南京航空航天大學

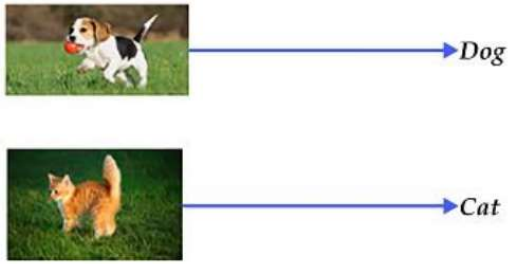
Nanjing University of Aeronautics and Astronautics

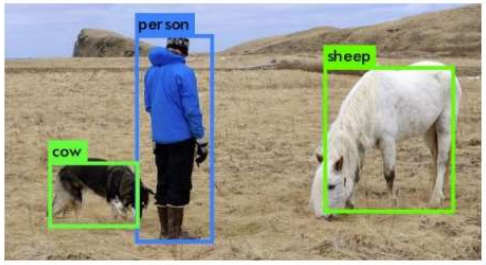
RankMatch: Fostering Confidence and Consistency in Learning with Noisy Labels

Ziyi Zhang[1,2], Weikai Chen[3], Chaowei Fang[4], Zhen Li[5], Lechao Chen[6], Liang Lin[2], Guanbin Li[2,7*]
[1]National Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China
[2]Sun Yat-sen University, Guangzhou, China, [3] Tencent America
[4] Xidian Univiersity, [5] The Chinese University of Hong Kong (Shenzhen), [6] Zhejiang Lab
[7] Research Institute, Sun Yat-sen University, Shenzhen, China
zhangziyi@lamda.nju.edu.cn, liguanbin@mail.sysu.edu.cn

ICCV 2023

Noisy Label Learning





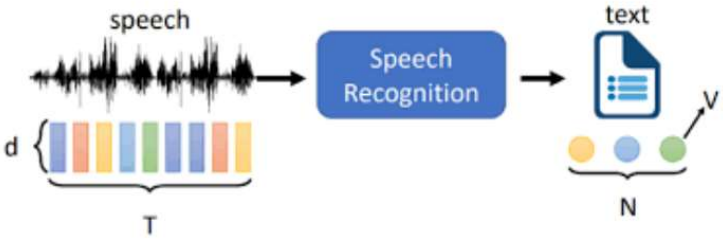


Image classification Object detection Speech recognition

Phenomena

- Real-world datasets contain erroneously labeled data samples.
- Well-labeled data is usually expensive.
- Learning from noisy labels significantly degrades DNNs' performances.

Problem

- How to train a robust model by using large-scale noisy data?

