

Selective-Supervised Contrastive Learning with Noisy Labels

Shikun Li^{1,2}, Xiaobo Xia³, Shiming Ge^{1,2*}, Tongliang Liu³

¹ Institute of Information Engineering, Chinese Academy of Sciences, China

² School of Cyber Security, University of Chinese Academy of Sciences, China

³ Trustworthy Machine Learning Lab, The University of Sydney, Australia

f lishikun, geshiming g@iie.ac.cn, xxia5420@uni.sydney.edu.au, tongliang.liu@sydney.edu.au

Abstract

Deep networks have strong capacities of embedding data into latent representations and finishing following tasks. However, the capacities largely come from high-quality annotated labels, which are expensive to collect. Noisy labels are more affordable, but result in corrupted representations, leading to poor generalization performance. To learn robust representations and handle noisy labels, we propose selective-supervised contrastive learning (Sel-CL) in this paper. Specifically, Sel-CL extends supervised contrastive learning (Sup-CL), which is powerful in representation learning, but is degraded when there are noisy labels. Sel-CL tackles the direct cause of the problem of Sup-CL. That is, as Sup-CL works in a pair-wise manner, noisy pairs built by noisy labels mislead representation learning. To alleviate the issue, we select confident pairs out of noisy ones for Sup-CL without knowing noise rates. In the selection process, by measuring the agreement between learned representations and given labels, we first identify confident examples that are exploited to build confident pairs. Then, the representation similarity distribution in the built confident pairs is exploited to identify more confident pairs out of noisy pairs. All obtained confident pairs are finally used for Sup-CL to enhance representations. Experiments on multiple noisy datasets demonstrate the robustness of the learned representations by our method, following the state-of-the-art performance. Source codes are available at <https://github.com/ShikunLi/Sel-CL>.

1. Introduction

Deep networks are powerful in various tasks, e.g., image recognition [19, 62], object detection [58], visual tracking [11] and text matching [3]. The power is largely attributed to the collection of large-scale datasets with high-quality annotated labels. In supervised learning, with the

Figure 1. Left: learning a classifier with ideal representations induced by clean labels. Right: learning a classifier with corrupted representations caused by noisy labels. Circles represent the representations of positive examples while triangles represent the representations of negative examples. When the representations are corrupted by noisy labels, the decision boundary of the classifier will be largely changed. Therefore, the learned classifier in this case cannot generalize well on test examples.

data (i.e., the instance and label pairs) in such datasets, deep networks first learn ideal latent representations of the instances and then complete following tasks with the representations [14, 57]. However, it is extremely expensive to obtain large-scale high-quality annotated labels. Alternatively, we can collect labels based on web search and user tags [36, 55]. These labels are cheap but inevitably noisy.

Noisy labels impair the generalization performance of deep networks [15, 60]. It is because, supervised by the datasets with noisy labels, the mislabeled data provide correct signals when inducing latent representations for the instances. The corrupted representations then cause inaccurate decisions for following tasks and hurt generalization [29, 52]. For example, as shown in Fig. 1, the corrupted representations result in an imprecise classification boundary. Therefore, it is crucial to induce robust latent representations of instances for learning with noisy labels, which is also our focus in this paper.

Recent works [7, 8, 13, 18, 65] show that, working in a pair-wise manner, contrastive learning (CL) methods can bring good latent representations to help following tasks. Based on whether supervised information is provided, the

* Shiming Ge is the corresponding author.

Figure 2. The illustration of the proposed Sel-CL, which progressively selects better con dent pairs for supervised contrastive learning based on the representation similarity. Without the noise rate prior, con dent examples are also obtained to help identify the pairs.

be the indicator of the event. Let $[z] = [z_1; \dots; z_g]$. Consider a classification task, there are C classes. We are given a noisily-labeled dataset $\mathcal{D} = \{f(x_i; y_i)\}_{i=1}^n$, where n is the sample size, x_i is the instance of the i -th example and $y_i \in [C]$ is the corresponding noisy label. For the associated but unobservable true label is denoted by y_i^* .

Overview. Following the two-stage learning style [13, 32, 41, 65], our aim is to induce robust pre-trained representations of instances for learning with noisy labels.

To achieve it, our approach progressively selects better con dent pairs G out of noisy pairs for performing supervised contrastive learning at each epoch. Without the noise rate prior, to help identify the pairs, con dent examples are also obtained in this process. The used pre-training network consists of three components: (1) a deep encoder (a convolutional neural network backbone) that maps the instance x_i to a high-dimensional representation z_i ; (2) a classifier head (a fully-connected layer followed by the softmax function) that receives z_i as an input and outputs class predictions $\hat{y}_i(x_i)$; (3) a linear or non-linear projection that maps z_i into a low-dimensional representation v_i . The illustration of the proposed Sel-CL can be seen in Fig. 2. After that, with the robust representations, we only keep the pre-trained deep encoder, and apply one new classifier head on the top to output the predictions for fine-tuning stage. In the following, we present our method step by step.

