Selective-Supervised Contrastive Learning with Noisy Labels（CVPR_2022）

Shikun Li , Xiaobo Xia , Shiming Ge, Tongliang Liu
Institute of Information Engineering, Chinese Academy of Sciences, China
School of Cyber Security, University of Chinese Academy of Sciences, China
Trustworthy Machine Learning Lab, The University of Sydney, Australia
{lishikun,geshiming}@iie.ac.cn, xxia5420@uni.sydney.edu.au, tongliang.liu@sydney.edu.au

2022. 3. 29 • ChongQing

gesis Leibniz-Institut für Sozialwissenschaften

Forschungszentrum · Research Center L3S

Reported by Lele Duan

Code & dataset : https://github.com/ShikunLi/Sel-CL

1.Background

2.Method

3.Experiments

Background

- *Noisy labels are more affordable, but result in corrupted representations , leading to poor generalization performance.*

- *To learn robust representations and handle noisy labels.*

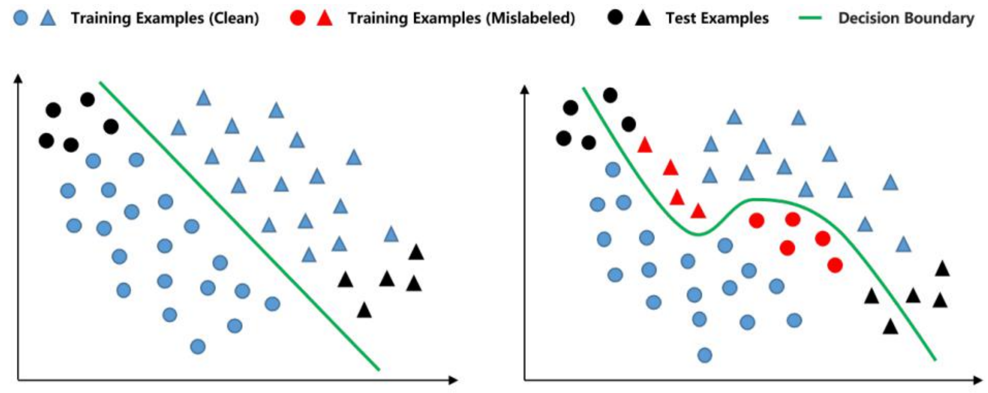

Figure 1. **Left**: learning a classifier with ideal representations induced by clean labels; **Right**: learning a classifier with corrupted representations caused by noisy labels. Circles represent the representations of positive examples while triangles represent the representations of negative examples. 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.