Noisy Label Learning

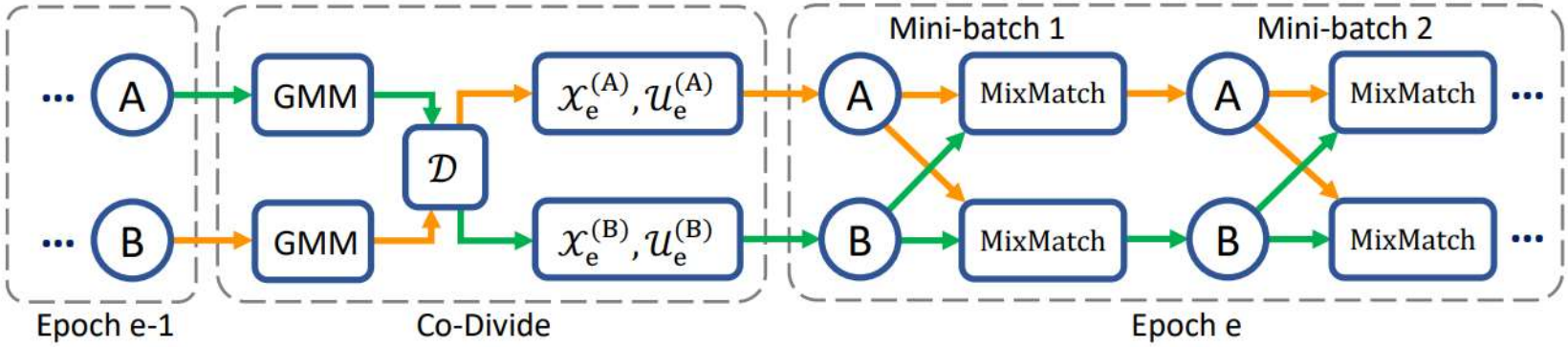
Loss correction

- Estimate noise transition matrix.
- Self-training with pseudo-labels.
- Design noisy-tolerant loss functions.

Example selection

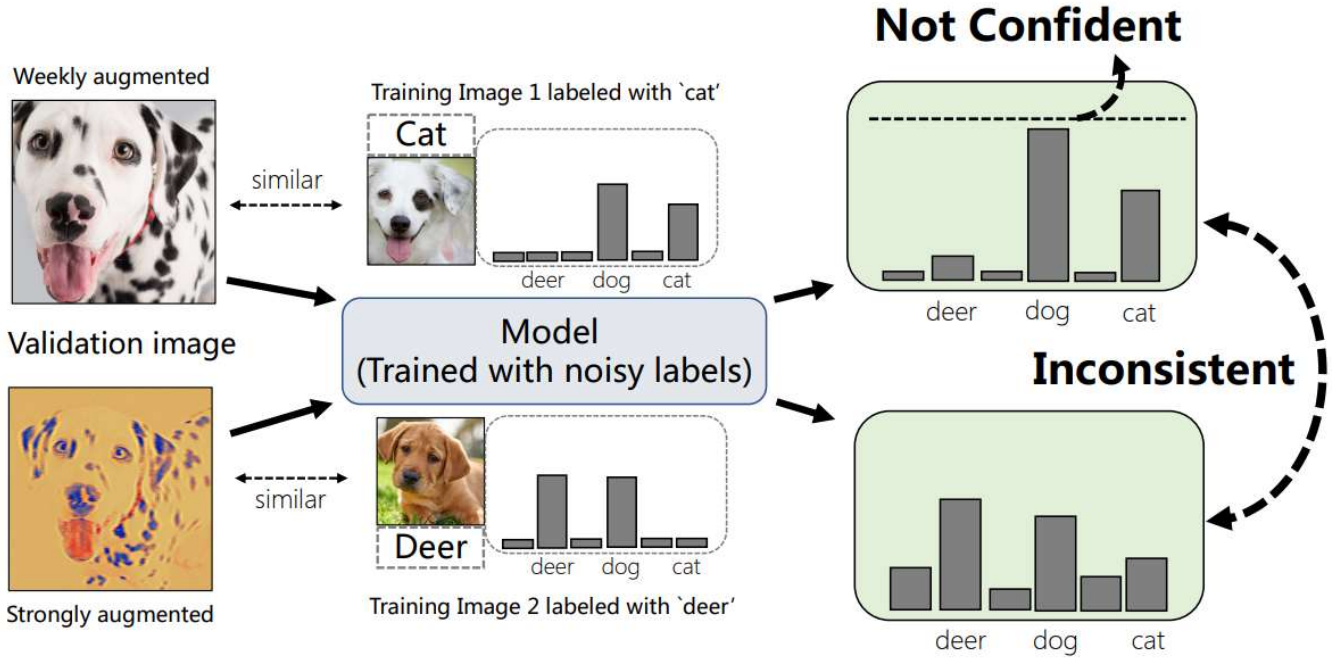
✓ Small loss criterion: memorization effect illuminates that DNNs learn clean and simple patterns faster than noisy labels.

✓ Dual-network strategy: ease confirmation bias by trains two networks simultaneously and let them select examples for each other.



Motivation

- The one-dimensional loss is an over-simplified metric that fails to accommodate the complex feature landscape of various samples.



- Obtaining robust and consistent representation for samples across categories and subjects is difficulty in noisy label learning.