3.1. Selecting Con dent Examples

To help identify the con dent pairs, we first select con dent examples base on the representation similarity. In addition, we warm up the training in the first few epochs to ob-

tain low-dimensional representations of instances for identifying con dent examples later. Specifically, we use Unsupervised Contrastive Learning (UnCL) [7] for the warm-up training. Note that our method is robust to the choice of warm-up methods (See Section 4.6).

The con dent examples are identified by measuring agreements between the obtained low-dimensional representations and given labels. For this goal, given two low-dimensional representations z_i and z_j , we first calculate the representation similarity between them by the cosine distance, i.e.,

$$d(z_i; z_j) = \frac{z_i \cdot z_j}{\|z_i\| \|z_j\|}. \quad (1)$$

Then, to quantify the agreement, as did in [41], for each example $(x_i; y_i)$, we aggregate the original label from its top- K neighbors based on the representation similarity to create a pseudo-label \hat{y}_i . That is to say, we count the original labels of its top- K neighbors ($K=250$ in the experiments) and correctly use the dominant class. In this way, we can better use of representation similarities to improve the detection of mislabeled examples [41]. We use the pseudo-labels to approximate clean class posterior probabilities,

$$q_c(x_i) = \frac{1}{K} \sum_{x_{k=1}^{N_i}} \mathbb{I}[\hat{y}_k = c]; c \in [C]; \quad (2)$$

where N_i denotes the neighbor set of closest instances to x_i according to the learned representation. Following [16], we exploit the cross-entropy loss to identify con dent examples. Denoted the set of con dent examples belonging to the c -th class as \mathcal{T}_c , we have

$$\mathcal{T}_c = \{f(x_i; y_i) \mid q_c(x_i) > \tau; i \in [n]; c \in [C]\}; \quad (3)$$

where ϵ_c is a threshold for the c -th class, which is dynamically defined to ensure a class-balanced set of identified confident examples. To achieve this goal, we use the fractile of per-class agreements between the corrected label \hat{y}_i and the original label y_i across all classes to determine how many examples should be selected for each class, i.e., $\sum_{i=1}^n I[\hat{y}_i = y_i] I[y_i = c]; c \in [C]$. Finally, we can get the confident example set including all classes, $T = \bigcup_{c=1}^C T_c$. This set is less noisy than original noisy datasets and therefore more reliable.

3.2. Selecting Confident Pairs

As it is hard to achieve a precise estimation of the noise rate of noisy pairs, we use identified confident examples to help select confident pairs. Specifically, we transform identified confident examples into a set of associated confident pairs. Denoted this set by G^0 , we have

$$G^0 = \{P_{ij} \mid y_i = y_j; (x_i; y_i); (x_j; y_j) \in T\}; \quad (4)$$

where P_{ij} is the pair built by the examples $(x_i; y_i)$ and $(x_j; y_j)$. As G^0 is built by T , it is reliable. It should be noted that, even though two examples in one pair have incorrect class labels, the similarity labels are still correct when two examples are misclassified to the same class [50]. Such pairs are valuable for representation learning. We thus identify more confident pairs from noisy pairs. We define whether two instances x_i and x_j belong to the same class as their similarity label, i.e., $s_{ij} = I[y_i = y_j]$. Note that similarity labels are extremely imbalanced, the positive ones are less than negative ones [50]. For convenience, we consider the noisy pairs with positive similarity labels. Denoted by the set of confident pairs identified from noisy positive pairs by G^{00} . We have

$$G^{00} = \{P_{ij} \mid s_{ij} = 1; d(z_i; z_j) > \epsilon; \quad (5)$$

where ϵ is a dynamic threshold to control the number of identified confident pairs; i and j are two indices sampled from all training data. To avoid the noise rate estimation of noisy positive pairs, we utilize the reliable information of to set ϵ . In more detail, the fractile of the representation similarities of the pairs in G^0 is used here. Finally, we can get the confident pair set $G = G^0 \cap G^{00}$.

3.3. Representation Learning with Selected Pairs

After selecting confident pairs, we can learn representations by utilizing them with supervised contrastive learning at each epoch. As selected confident pairs can be less noisy, this selective-supervised paradigm can enhance representations to handle noisy labels.

Contrastive learning. Following the original Sup-CL [27], we randomly sample n instances in each mini-batch to apply two random data augmentation operations to each in-

stance, thus generating two data views. The resulting training mini-batch data is $\{(x_i; y_i)g_{i=1}^{2N}\}$, where $i \in [2N]$ is the index of an arbitrary augmented instance. Given we perform supervised contrastive learning with selected confident positive pairs:

$$L = \sum_{i \in [2N]} L_i(z_i) = \sum_{i \in [2N]} \frac{1}{|G(i)|} \sum_{g \in G(i)} \log P \frac{\exp(z_i \cdot z_g)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a)}; \quad (6)$$

where $A(i)$ means the set of indices excluding i.e., $A(i) = [n] \setminus \{i\}$; $G(i) = \{g \mid g \in G, g \neq i\}$; $P_{i \in G^0} \in [0, 1]$ and g^0 are the original indices of x_i and x_g in \mathcal{D} , respectively. $\tau \in \mathbb{R}^+$ is a temperature parameter. Note that expect for the examples that are involved in selected confident pairs, for the other examples, we perform unsupervised contrastive learning [7] on them.