➢ Over view

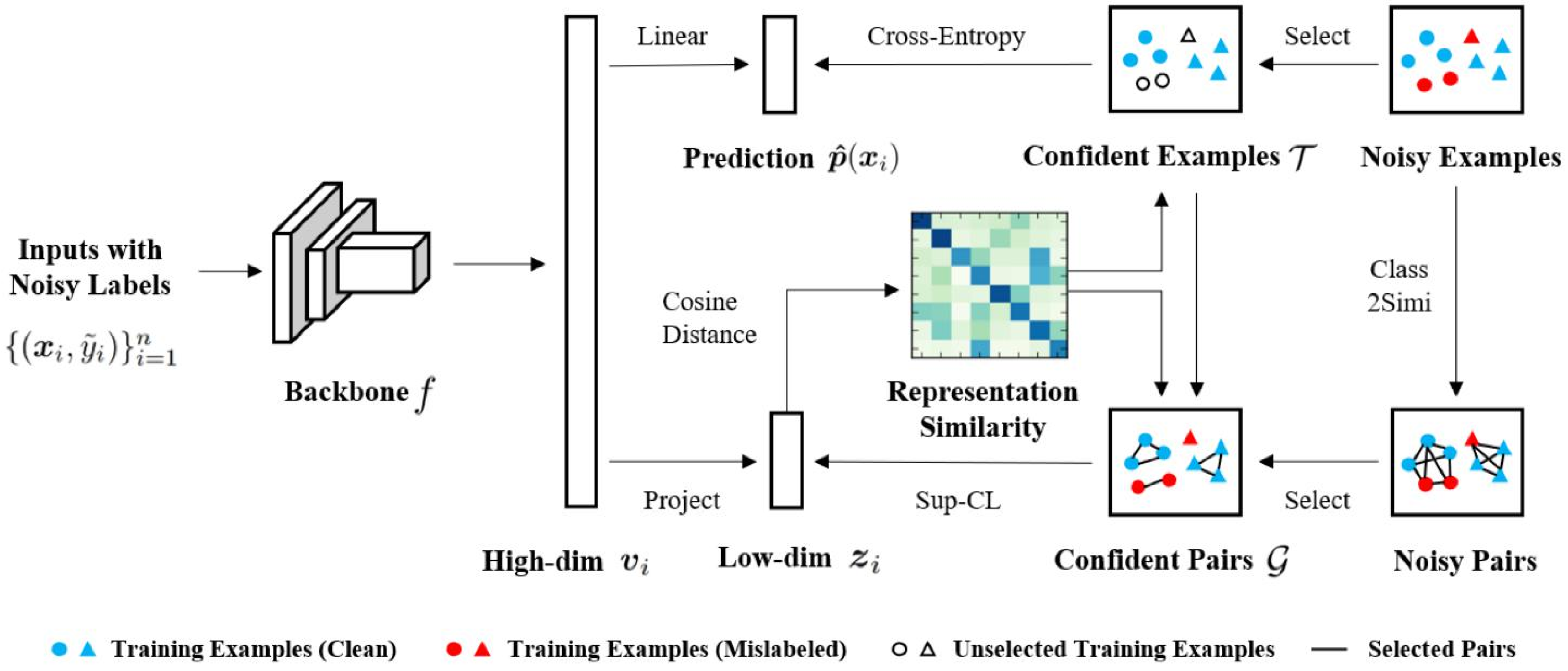

Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

Selecting Confident Examples

- Representation similarity:

$$d(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{z}_i \boldsymbol{z}_j^\top}{\|\boldsymbol{z}_i\| \, \|\boldsymbol{z}_j\|}. \quad (1)$$

- Pseudo-label:

$(\boldsymbol{x}_i, \tilde{y}_i)$ pseudo-label $\hat{y}_i$

$$\hat{q}_c(\boldsymbol{x}_i) = \frac{1}{K} \sum_{\substack{k=1 \\ \boldsymbol{x}_k \in \mathcal{N}_i}}^{K} \mathbb{I}[\hat{y}_k = c], c \in [C], \quad (2)$$

where $\mathcal{N}_i$ denotes the neighbor set of K closest instances to $\boldsymbol{x}_i$ according to the learned representation. Following [16], we exploit the cross-entropy loss ℓ to identify confident examples. Denoted the set of confident examples belonging to the c-th class as $\mathcal{T}_c$, we have

$$\mathcal{T}_c = \{(\boldsymbol{x}_i, \tilde{y}_i) \mid \ell(\hat{\boldsymbol{q}}(\boldsymbol{x}_i), \tilde{y}_i) < \gamma_c, i \in [n]\}, c \in [C], \quad (3)$$