RankMatch

Sample-selection via Confidence Voting (SCV)

$\mathbf{p}[c]$ is the c-th element of $\mathbf{p}$ $(= P(F(\mathbf{x}, \theta)))$

- Basic confident samples:

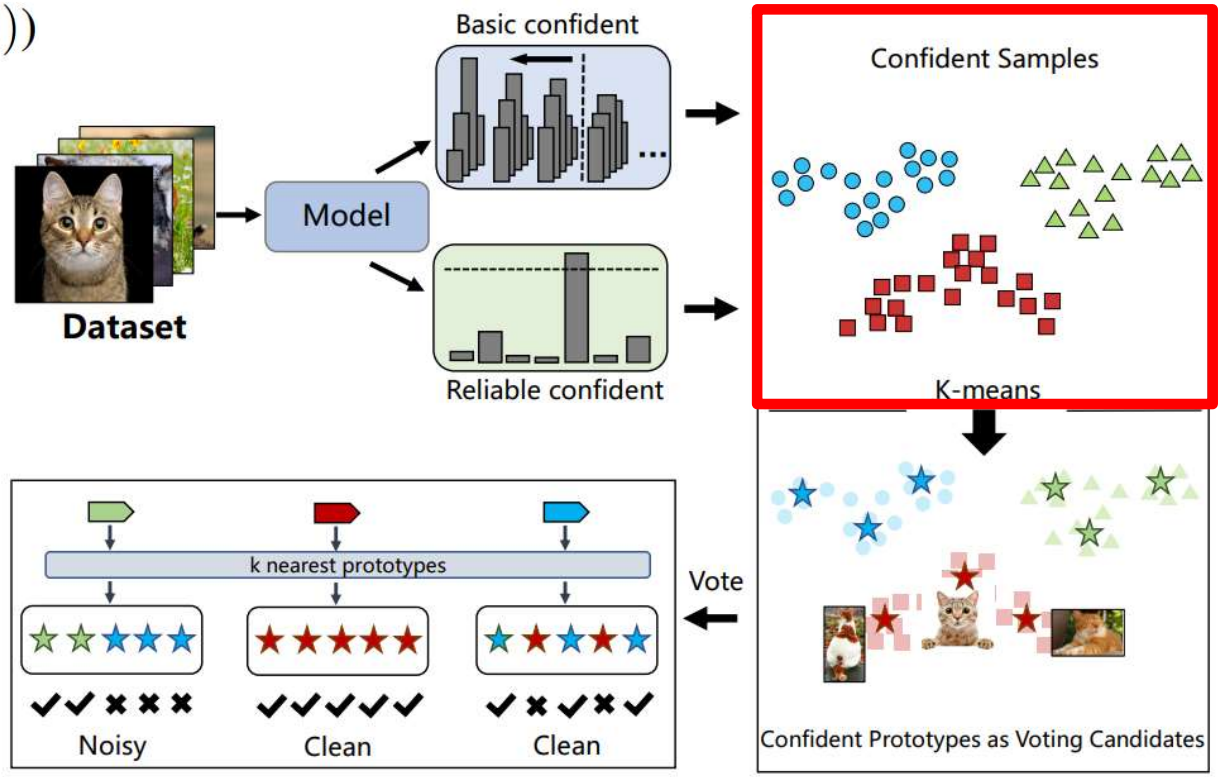
$$\mathcal{I}_b^c = \underset{|\hat{\mathcal{X}}|=B, \hat{\mathcal{X}} \subseteq \mathcal{X}}{\arg\max} \sum_{\mathbf{x} \in \hat{\mathcal{X}}} \mathbf{p}[c],$$

- Reliable confident samples:

$$\mathcal{I}_r^c = \{\mathbf{x}_i | \mathbf{p}_i[c] > \tau, \mathbf{x}_i \in \mathcal{X}\},$$

- Confident samples:

$$\mathcal{I}^c = \mathcal{I}_b^c \cup \mathcal{I}_r^c.$$



RankMatch

Sample-selection via Confidence Voting (SCV)

