss modeling methods in the literature is very sensitive to the choice of warmup epoch. Being too early or too late to end the warmup stage will cause either insufficient learning or over-fitting to noisy labels. As is shown in Figure 4, our method is relatively nonsensitive to the choice of warmup epoch. Under the setting of 20% and 50% noise rate, the precision of labels produced by FGCM keeps at a high point even when the model begins to be over-fitted. When the noise rate is extremely high (e.g., 80%) the precision of labels produced by FGCM slowly decreases.

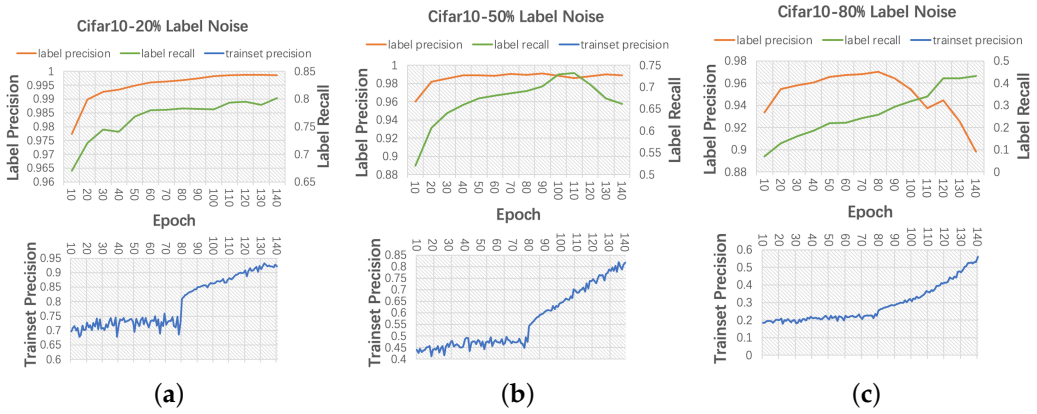

Figure 4. Precision and recall of labels produced by FGCM under different noise rates on CIFAR-10 dataset vs. the precision on trainset during training. The orange curves represent the precision of labels of the labeled trainset produced by FGCM. The green curves show the recall of the labeled trainset produced by FGCM. The blue curves illustrate the precision of models on the original trainset. (**a**) Results on CIFAR-10 with 20% symmetric label noise. (**b**) Results on CIFAR-10 with 50% symmetric label noise. (**c**) Results on CIFAR-10 with 80% symmetric label noise.

4.3.2. Ablation Study on Cluster Number

Tables 7 and 8 illustrate the precision and recall of filtered labels obtained using FGCM with different clustering numbers under different noise types and rates on CIFAR-10.

For ANNO-GMM-2, we cluster samples into two categories only using the confidence of annotation label, which is the common practice adopted by previous methods [9,19]. The set with higher confidence is chosen. For FGCM-GMM-3, we cluster samples into three categories and use the set with highest sum of confidence (i.e., CS) as well as the set with biggest difference between confidence of annotation and pseudo label (i.e., NS); For FGCM-GMM-4, we cluster samples into four categories, filtering CS, NS, and CH. The results show that by clustering samples into five categories, we can achieve the highest precision of filtered labels on all the settings while the recall is still acceptable. The comparison between ANNO-GMM-2 and FGCM-GMM-5 demonstrates the effectiveness of our proposed fine-grained confidence modeling, since FGCM can achieve higher precision than ANNO-GMM-2 while the recall is far higher than that of ANNO-GMM-2. Notably, though the label precision of ANNO-GMM-2 achieves 99.79% under 40% asymmetric noise, the clean set only consists of samples from five classes while all the samples from the other five classes are in the noisy set.

Table 7. Precision of obtained labels using different clustering methods and cluster numbers on CIFAR-10 with different type and ratios of noise. ANNO-GMM-2 denotes experiment that only uses the confidence of annotation label to divide samples into two categories.