$$\mathcal{T} = \cup_{c=1}^{C} \mathcal{T}_c$$

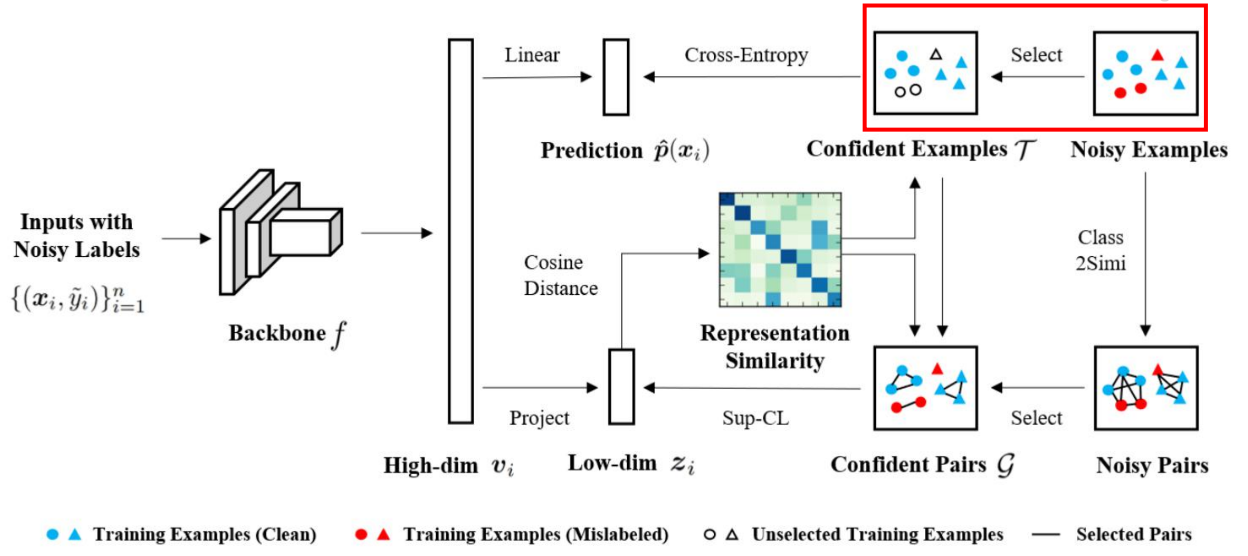

Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

where γ_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 α fractile of per-class agreements between the corrected label $\hat{y}_i$ and the original label $\tilde{y}_i$ across all classes to determine how many examples should be selected for each class, i.e., $\sum_{i=1}^{n} \mathbb{I}[\hat{y}_i = \tilde{y}_i][\tilde{y}_i = c], c \in [C]$. Finally, we can get the confident example set including all classes, *i.e.*, $\mathcal{T} = \cup_{c=1}^{C} \mathcal{T}_c$. This set is less noisy than original noisy datasets and therefore more reliable.

Selecting Confident Pairs

- From $\mathcal{T}$:

$$\mathcal{G}' = \{P_{ij} \mid \tilde{y}_i = \tilde{y}_j, (\boldsymbol{x}_i, \tilde{y}_i), (\boldsymbol{x}_j, \tilde{y}_j) \in \mathcal{T}\}, \qquad (4)$$

where P_{ij} is the pair built by the examples $(\boldsymbol{x}_i, \tilde{y}_i)$ and $(\boldsymbol{x}_j, \tilde{y}_j)$. As $\mathcal{G}'$ is built by $\mathcal{T}$, it is reliable. It should

- From noisy pairs:

$$\tilde{s}_{ij} = \mathbb{I}[\tilde{y}_i = \tilde{y}_j]$$

$$\mathcal{G}'' = \{P_{ij} \mid \tilde{s}_{ij} = 1, d(\boldsymbol{z}_i, \boldsymbol{z}_j) > \gamma\}, \qquad (5)$$

where γ 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 $\mathcal{T}$ to set γ. In more detail, the β fractile of the representation similarities of the pairs in $\mathcal{G}'$ is used here. Finally, we can get the confident pair set $\mathcal{G} = \mathcal{G}' \cup \mathcal{G}''$.

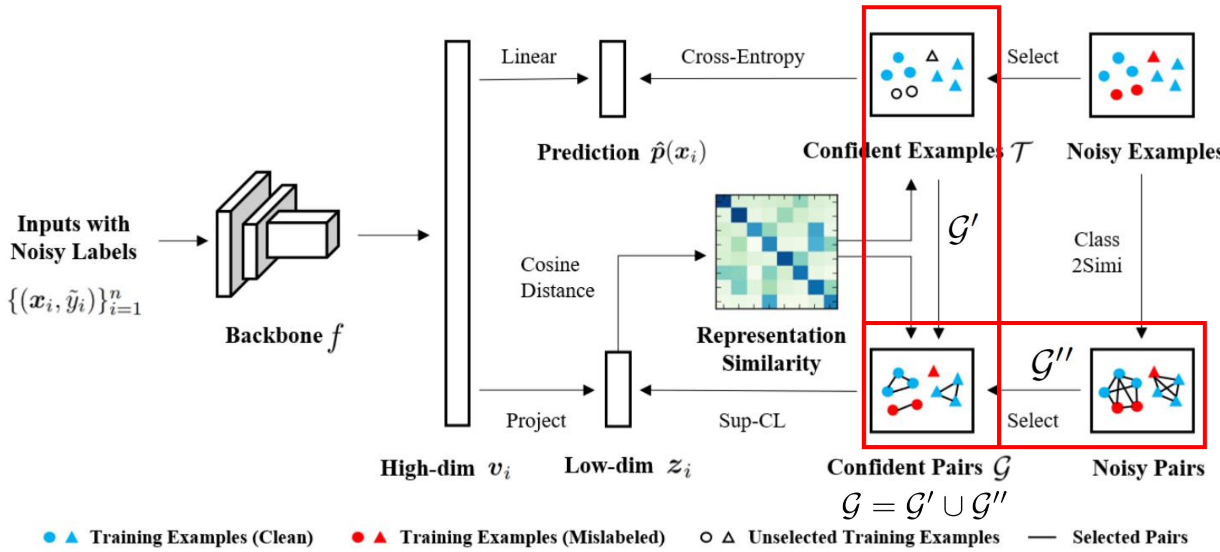

Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

Representation Learning with Selected Pairs

- *Contrastive learning:*

$$\mathcal{L} = \sum_{i \in I} \mathcal{L}_i(\boldsymbol{z}_i)$$
$$= \sum_{i \in I} \frac{-1}{|\mathcal{G}(i)|} \sum_{g \in \mathcal{G}(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_g / \tau)}{\sum_{a \in A(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)}, \tag{6}$$