- Confident Prototype Generation:

$$\phi^c = \frac{1}{|\mathcal{I}^c|} \sum_{\mathbf{x} \in \mathcal{I}^c} F(\mathbf{x}, \theta).$$

- Split the confident samples into K clusters:

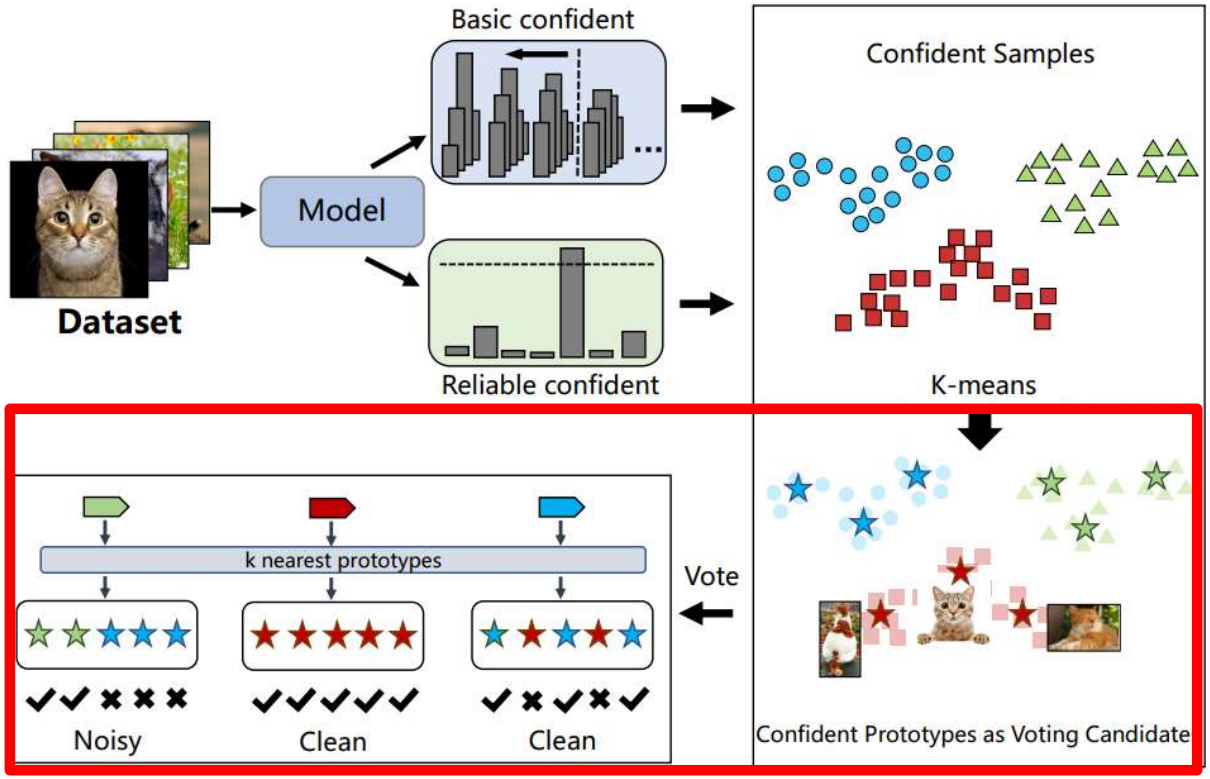
$$\Phi^c = \{\phi_j^c\}_{j=1}^K.$$

- Confidence Voting:

$$\mathcal{V}_i = \underset{|\hat{\Phi}|=k, \hat{\Phi} \subseteq \Phi}{\arg\max} \sum_{\phi \in \hat{\Phi}} \cos(\mathbf{f}_i, \phi).$$

- Clean set:

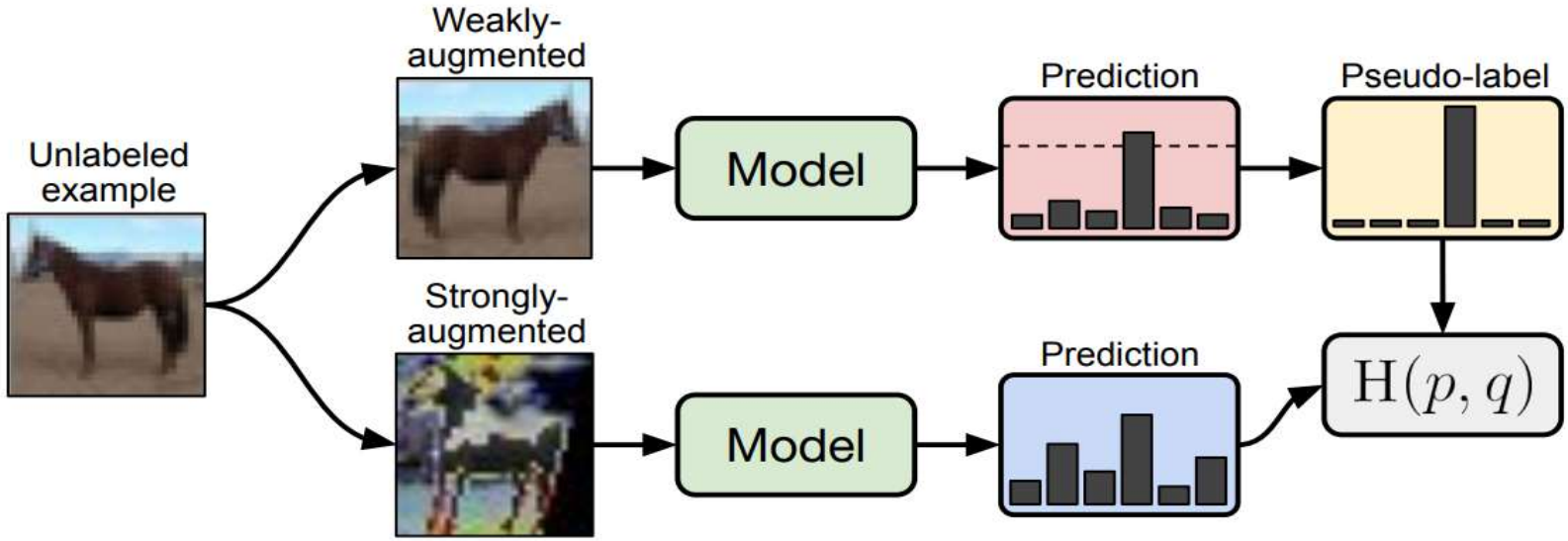
$$\mathcal{D}_{\text{cln}} = \{(\mathbf{x}_i, \widetilde{\mathbf{y}}_i) | \widetilde{\mathbf{y}}_i = \mathbf{y}'_i, \forall (\mathbf{x}_i, \widetilde{\mathbf{y}}_i) \in \mathcal{D}\}$$