Methods-Cluster Num	Sym 20%	Sym 50%	Sym 80%	Asym 40%
ANNO-GMM-2	99.70	97.94	96.73	99.79
FGCM-GMM-3	97.67	97.20	96.19	95.94
FGCM-GMM-4	99.10	97.42	95.40	95.40
FGCM-KMEANS-5	99.75	98.33	95.93	95.13
FGCM-GMM-5	99.75	99.02	97.23	97.04

Table 8. Recall of obtained labels using different clustering methods and cluster numbers on CIFAR-10 with different type and ratios of noise. ANNO-GMM-2 denotes experiment that only uses the confidence of annotation label to divide samples into two categories.

Methods-Cluster Num	Sym 20%	Sym 50%	Sym 80%	Asym 40%
ANNO-GMM-2	77.05	52.09	16.44	41.39
FGCM-GMM-3	79.68	73.28	35.84	58.72
FGCM-GMM-4	83.80	71.04	32.57	63.20
FGCM-KMEANS-5	77.56	66.40	33.78	60.83
FGCM-GMM-5	79.35	71.42	29.57	60.59

4.3.3. Ablation Study on Clustering Methods

The last two rows in Tables 7 and 8 perform a comparison between different clustering methods in FGCM. In FGCM-KMEANS-5, we use Kmeans, one of the most popular clustering algorithms, to cluster samples into five clusters based on the joint confidence scores with the maximum number of iterations and relative tolerance set to 2000 and 0.0001, respectively.

As stated in Tables 7 and 8, GMM outperforms Kmeans. The accuracy of labels produced by GMM is higher than Kmeans especially under high noise ratios while the recall of labels is comparable. GMM is more flexible in the sharpness of distribution and more robust to outliers compared with Kmeans. Therefore, we employ GMM in this paper.

4.4. Training Time Analysis

In this section, we compare the training time of FGCM on CIFAR-10 with 50% symmetrical noise with previous methods. The results are listed in Table 9. We use a single Nvidia Tesla V100 GPU to train all the models. Co-teaching+ is the fastest but the trained performance of other methods is much higher than Co-teaching+. The training time of

FGCM is slightly longer than DivideMix yet both of FGCM and DivideMix are much faster than Sel-CL+.

Table 9. Comparison of total Training Time on CIFAR-10 with 50% symmetrical label noise, using a single Nvidia Telsa V100 GPU.

Co-Teaching+	Pcorr	Meta-Learning	DivideMix	Sel-CL+	FGCM
4.3 h	6.0 h	8.6 h	5.2 h	7.2 h	5.4 h

5. Conclusions and Future Work

In this paper, in order to train deep neural networks robustly on noisily labeled datasets by recalling more reusable training samples, we propose to categorize training samples into five kinds (i.e., Clean Simple (CS), Noisy Simple (NS), Clean Hard (CH), Noisy Hard with Relevant annotations (NHR), and Noisy Hard with Irrelevant lables (NHI)). Then, we find there are significant differences among the confidence trends of samples from the five kinds during the training process. Thus, we propose a simple yet effective framework for noisy label learning called Fine-Grained Confidence Modeling (FGCM) where samples in the training set are categorized into fine-grained categories based on the difficulty experienced by DNN models when learning the samples and label correctness.

In every epoch after a warmup process, FGCM clusters training samples into the five fine-grained categories by fitting a five-component Gaussian Mixture Model on the prediction confidence of both annotation labels and the maximum repeating pseudo labels. Then, the generated clusters are refined using their prior assumptions (e.g., annotation labels and the maximum repeating pseudo labels of NS samples should be different). Finally, a new training set with high label accuracy can be obtained by screening out instances from NHR and NHI whose labels are ambiguous, rectifying the labels of NS samples with their maximum repeating pseudo labels. By iteratively repeating the process, FGCM can gradually recall more reusable training samples from NHR and NHI. Moreover, semi-training is employed with samples from NHR and NHI treated as unlabeled data. In this way, FGCM can maximize the use of supervised labels and training samples, meanwhile preventing the model from over-fitting to noisy labels. Extensive experiments on CIFAR-10, CIFAR-100, clothing1M and Webvision demonstrate the effectiveness of FGCM.