Additionally, to further make representation learning robust, following MOIT [41], a Mixup technique [61] is added into our framework. Mixup performs convex combination of pairs of examples as $\tilde{x}_i = \alpha x_a + (1 - \alpha)x_b$, where $\alpha \in [0, 1]$ $\text{Beta}(\beta_m; \beta_m)$ and x_i denotes the training example that combines two mini-batch examples x_a and x_b . A linear relation in the contrastive loss is imposed as

$$L_i^{\text{MIX}}(z_i) = L_a(z_i) + (1 - \alpha)L_b(z_i); \quad (7)$$

where L_a and L_b have the same form as L_i in Eq. 6. It should be noted that the selection of positive/negative pairs involves considering a unique label for each mixed example [41]. Nevertheless, the input example always contain two labels, where α determines the dominant one. We assign this dominant label to every example for positive/negative sampling.

Classification learning. A classification objective with confident examples is also employed to stabilize the convergence and achieve better representations. Given the confident examples of T , classification learning is conducted by using

$$L^{\text{CLS}} = \sum_{(x_i; y_i) \in T} L_i^{\text{cls}}(x_i) = \sum_{(x_i; y_i) \in T} -\log(\hat{p}(x_i; y_i)); \quad (8)$$

where x_i can also refer to the augmented image.

Moreover, inspired by recent methods [22, 50] that learn classifiers with similarity labels, we add a learning objective to directly learn from similarity labels using classifier predictions. Given a multiviewed mini-batch data $\{(x_i; y_i)g_{i=1}^{2N}\}$, the added similarity loss is:

$$L^{\text{SIM}} = \sum_{i \in [2N]} \sum_{j \in A(i)} -\log(\hat{p}(x_i) \hat{p}(x_j) I[P_{ij} \in G]); \quad (9)$$

Lastly, combining the above analyses, the total objective loss is:

$$\mathcal{L}^{\text{ALL}} = \mathcal{L}^{\text{MIX}} + c \mathcal{L}^{\text{CLS}} + s \mathcal{L}^{\text{SIM}}; \quad (10)$$

where c and s are loss weights, which we set $a_s = 1$; $s = 0.01$ in all experiments. Note that, by alternately identifying con dent pairs and learning robust representations at each epoch, it ful ls a positive cycle that better con dent pairs will result in better learned representations and better representations will identify better con dent pairs. The similar idea of the positive cycle is shared by [2].

3.4. Classification Fine-tuning

With the pre-trained robust representation before, we only keep the pre-trained deep encoder and apply one new classifier head on top to form a classifier network and output the predictions. The existing methods on handling noisy labels can be applied to re-tune the classifier. To avoid the complexity of the method, we re-tune the deep networks on identified con dent examples, with a simple robust loss function [41]. In order to better distinguish the two stages in our framework, we call the proposed method with re-tuning as Sel-CL+. Also, we empirically verify that our method is also applicable with complex re-tuning algorithms e.g., DivideMix [30] and ELR+ [37].

4. Experiments

In this section, we employ comprehensive experiments to verify the effectiveness of our method.

4.1. Datasets and Implementation Details

Simulated noisy datasets We validate our method on two simulated noisy datasets, CIFAR-10 [28] and CIFAR-100 [28]. Both CIFAR-10 and CIFAR-100 contain 50k training images and 10k test images of the size 32 × 32. Following previous works [30, 41, 46], we consider two settings of simulated noisy labels: symmetric and asymmetric noise. Symmetric noise is generated by randomly replacing the labels for a percentage of the training data with all possible labels. Asymmetric noise uses label flips to incorrect classes: “truck → automobile, bird → airplane, deer → horse, cat → dog” in CIFAR-10, whereas in CIFAR-100 label flips are done circularly within the super-classes.

For CIFAR-10/100 datasets, we use a PreAct ResNet-18 network, and train it using SGD with a momentum of 0.9, a weight decay of 10^{-4} , and a batch size of 128. The network is trained for 250 epochs and the warm-up training has 1 epoch. We set the initial learning rate as 0.1, and reduce it by a factor of 10 after 125 and 200 epochs. The re-tuning stage of Sel-CL+ has 70 epochs, where the learning rate is 0.001. We always use the Mixup parameter $\alpha = 1$, scalar temperature $\tau = 0.1$, and loss weights $c = 1$; $s = 0.01$. The noise detection fractiles on the synthetic datasets are

Table 1. Noise detection fractiles on simulated noisy CIFAR-10 and CIFAR-100.

Fractiles	CIFAR-10		CIFAR-100	
	Sym.	Asym.	Sym.	Asym.
	50%	50%	75%	25%
	25%	25%	35%	0%

Table 2. Weighted KNN evaluations (%) on CIFAR-100. The best results are in bold.