RankMatch

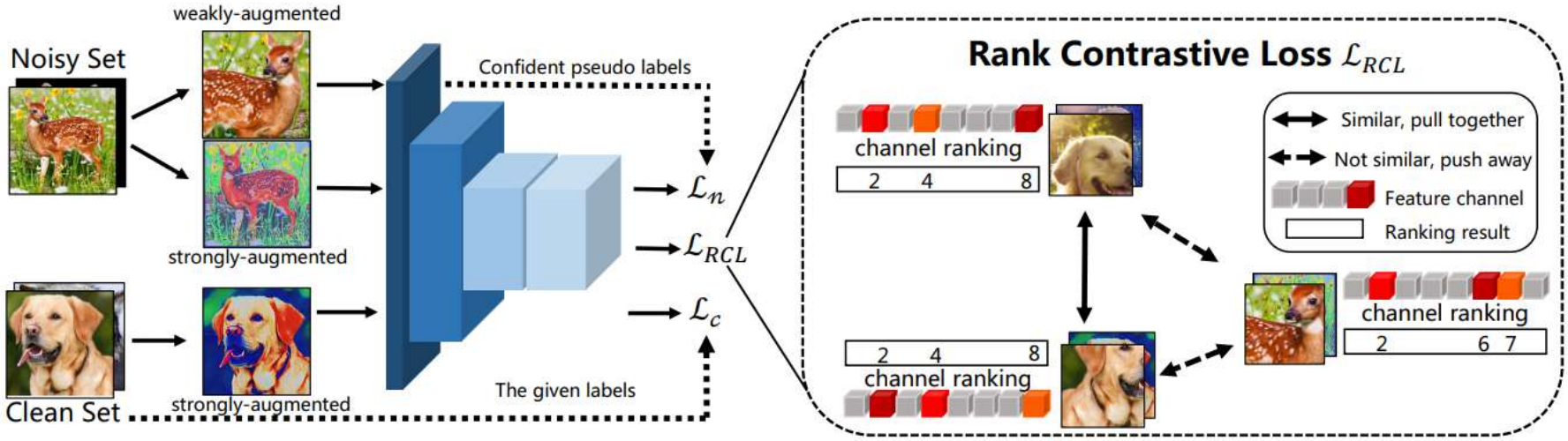
- For clean samples: $\mathcal{L}_c = -\frac{1}{|\mathcal{D}_{\text{cln}}|} \sum_{(\mathbf{x},\widetilde{\mathbf{y}})\in\mathcal{D}_{\text{cln}}} \widetilde{\mathbf{y}} \circ \log(\mathbf{p}^s),$
- For noisy samples with high prediction confidence:

$$\hat{\mathcal{D}}_{\text{nsy}} = \{(\mathbf{x}_i, \hat{\mathbf{y}}_i) | \forall \mathbf{x}_i \in \mathcal{X}_{\text{nsy}}, \max_c \{\mathbf{p}^w[c]\} > \nu, \hat{\mathbf{y}}_i = c\}, \quad \mathcal{L}_n = -\frac{1}{|\hat{\mathcal{D}}_{\text{nsy}}|} \sum_{(\mathbf{x},\hat{\mathbf{y}})\in\hat{\mathcal{D}}_{\text{nsy}}} \hat{\mathbf{y}} \circ \log(\mathbf{p}^s).$$



RankMatch

- For hard-to-learn noisy samples: (**Rank Contrastive Loss**)
  - ✓ Every convolution kernel filters out certain kinds of attributes in the input image.
  - ✓ Similar visual patterns are prone to activate the same representation channel of response maps produced by convolution layers.
  - ✓ The indices of feature elements rank ordered in accordance with their magnitudes, can serve as a metric for assessing the pairwise representations similarity.



RankMatch

- For hard-to-learn noisy samples: (**Rank Contrastive Loss**)

$$\mathbf{f}_i[n] \text{ denote the } n\text{-th channel of feature } \mathbf{f}_i.$$

➢ Principal feature dimensions:

$$\mathcal{R}_i = \underset{|\hat{\mathcal{R}}|=r,\ \hat{\mathcal{R}}\subset\{1,2,\ldots,L\}}{\arg\max} \sum_{n\in\hat{\mathcal{R}}} \mathbf{f}_i[n],$$

➢ Similarity:

$$s_{ij} = \begin{cases} 1 & \text{if } \mathcal{R}_i = \mathcal{R}_j; \\ 0 & \text{otherwise.} \end{cases}$$

➢ Rank Contrastive Loss (binary cross-entropy loss):

$$\mathcal{L}_{\text{RCL}} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[s_{ij}log(\mathbf{p}_i\circ\mathbf{q}_j)+(1-s_{ij})log(1-\mathbf{p}_i\circ\mathbf{q}_j)],$$

where $\mathbf{p}_i = P(F(\mathbf{v}_i^w, \theta))$ and $\mathbf{q}_i = P(F(\mathbf{v}_i^s, \theta))$.