where $A(i)$ means the set of indices excluding i, i.e., $A(i) = I \backslash \{i\}$; $\mathcal{G}(i) = \{g \mid g \in A(i), P_{i'g'} \in \mathcal{G}\}$, and i' and g' are the original indices of $\boldsymbol{x}_i$ and $\boldsymbol{x}_g$ in $\widetilde{\mathcal{D}}$, respectively. $\tau \in \mathbb{R}^+$ is a temperature parameter. Note that except for the examples that are involved in selected confident pairs, for the other examples, we perform unsupervised contrastive learning [7] on them.

$$\mathcal{L}_i^{\mathrm{MIX}}(\boldsymbol{z}_i) = \lambda \mathcal{L}_a(\boldsymbol{z}_i) + (1 - \lambda)\mathcal{L}_b(\boldsymbol{z}_i), \tag{7}$$

where $\mathcal{L}_a$ and $\mathcal{L}_b$ have the same form as $\mathcal{L}_i$ in Eq. 6. It

$$\lambda \in [0, 1] \sim \mathrm{Beta}(\alpha_m, \alpha_m)$$

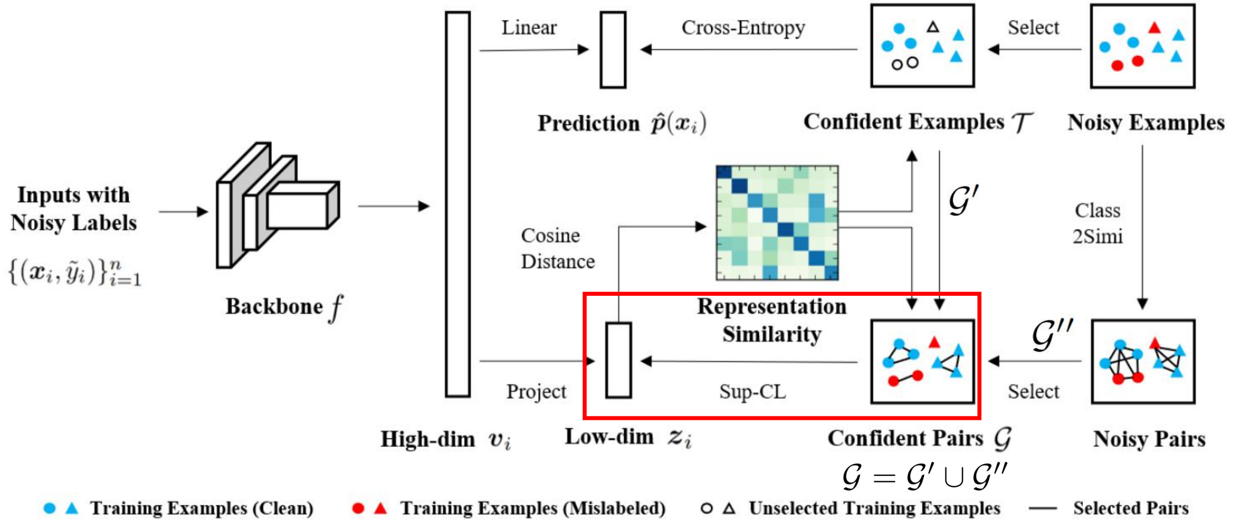

Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

Representation Learning with Selected Pairs

- Classification learning:

$$\mathcal{L}^{\mathrm{CLS}} = \sum_{(\boldsymbol{x}_i,\tilde{y}_i)\in\mathcal{T}} \mathcal{L}_i^{cls}(\boldsymbol{x}_i) = \sum_{(\boldsymbol{x}_i,\tilde{y}_i)\in\mathcal{T}} \ell(\hat{\boldsymbol{p}}(\boldsymbol{x}_i),\tilde{y}_i), \quad (8)$$

where $\boldsymbol{x}_i$ can also refer to the augmented image.