Methods	Clean	Symmetric		Asymmetric	
	0%	20%	80%	10%	40%
Uns-CL [7]	56.23	—	—	—	—
Sup-CL [27]	72.66	58.32	41.00	71.11	68.00
MOIT [41]	<u>77.48</u>	67.42	55.58	74.86	72.60
Sel-CL	77.94	75.36	62.49	76.77	72.71

shown in Tab. 1. We apply the strong data augmentations SimAug [7] in the pre-training stage, and the standard weak data augmentations in the re-tuning stage. To save computing resources, the MOCO trick [18] is used, and the size of queue is set to 30k.

Real-world noisy dataset. We validate our method on a real-world noisy dataset, i.e., WebVision [35]. WebVision contains 2.4 million images crawled from the web using the 1,000 concepts in ImageNet ILSVRC12. Following previous works [30, 41], we compare baseline methods on the first 50 classes of the Google image subset, called WebVision-50. For WebVision-50, we use a standard ResNet-18, and train it using SGD with a momentum of 0.9, a weight decay of 10^{-4} , and a batch size of 64. The network is trained for 130 epochs and the warm-up training has 5 epochs. We set the initial learning rate as 0.1, and reduce it by a factor of 10 after 80 and 105 epochs. The re-tuning stage has 50 epochs, where the learning rate is 0.001. We use the Mixup parameter $\alpha = 1$, scalar temperature $\tau = 0.1$, loss weights $c = 1$; $s = 0.01$. The noise detection fractiles are $c = 40\%$ and $s = 0\%$. The data augmentation techniques are similar to those of the CIFAR-10/100 datasets. The MOCO trick [18] also is used, and the size of queue is 60k.

For all baselines, we use similar configurations for the same datasets for a fair comparison. We acquire the results of some baselines based on published codes. The results with underlines mean that we obtain them based on published codes. Implementation details for baselines are provided in Appendix B. We also borrow other experimental results from the related works [30, 37, 41, 42, 51, 65].

4.2. Representation Learning Evaluations

We start by analyzing the behaviors of different contrastive learning methods in the presence of noisy labels. We show how the selective module for con dent pairs im-

Table 3. Comparison with state-of-the-art methods in the test accuracy (%) on CIFAR-10 and CIFAR-100. The best results are in

Dataset	CIFAR-10									CIFAR-100								
	Symmetric					Asymmetric				Symmetric					Asymmetric			
	20%	50%	80%	90%	10%	20%	30%	40%	10%	20%	50%	80%	90%	10%	20%	30%	40%	10%
Cross-Entropy	82.7	57.9	26.1	16.8	88.8	86.1	81.7	76.0	61.8	37.3	8.8	3.5	68.1	63.6	53.3	44.5		
Mixup [61]	92.3	77.6	46.7	43.9	93.3	88.0	83.3	77.7	66.0	46.6	17.6	8.1	72.4	65.1	57.6	48.1		
Forward [43]	83.1	59.4	26.2	18.8	90.4	86.7	81.9	76.7	61.4	37.3	9.0	3.4	68.7	63.2	54.4	45.3		
GCE [64]	86.6	81.9	54.6	21.2	89.5	85.6	80.6	76.0	59.2	47.8	15.8	7.2	68.0	58.6	51.4	42.9		
P-correction [59]	92.0	88.7	76.5	58.2	93.1	92.9	92.6	91.6	68.1	56.4	20.7	8.8	76.1	68.9	59.3	48.3		
M-correction [1]	93.8	91.9	86.6	68.7	89.6	91.8	92.2	91.2	73.4	65.4	47.6	20.5	67.1	64.5	58.6	47.4		
DivideMix [30]	95.0	93.7	92.4	74.2	93.8	93.2	92.5	91.4	74.8	72.1	57.6	29.2	69.5	69.2	68.3	51.0		
ELR [37]	93.8	92.6	88.0	63.3	94.4	93.3	91.5	85.3	74.5	70.2	45.2	20.5	75.8	74.8	73.6	70.0		
GCE (Uns-CL init.) [13]	90.0	89.3	73.9	36.5	91.1	87.3	82.2	78.1	68.1	53.3	22.1	8.9	70.2	60.2	52.6	44.1		
ELR (Uns-CL init.)	94.4	93.0	88.3	86.2	95.0	94.7	94.4	93.3	76.2	71.9	57.9	40.8	77.2	75.5	74.3	70.4		
MOIT+ [41]	94.1	91.8	81.1	74.7	94.2	94.3	94.3	93.3	75.9	70.6	47.6	41.8	77.4	76.4	75.1	74.0		
Sel-CL+	95.5	93.9	89.2	81.9	95.6	95.2	94.5	93.4	76.5	72.4	59.6	48.8	78.7	77.5	76.4	74.2		

(a)

(b)

(c)

(d)

Figure 3. The learning process of Sel-CL on CIFAR-100 with 20% symmetric noise. (a) the number of selected examples per epoch; (b) the number of selected positive pairs per epoch; (c) the label precision of selected examples and pairs per epoch; (d) weighted KNN evaluations of Sel-CL (%) per epoch.

pacts the learned representations. We evaluate the quality of representations using a weighted KNN (K=200) evaluation in unsupervised learning [24].