RankMatch

- For hard-to-learn noisy samples: (**Rank Contrastive Loss**)

$$\mathcal{L}_{\text{RCL}} = \overbrace{-\frac{1}{N^2}\sum_{i=1}^{N}[s_{ii}\log(\mathbf{p}_i \circ \mathbf{q}_i) + (1-s_{ii})\log(1-\mathbf{p}_i \circ \mathbf{q}_i)]}^{\text{Consistency regularization}}$$
$$-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}[s_{ij}\log(\mathbf{p}_i \circ \mathbf{q}_j) + (1-s_{ij})\log(1-\mathbf{p}_i \circ \mathbf{q}_j)]. \quad (12)$$

Since $s_{ii} = 1$ always holds, organize the first term as:

$$-\frac{1}{N^2}\sum_{i=1}^{N}\log(\mathbf{p}_i \circ \mathbf{q}_i) = -\frac{1}{N^2}\sum_{i=1}^{N}\log(\mathbf{p}_i^w \circ \mathbf{p}_i^s). \quad (13)$$

Minimizing this term is equal to $\max\{\mathbf{p}_i^w \circ \mathbf{p}_i^s\}$.

- Overall loss function:

$$\mathcal{L} = \mathcal{L}_c + \lambda_n \mathcal{L}_n + \mathcal{L}_{\text{RCL}} + \mathcal{L}_{div}.$$

$$\mathcal{L}_{div} = \sum_{c=1}^{C}\frac{1}{C}\log\left(\frac{1}{C}\Big/\frac{\sum_{i=1}^{N}\mathbf{p}_i^w[c]}{N}\right).$$

Experiment

Dataset		CIFAR-10					CIFAR-100			
Method/Noise ratio		20%	50%	80%	90%	Asym. 40%	20%	50%	80%	90%
Cross-Entropy	Best	86.8	79.4	62.9	42.7	85.0	62.0	46.7	19.9	10.1
	Last	82.7	57.9	26.1	16.8	72.3	61.8	37.3	8.8	3.5
PENCIL [56]	Best	92.4	89.1	77.5	58.9	88.5	69.4	57.5	31.1	15.3
	Last	92.0	88.7	76.5	58.2	88.1	68.1	56.4	20.7	8.8
Meta-Learning [25]	Best	92.9	89.3	77.4	58.7	89.2	68.5	59.2	42.4	19.5
	Last	92.0	88.8	76.1	58.3	88.4	67.7	58.0	40.1	14.3
DivideMix [24]	Best	96.1	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
	Last	95.7	94.4	92.9	75.4	92.1	76.9	74.2	59.6	31.0
ELR [30]	Best	95.8	94.8	93.3	78.7	93.0	77.6	73.6	60.8	33.4
MOIT [38]	Best	94.1	91.1	75.8	70.1	93.2	75.9	70.1	51.4	24.5
RRL [26]	Last	96.4	95.3	93.3	77.4	93.3	**80.3**	76.0	61.1	33.1
CTRR [28]	Best	93.3	-	86.7	84.3	89.0	70.1	-	43.7	-
RankMatch	Best	**96.5**	**95.6**	**94.5**	**92.6**	**94.7**	79.5	**77.9**	**67.6**	**50.6**
	Last	**96.4**	**95.4**	**94.2**	**92.1**	**94.4**	79.3	**77.6**	**67.2**	**49.9**

Table 1: Comparison between RankMatch and state-of-the-art methods on CIFAR-10 and CIFAR-100 under symmetric and asymmetric noise. "Best" refers to the best test accuracy across all epochs and "Last" is the averaged test accuracy over the last 10 epochs.

Experiment

Method	Test Accuracy
Cross-Entropy	69.21
PENCIL [56]	73.49
DivideMix [24]	74.76
RRL [26]	74.97
DSOS [1]	73.63
UNICON [19]	74.98
SOP [31]	73.50
SFT [50]	75.08
RankMatch	**75.22**

Table 2: Comparison with state-of-the-art methods on Clothing1M. Baseline results are copied from original papers.