- Learn classifiers with similarity labels:

$$\mathcal{L}^{\mathrm{SIM}} = \sum_{i\in I}\sum_{j\in A(i)} \ell(\hat{\boldsymbol{p}}(\boldsymbol{x}_i)\hat{\boldsymbol{p}}(\boldsymbol{x}_j), \mathbb{I}[P_{i'j'}\in\mathcal{G}]). \quad (9)$$

$$\mathcal{L}^{\mathrm{ALL}} = \mathcal{L}^{\mathrm{MIX}} + \lambda_c\mathcal{L}^{\mathrm{CLS}} + \lambda_s\mathcal{L}^{\mathrm{SIM}}, \quad (10)$$

where λ_c and λ_s are loss weights, which we set as $\lambda_c = 1, \lambda_s = 0.01$ in all experiments. Note that, by alternately

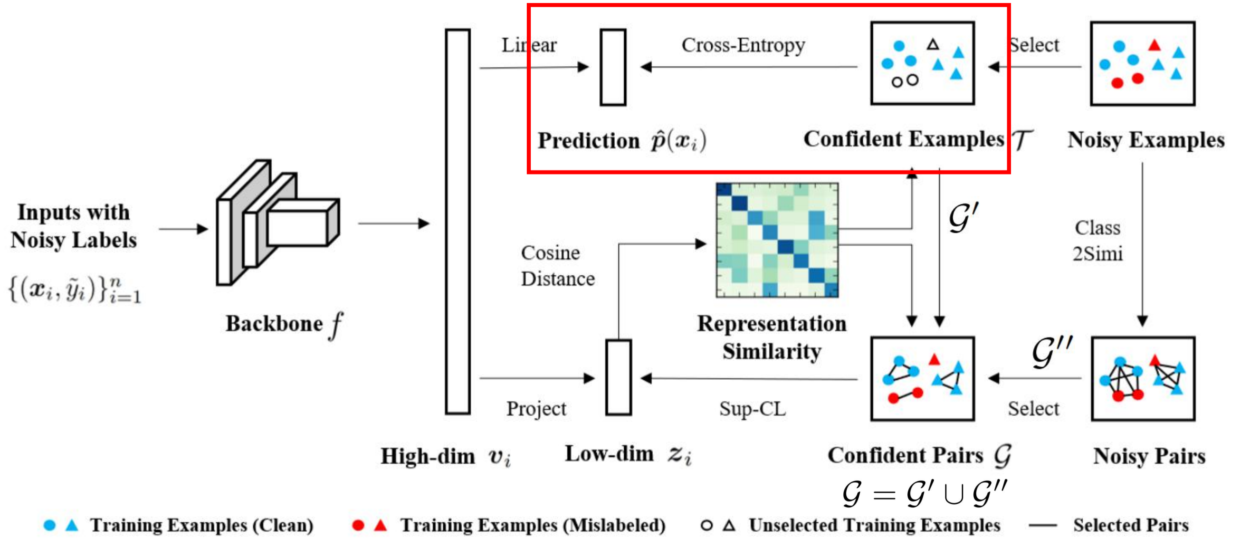

Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

Experiment

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

Fractiles	CIFAR-10		CIFAR-100	
	Sym.	Asym.	Sym.	Asym.
α	50%	50%	75%	25%
β	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]	77.48	67.42	55.58	74.86	72.60
Sel-CL	**77.94**	**75.36**	**62.49**	**76.77**	**72.71**

Experiment

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

Dataset	CIFAR-10								CIFAR-100							
Methods/Noise rate	Symmetric				Asymmetric				Symmetric				Asymmetric			
	20%	50%	80%	90%	10%	20%	30%	40%	20%	50%	80%	90%	10%	20%	30%	40%
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**