The results are shown in Tab. 2. As can be seen, when there is no noise, MOIT greatly improves the quality of learned representations. Benefiting from the application of the similarity loss, Sel-CL achieves 0.46% quality gains. Then, for noisy datasets, Sup-CL is seriously affected by noisy labels. Although MOIT improves its robustness through the Mixup technique and semi-supervised strategy, its performance is still unsatisfactory. For example, in the case with 80% symmetric noise, the weighted KNN accu-

racy of MOIT is 55.58%, which is even less than the result of Uns-CL, i.e., 56.23%. In contrast, our Sel-CL significantly alleviates the side effect of noisy labels, by selecting content pairs for supervised contrastive learning. We can see that Sel-CL consistently outperforms baselines across different cases.

We illustrate the numbers of the selected content examples and pairs, the label precision of selected content examples and pairs, and the quality of learned representations during training. The experiments are conducted with 20% symmetric noise. The results are provided in Fig. 3. It can be seen that the selection of content examples/pairs and representation learning form a positive cycle. Sel-CL can progressively achieve effective selection and improve learned representations.

4.3. Results on Simulated Noisy Datasets

We compare the proposed Sel-CL+ with multiple baselines. We report the averaged test accuracy over the last 10 epochs. All baselines use the PreAct ResNet-18 network. Note that for a fair comparison, we report the results of DivideMix [30] without model ensembles, and equip ELR [37] with Mixup and weight averaging techniques.

As shown in Tab. 3, we first observe that the methods pre-trained by Uns-CL [7] can outperform the original ones, e.g., GCE vs. GCE (Uns-CL init.) and ELR vs. ELR (Uns-CL init.). When the noise level is high, e.g., 80% and 90%, the improvement is clearer, which is consistent with the observations in related works [13, 65]. For MOIT+ which exploits Sup-CL by adding regularization, its performance is promising. For symmetric noise, MOIT+ achieves comparable results with other state-of-the-art algorithms, e.g., ELR and DivideMix. For asymmetric noise, Sel-CL+ consistently achieves

Table 4. Weighted KNN evaluation (%) of Sel-CL with different selected sets on CIFAR-10/100 with 40% asymmetric noise.

Sel-CL with different selected sets	CIFAR-10	CIFAR-100
All examples and all pairs	90.58	68.66
Con dent examples $\bar{\mathbf{T}}$ and pairs \mathbf{G}^0	90.64	70.25
Con dent examples $\bar{\mathbf{T}}$ and pairs \mathbf{G}^0 [G ⁰⁰	92.97	72.71
Clean examples and associated pairs	94.21	71.45
Clean examples and all pairs	94.76	73.43
Clean examples and clean pairs	95.52	76.56

the best performance over all baselines. The results verify the effectiveness of our method well.

4.4. Pair-wise Selection Analysis

As the results on simulated noisy datasets show that Sel-CL+ work better in the case of asymmetric noise, where label flips happens between semantically similar classes. It is because Sel-CL not only exploits con dent pairs associated with \mathbf{T} , but also extracts more con dent positive pairs based on the representation similarity. To detail this phenomenon, we conduct some analysis about the pair-wise selection in our approach under the asymmetric noise cases.

There are several observations provided in Tab. 4. The results show that, it is useful to employ the con dent examples and associated positive pairs. However, the improvement is limited compared with selecting clean examples and clean pairs. Our method further employs which is based on the similarity representations. By this procedure, we can make better use of the con dent pairs whose class labels are incorrect but similarity labels could be correct, which is much common in the asymmetric cases and real-world scenarios. Besides, an interesting phenomenon is that it is not bad for selecting clean examples and all pairs, since the noise rate of pairs is smaller. This phenomenon may explain why MOIT+ also has advantages with asymmetric noise, which uses all noisy pairs.

4.5. Results on the Real-World Noisy Dataset

We validate our method on the real-world dataset WebVision-50 [35], which contains noisily-labeled images collected from Flickr and Google. As shown in Tab. 5, Sel-CL+ achieves the best results on both top-1 and top-5 accuracy on the WebVision validation set and ImageNet ILSVRC12 validation set than the other state-of-the-art i.e., DivideMix [30] (be better with symmetric noise) and methods, including four recent methods that also utilize ELR+ [37] (be better with asymmetric noise). As shown in Tab. 8, using more advanced ne-tuning methods can further improve the performance. Both the representations obtained by Uns-CL [7] and Sel-CL can promote the robustness of two methods, and our Sel-CL brings better results.

4.6. Ablation Study and Discussions

In uence of each component. To study the impact of each component in our method, we use CIFAR-100 for evaluations. We report test accuracy/weighted KNN evaluations (%) for Sel-CL, and test accuracy (%) for Sel-CL+

Table 5. Accuracy (%) on the WebVision and ILSVRC2012 validation sets. The model is trained on WebVision-50. The best results are in bold.