Limitations Despite FGCM significantly outperforming previous methods, the performance will downgrade on noisy datasets with class imbalance issues, which is common in real-world datasets. The confidence of training samples will be influenced by the long-tailed distribution between different classes, making it more difficult to distinguish clean samples. Furthermore, this is one of the reasons why current noisy label learning methods struggle to achieve better performance on real-world datasets such as Clothing1M and Webvision.

Future Works Our future works will focus on three aspects: (1) The robust training of DNNs on noisily labeled imbalance datasets. (2) The training time costed by existing methods is still high especially on large scale datasets. More efficient algorithms should be developed. (3) Currently, most works on noisy label learning are focused on the task of image classification. In future, techniques of noisy label learning may extend to other visual tasks [46] or other modalities such as natural language processing [47].

Author Contributions: Conceptualization, S.Y. and X.T.; methodology, S.Y.; validation, S.Y., X.T. and R.J.; formal analysis, S.Y. and R.J.; writing—original draft preparation, S.Y.; writing—review and editing, X.T., R.J. and Y.C.; supervision, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Foundation of China under Grant 31627802 and the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this work included, i.e., CIFAR-10, CIFAR-100, Clothing1M and WebVision image sets, are openly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M.S. A closer look at memorization in deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 7–9 August 2017.
2. Toneva, M.; Sordoni, A.; des Combes, R.T.; Trischler, A.; Bengio, Y.; Gordon, G. An empirical study of example forgetting during deep neural network learning. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
3. Zhang, C.; Recht, B.; Bengio, S.; Hardt, M.; Vinyals, O. Understanding deep learning requires rethinking generalization. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *CACM* **2017**, *6*, 84–90. [CrossRef]
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016.
6. Tanaka, D.; Ikami, D.; Yamasaki, T.; Aizawa, K. Joint optimization framework for learning with noisy labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
7. Liu, T.; Tao, D. Classification with Noisy Labels by Importance Reweighting. *IEEE Trans. PAMI* **2016**, *38*, 447–461. [CrossRef] [PubMed]
8. Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.E.; McGuinness, K. Unsupervised label noise modeling and loss correction. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
9. Zhang, Z.; Zhang, H.; Arik, S.O.; Lee, H.; Pfister, T. Distilling effective supervision from severe label noise. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020.
10. Gu, N.; Fan, M.; Meng, D. Robust Semi-Supervised Classification for Noisy Labels Based on Self-Paced Learning. *IEEE SPL* **2016**, *23*, 1806–1810. [CrossRef]
11. Yao, J.; Wang, J.; Tsang, I.; Zhang, Y.; Sun, J.; Zhang, C.; Zhang, R. Deep Learning from Noisy Image Labels with Quality Embedding. *IEEE Trans. IP* **2019**, *28*, 1909–1922. [CrossRef] [PubMed]
12. Li, S.; Xia, X.; Ge, S.; Liu, T. Selective-Supervised Contrastive Learning with Noisy Labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022.
13. Yi, R.; Huang, Y. TC-Net: Detecting Noisy Labels via Transform Consistency. *IEEE Trans. Multimed.* **2021**, *24*, 4328–4341. [CrossRef]
14. Bernhardt, M.; Castro, D.C.; Tanno, R.; Schwaighofer, A.; Tezcan, K.C.; Monteiro, M.; Bannur, S.; Lungren, M.P.; Nori, A.; Glocker, B.; et al. Active label cleaning for improved dataset quality under resource constraints. *Nat. Commun.* **2022**, *13*, 1161. [CrossRef] [PubMed]
15. Liu, S.; Niles-Weed, J.; Razavian, N.; Fernandez-Granda, C. Early-learning regularization prevents memorization of noisy labels. In Proceedings of the Conference on Neural Information Processing Systems, Virtual, 6–12 December 2020.
16. Jiang, L.; Zhou, Z.; Leung, T.; Li, L.; Li, F.-F. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
17. Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Proceedings of the Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018.
18. Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.W.; Sugiyama, M. How does disagreement help generalization against label corruption? In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
19. Li, J.; Socher, R.; Hoi, S. Dividemix: Learning with noisy labels as semi-supervised learning. In Proceedings of the International Conference on Learning Representations, Virtual, 26–30 April 2020.
20. Patrini, G.; Rozza, A.; Menon, A.K.; Nock, R.; Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017.
21. Chen, X.; Gupta, A. Webly supervised learning of convolutional networks. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
22. Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; Wang, X. Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
23. Ghosh, A.; Kumar, H.; Sastry, P. Robust loss functions under label noise for deep neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
24. Zhang, Z.; Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of the Conference on Neural Information Processing Systems, Montreal, QC, Canada, 2–8 December 2018.
25. Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–31 October 2019.