Experiment

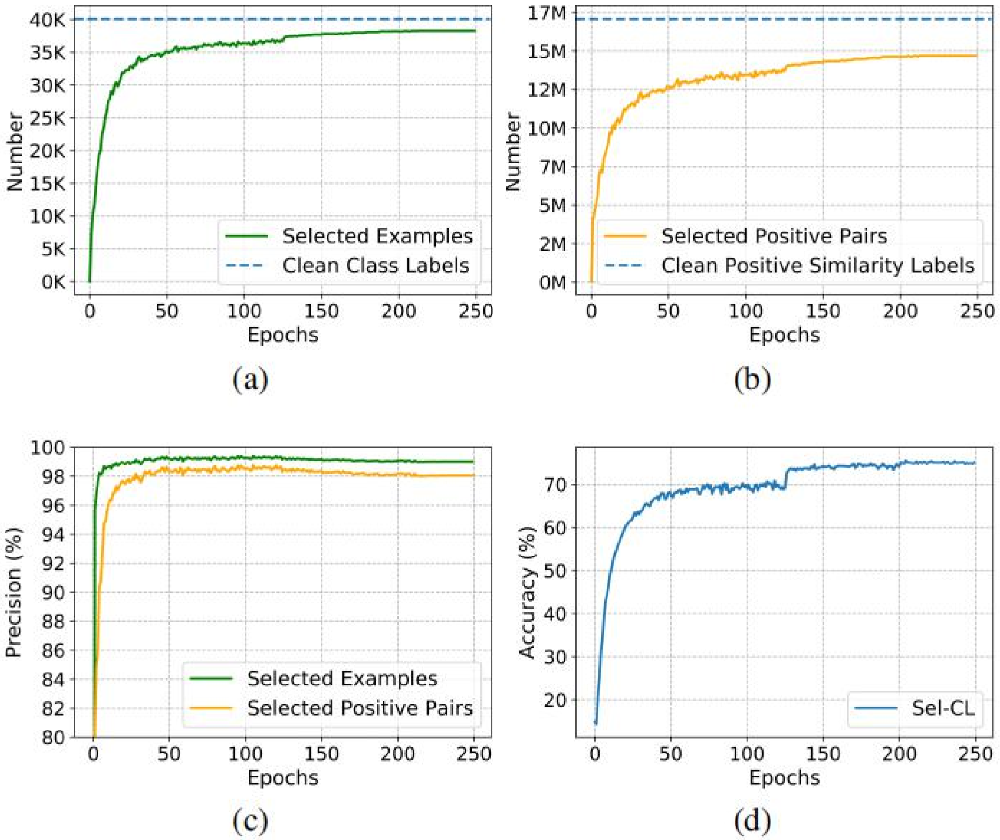

Figure 3. The learning process of Sel-CL on CIFAR-100 with 20% symmetric noise. (a) the number of selected examples *vs.* epochs; (b) the number of selected positive pairs *vs.* epochs; (c) the label precision of selected examples and pairs (%) *vs.* epochs; (d) weighted KNN evaluations of Sel-CL (%)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

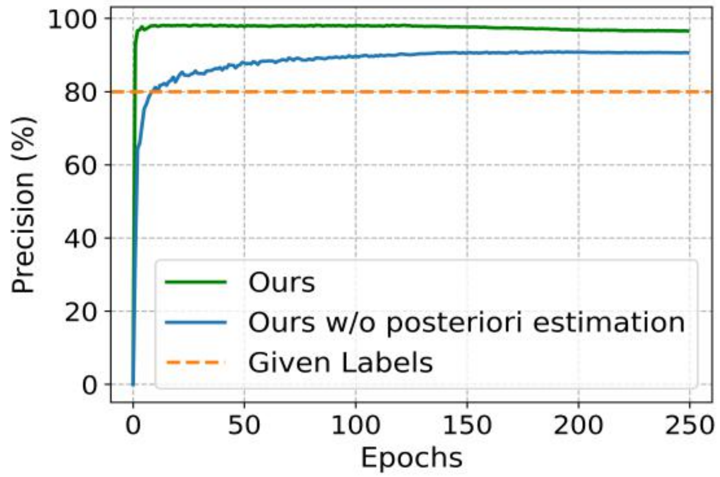

Figure 4. Comparison of label precision. The experiments are conducted on CIFAR-100 with 20% asymmetric noise.

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
Confident examples $\mathcal{T}$ and pairs $\mathcal{G}'$	90.64	70.25
Confident examples $\mathcal{T}$ and pairs $\mathcal{G}' \cup \mathcal{G}''$	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

Experiment

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 Classfier Learning	— /69.9	— /70.2
Sel-CL w/o $\mathcal{L}^{SIM}$	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**

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

Dataset	CIFAR-10			CIFAR-100		
Noise type	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 in **bold**.

Dataset	CIFAR-10		CIFAR-100	
Noise type	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. Comparesion with one-stage methods in the test accuracy (%). † denotes fine-tuning using AugMix data augmentation. The best results are in **bold**.

Dataset	CIFAR-10			CIFAR-100		
Noise type	Sym.		Asym.	Sym.		Asym.
Noise rate	20%	90%	40%	20%	90%	40%
ProtoMix [31]	95.8	75.0	91.9	79.1	29.3	<u>48.8</u>
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+†	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 precsion (%) 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

Thank you!