Methods	WebVision		ILSVRC12	
	top-1	top-5	top-1	top-5
Forward [43]	61.12	82.68	57.36	82.36
Decoupling [40]	62.54	84.74	58.26	82.26
D2L [39]	62.68	84.00	57.80	81.36
MentorNet [26]	63.00	81.40	57.80	79.92
Co-teaching [16]	63.58	85.20	61.48	84.70
Iterative-CV [6]	65.24	85.34	61.60	84.98
DivideMix [30]	77.32	91.64	75.20	90.84
ELR [37]	76.26	91.26	68.71	87.84
ELR+ [37]	77.78	91.68	70.29	89.76
ELR (Uns-CL init.)	79.93	92.00	71.23	88.23
ProtoMix [31]	76.3	91.5	73.3	91.2
MoPro [32]	77.59	—	76.31	—
NGC [51]	79.16	91.84	74.44	91.04
Sel-CL+	79.96	92.64	76.84	93.04

Table 6. Ablation study for Sel-CL and Sel-CL+ on CIFAR-100. The best results are in bold.

Methods	Sym. 20%	Asym. 40%
Sel-CL w/o Mixup Data Aug.	70.3/70.6	64.2/66.2
Sel-CL w/o MOCO Trick	73.3/74.1	69.2/71.5
Sel-CL w/o Selection	67.2/68.9	49.9/68.7
Sel-CL w/o Class er Learning	— /69.9	— /70.2
Sel-CL w/o L ^{SIM}	74.5/74.9	71.8/72.5
Sel-CL	74.9/75.4	72.0/72.7
Sel-CL+ w/ Strong Data Aug.	74.5	72.7
Sel-CL+ w/o Retraining Cls.	76.4	73.4
Sel-CL+	76.5	74.2

in Tab. 6. It shows the contribution of each component into our method.

Discussions on warm-up methods. We use different warm-up methods, i.e., Uns-CL [7] and Sup-CL [27] for Sel-CL+. The experiments are conducted on CIFAR-10 and CIFAR-100 with symmetric and asymmetric noise. The results in Tab. 7 demonstrate that Sel-CL+ is robust to the choice of warm-up methods.

Discussions on ne-tuning methods. We further test two complex but advanced methods for ne-tuning stage, i.e., DivideMix [30] (be better with symmetric noise) and methods, including four recent methods that also utilize ELR+ [37] (be better with asymmetric noise). As shown in Tab. 8, using more advanced ne-tuning methods can further improve the performance. Both the representations obtained by Uns-CL [7] and Sel-CL can promote the robustness of two methods, and our Sel-CL brings better results.

Comparison with one-stage methods for handling noisy labels. We compare our method with two recent one-stage methods, which also employ contrastive learning to handle noisy labels. For a fair comparison, we use AugMix [21]

Table 7. Comparison with different warm-up methods in the test accuracy (%) of Sel-CL+.

Dataset	CIFAR-10			CIFAR-100		
	Sym.		Asym.	Sym.		Asym.
Noise rate	20%	90%	40%	20%	90%	40%
Uns-CL [7]	95.5	81.9	93.4	76.5	48.8	74.2
Sup-CL [27]	95.5	81.6	93.4	76.8	51.4	74.5

Table 8. Comparison with using different fine-tuning methods in the test accuracy (%). The best results are **bold**.

Dataset	CIFAR-10		CIFAR-100	
	Sym.	Asym.	Sym.	Asym.
Noise rate	20%	40%	20%	40%
DivideMix [30]	95.7	92.1	76.9	53.8
ELR+ [37]	94.6	93.0	77.5	72.2
DivideMix (Uns-CL init.) [65]	96.2	90.8	78.3	52.9
ELR+ (Uns-CL init.) [65]	94.8	94.3	77.7	72.3
DivideMix (Sel-CL init.)	96.3	91.6	78.7	55.2
ELR+ (Sel-CL init.)	95.2	94.6	77.7	72.9

Table 9. Comparison with one-stage methods in the test accuracy (%). ψ denotes fine-tuning using AugMix data augmentation. The best results are **bold**.

Dataset	CIFAR-10			CIFAR-100		
	Sym.		Asym.	Sym.		Asym.
Noise rate	20%	90%	40%	20%	90%	40%
ProtoMix [31]	95.8	75.0	91.9	79.1	29.3	48.8
NGC [51]	95.9	80.5	90.6	79.3	29.8	–
Sel-CL+	95.4	67.5	92.8	76.4	35.5	74.2
Sel-CL ψ	95.2	67.4	92.5	76.0	35.4	74.2

Table 10. Comparison with different example selection strategies in the test accuracy/label precision (%) of Sel-CL+.

Noise type	Sym.		Asym.	
Noise rate	20%	90%	20%	40%
w pseudo-labels	76.5/99.1	48.8/62.0	77.5/97.5	74.2/92.2
w/o pseudo-labels	76.5/99.1	46.2/55.4	76.8/96.6	69.4/83.9

as data augmentations of the pre-training stage. As shown in Tab. 9, our approach has advantages in the asymmetric noise cases. In addition, we also try to adopt AugMix augmentation in the fine-tuning stage and find that the weak augmentation is more suitable, which is consistent with the ablation study. The results may reflect the difference between representation learning and classification learning.