26. Chen, P.; Liao, B.B.; Chen, G.; Zhang, S. Understanding and utilizing deep neural networks trained with noisy labels. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
27. Nguyen, D.T.; Mummadi, C.K.; Ngo, T.P.N.; Nguyen, T.H.P.; Beggel, L.; Brox, T. SELF: Learning to filter noisy labels with self-ensembling. In Proceedings of the International Conference on Learning Representations, Virtual, 26–30 April 2020.
28. Ortego, D.; Arazo, E.; Albert, P.; O'Connor, N.E. Multi-objective interpolation training for robustness to label noise. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 20–25 June 2021.
29. Permuter, H.; Francos, J.; Jermyn, I. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognit.* **2006**, *39*, 695–706. [CrossRef]
30. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *R. Stat. Soc.* **1977**, *39*, 1–38.
31. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In Proceedings of the Conference on Neural Information Processing Systems, Virtual, 6–12 December 2020.
32. Krizhevsky, A.; Nair, V.; Hinton, G. *CIFAR-10 CIFAR-100*; Canadian Institute for Advanced Research: Toronto, ON, Canada, 2021.
33. Li, W.; Wang, L.; Li, W.; Agustsson, E.; Gool, L.V. Webvision database: Visual learning and understanding from web data. *arXiv* **2017**, arXiv:1708.02862.
34. Song, H.; Kim, M.; Park, D.; Lee, J. Prestopping: How does early stopping help generalization against label noise? In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016.
37. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
38. Yi, K.; Wu, J. Probabilistic end-to-end noise correction for learning with noisy labels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
39. Kim, T.; Ko, J.; Choi, J.; Yun, S.Y. Fine samples for learning with noisy labels. In Proceedings of the Conference on Neural Information Processing Systems, Virtual, 7–10 December 2021.
40. Li, J.; Wong, Y.; Zhao, Q.; Kankanhalli, M.S. Learning to learn from noisy labeled data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
41. Wu, Y.; Shu, J.; Xie, Q.; Zhao, Q.; Meng, D. Learn To Purify Noisy Labels via Meta Soft Label Corrector. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021.
42. Zhang, W.; Wang, Y.; Qiao, Y. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
43. Malach, E.; Shalev-Shwartz, S. Decoupling "when to update" from "how to update". In Proceedings of the Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–7 December 2017.
44. Ma, X.; Wang, Y.; Houle, M.; Zhou, S.; Erfani, S.; Xia, S.; Wijewickrema, S.; Bailey, J. Dimensionality-driven learning with noisy labels. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
45. Li, J.; Xiong, C.; Hoi, S. Learning from noisy data with robust representation learning. In Proceedings of the International Conference on Computer Vision, Virtual, 11–17 October 2021.
46. Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; Peng, X. Learning with Twin Noisy Labels for Visible-Infrared Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022.
47. Wang, Y.; Baldwin, T.; Verspoor, K. Noisy Label Regularisation for Textual Regression. In Proceedings of the International Conference on Computational Linguistics, Gyeongju, Korea, 10–12 October 2022.