Role of pseudo-labels. Following [41], we use the representation similarity with the weighted KNN to create pseudo labels. In this way, we can make better use of representation similarities. To make it clear, we use CIFAR-100 for evaluations. We report test accuracy/label precision evaluations (%) for Sel-CL+ with and without pseudo-labels, and here label precision is the average over all epochs. As Tab. 10 shows, this practice results in better performance

due to improved label precision. Besides, we also can just use the pseudo labels to define confident examples rather than to estimate clean class posterior probabilities. By comparing them from empirical observations, our method brings a higher label precision, which is shown in Fig. 4.

Figure 4. Comparison of label precision. The experiments are conducted on CIFAR-100 with 20% asymmetric noise.

The ablation studies about the hyperparameters are provided in Appendix A, which show that our method is robust to the choices of the hyperparameters.

5. Limitations

Our work still has certain limitations, including: 1) This work exploits contrastive learning, and therefore, the performance of our approach relies on adequate data augmentation and large amounts of negative samples. A larger batch size or memory bank is needed, which places higher demands on the storage of computing devices. 2) The use of the KNN algorithm brings greater computational consumption. In this work, we have used some faster KNN algorithms (see source codes) to alleviate the above issue, which facilitates the application of our method to large-scale datasets.

6. Conclusion

This paper proposes selective-supervised contrastive learning (Sel-CL), a new method to handle noisy labels in training data by learning robust pre-trained representations. We make use of the pair-wise characteristic of contrastive learning to better enhance network robustness. Without the noise rate prior, the confident pairs are selected out of noisy pairs for supervised contrastive learning. We demonstrate the state-of-the-art performance of our method with extensive experiments on multiple noisy datasets. For future work, we are interested in extending our method to other tasks such as object detection and text matching.

Acknowledgements This work was partially supported by grants from the Beijing Natural Science Foundation (19L2040), National Key Research and Development Plan (2020AAA0140001), and National Natural Science Foundation of China (61772513). Shiming Ge was supported by the Youth Innovation Promotion Association, CAS. Xiaobo Xia was supported by ARC Projects DE-190101473. Tongliang Liu was partially supported by ARC Projects DP180103424, DE-190101473, and DP-220102121.

References

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, pages 312–321, 2019. 6
- [2] Yingbin Bai and Tongliang Liu. Me-momentum: Extracting hard confident examples from noisily labeled data. *ICCV*, 2021. 5
- [3] Haolan Chen, Fred X. Han, Di Niu, Dong Liu, Kunfeng Lai, Chenglin Wu, and Yu Xu. MIX: multi-channel information crossing for text matching. *IKDD*, pages 110–119, 2018. 1
- [4] Ling-Hao Chen, He Li, and Wenhao Yang. Anomman: Detect anomaly on multi-view attributed networks. *arXiv preprint arXiv:2201.02822*, 2022. 2
- [5] Pengfei Chen, Guangyong Chen, Junjie Ye, Pheng-Ann Heng, et al. Noise against noise: stochastic label noise helps combat inherent label noise. *ICLR*, 2021. 2
- [6] Pengfei Chen, Benben Liao, Guangyong Chen, et al. Understanding and utilizing deep neural networks trained with noisy labels. *ICML*, pages 1833–1841, 2019. 7
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, pages 1597–1607, 2020. 1, 2, 3, 4, 5, 6, 7, 8
- [8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, pages 22243–22255, 2020. 1, 2
- [9] De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. *CVPR*, 2022. 2
- [10] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. *ICLR*, 2021. 2
- [11] Shiming Ge, Chunhui Zhang, Shikun Li, Dan Zeng, and Dacheng Tao. Cascaded correlation refinement for robust deep tracking. *TNNLS* 32(3):1276–1288, 2021. 1
- [12] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, pages 1919–1925, 2017. 2
- [13] Aritra Ghosh and Andrew Lan. Contrastive Learning Improves Model Robustness Under Label Noise. *CVPR Workshop*, 2021. 1, 3, 6
- [14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1
- [15] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. *ICML*, pages 4006–4016, 2020. 1, 2
- [16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8536–8546, 2018. 2, 3, 7
- [17] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *ICCV*, pages 5138–5147, 2019. 2
- [18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, pages 9726–9735, 2020. 1, 2, 5
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016. 1
- [20] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *NeurIPS*, pages 10477–10486, 2018. 2
- [21] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *ICLR*, 2020. 7
- [22] Yen-chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification class labels without multi-class labels. *ICLR*, 2019. 4
- [23] Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. *ICLR*, 2020. 2
- [24] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *ICML*, pages 2849–2858, 2019. 6
- [25] Jinchuan Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In *ICCV*, pages 3326–3334, 2019. 2
- [26] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *ICML*, pages 2309–2318, 2018. 2, 7
- [27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Dilip Krishnan, and Ce Liu. Supervised contrastive learning. In *NeurIPS*, pages 1–23, 2020. 2, 4, 5, 7, 8
- [28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [29] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, et al. Robust inference via generative classifiers for handling noisy labels. In *ICML*, pages 3763–3772, 2019. 1
- [30] Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 2, 5, 6, 7, 8
- [31] Junnan Li, Caiming Xiong, and Steven Hoi. Learning from noisy data with robust representation learning. *ICCV*, 2021. 2, 7, 8
- [32] Junnan Li, Caiming Xiong, and Steven C.H. Hoi. MoPro: Webly supervised learning with momentum prototypes. In *ICLR*, 2021. 2, 3, 7
- [33] Shikun Li, Shiming Ge, Yingying Hua, Chunhui Zhang, Hao Wen, Tengfei Liu, and Weiqiang Wang. Coupled-view deep classifier learning from multiple noisy annotators. In *AAAI*, pages 4667–4674, 2020. 2

- [34] Shikun Li, Tongliang Liu, Jiyong Tan, Dan Zeng, and Shim-ing Ge. Trustable co-label learning from multiple noisy an-notators. *TMM*, 2021. [2](#)
- [35] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learn-ing and understanding from web data. *arXiv preprint arXiv:1708.02862* 2017. [5](#), [7](#)
- [36] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, pages 1928–1936, 2017. [1](#)
- [37] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Car-los Fernandez-Granda. Early-learning regularization pre-vents memorization of noisy labels. *NeurIPS* 2020. [2](#), [5](#), [6](#), [7](#), [8](#)
- [38] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *PAMI*, 38(3):447–461, 3 2016. [2](#)
- [39] Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah M. Erfani, Shu-Tao Xia, Sudanthi N. R. Wijewick-rema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, pages 3361–3370, 2018. [7](#)
- [40] Eran Malach and Shai Shalev-Shwartz. Decoupling” when to update” from” how to update”. In *NeurIPS* pages 960–970, 2017. [7](#)
- [41] Diego Ortego, Eric Arazo, Paul Albert, Noel E. O’Connor, et al. Multi-objective interpolation training for robustness to label noise. In *CVPR* 2021. [2](#), [3](#), [4](#), [5](#), [6](#), [8](#)
- [42] Diego Ortego, Eric Arazo, Paul Albert, Noel E. O’Connor, and Kevin McGuinness. Towards Robust Learning with Dif-ferent Label Noise Distributions. *ICPR*, pages 7020–7027, 2021. [5](#)
- [43] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. *CVPR*, pages 2233–2241, 2017. [6](#), [7](#)
- [44] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtas-un. Learning to reweight examples for robust deep learning. In *ICML*, pages 4331–4340, 2018. [2](#)
- [45] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *NeurIPS* pages 1917–1928, 2019. [2](#)
- [46] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiy-oharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pages 5552–5560, 2018. [2](#), [5](#)
- [47] Ruijia Wang, Shuai Mou, Xiao Wang, Wanpeng Xiao, Qi Ju, Chuan Shi, and Xing Xie. Graph structure estimation neural networks. In *WWW*, pages 342–353, 2021. [2](#)
- [48] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yan-dong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. *arXiv preprint arXiv:2111.15174* 2021. [2](#)
- [49] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Ro-bust long-tailed learning under label noise. *arXiv preprint arXiv:2108.11569* 2021. [2](#)
- [50] Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Ming-ming Gong, Nannan Wang, Haifeng Liu, and Gang Niu. Class2Simi: A New Perspective on Learning with Label Noise. In *ICML*, 2021. [4](#)
- [51] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. NGC: A uni ed framework for learning with open-world noisy data. *ICCV*, 2021. [2](#), [5](#), [7](#), [8](#)
- [52] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. *ICLR*, 2021. [1](#)
- [53] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Ming-ming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *NeurIPS* 2020. [2](#)
- [54] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *NeurIPS* pages 6835–6846, 2019. [2](#)
- [55] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015. [1](#)
- [56] Hansi Yang, Quanming Yao, Bo Han, Gang Niu, Hansi Yang, Bo Han, Gang Niu, and James Kwok. Searching to Exploit Memorization Effect in Learning from Corrupted Labels. In *ICML*, 2020. [2](#)
- [57] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *ICLR*, 2021. [1](#)
- [58] Shuo Yang, Peize Sun, Yi Jiang, Xiaobo Xia, Ruiheng Zhang, Zehuan Yuan, Changhu Wang, Ping Luo, and Min Xu. Objects in semantic topology. *ICLR*, 2022. [1](#)
- [59] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise cor-rection for learning with noisy labels. *CVPR*, pages 7017–7025, 2019. [6](#)
- [60] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning re-quires rethinking generalization. *ICLR*, 2017. [1](#)
- [61] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimiza-tion. In *ICLR*, 2018. [2](#), [4](#), [6](#)
- [62] Kangkai Zhang, Chunhui Zhang, Shikun Li, Dan Zeng, and Shiming Ge. Student network learning via evolutionary knowledge distillation. *TCSVT* 2021. [1](#)
- [63] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *ICLR*, 2021. [2](#)
- [64] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pages 8778–8788, 2018. [2](#), [6](#)
- [65] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M. Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *WACV*, 2022. [1](#), [3](#), [5](#), [6](#), [8](#)
- [66] Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *ICML*, pages 11447–11457, 2020. [2](#)