Method	WebVision		ILSVRC12	
	top1	top5	top1	top5
MentorNet [18]	63.00	81.40	57.80	79.92
Co-teaching [14]	63.58	85.20	61.48	84.70
Iterative-CV [7]	65.24	85.34	61.60	84.98
DivideMix [24]	77.32	91.64	75.20	90.84
RRL [26]	77.8	91.3	74.4	90.9
DSOS [1]	77.76	92.04	74.36	90.80
SOP [31]	76.6	-	69.1	-
UNICON [19]	77.60	93.44	75.29	93.72
RankMatch	**79.91**	**93.61**	**77.39**	**94.26**

Table 3: Comparison with state-of-the-art methods on (mini) WebVision dataset. Numbers denote top-1 (top-5) accuracy (%) on the WebVision and ImageNet ILSVRC12 validation set. Results for baselines are copied from the corresponding papers.

Experiment

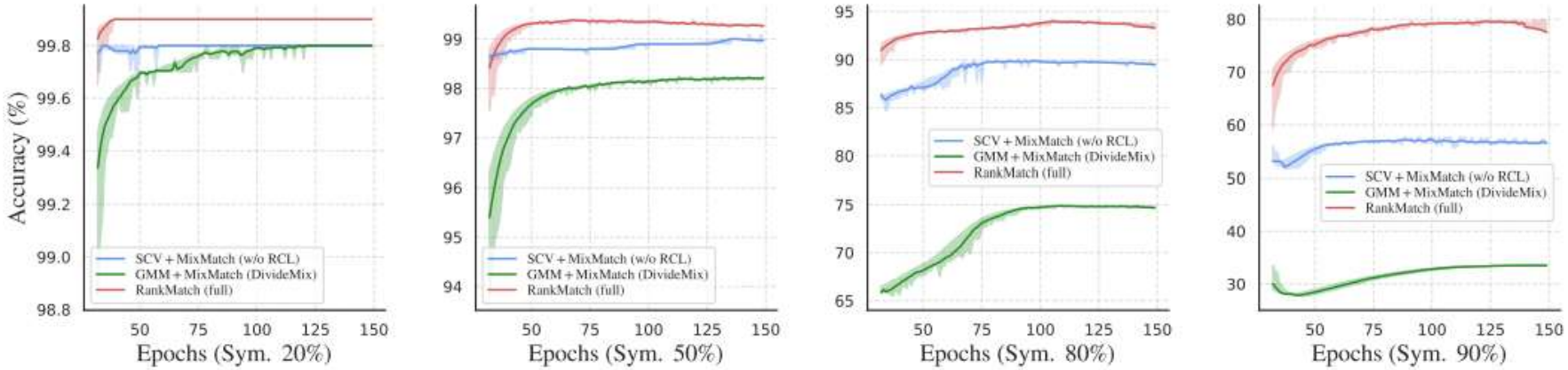


Figure 5: Accuracy (%) of the selected clean samples under different symmetric noise rates on CIFAR-100 datasets. Our full method RankMatch is compared with two baselines. One is DivideMix, denoted as GMM and MixMatch. The other is replacing GMM with our proposed Sample-selection via Confident Voting (SCV) method. Compared with GMM, SCV is more efficient to identify the clean samples. And our full model RankMatch has significant improvement in the sample selection stage under high noise rate.

Experiment

Dataset		CIFAR-10		CIFAR-100	
Noise ratio		50%	90%	50%	90%
RankMatch	Best	**95.6**	**92.6**	**77.9**	**50.6**
	Last	**95.4**	**92.1**	**77.6**	**49.9**
w/o SCV	Best	95.1	89.6	75.3	45.1
	Last	94.9	89.1	74.9	44.9
SCV w/o K-means	Best	95.4	91.8	76.5	47.9
	Last	95.2	91.4	75.2	47.5
w/o RCL	Best	95.2	90.8	76.1	45.6
	Last	94.9	88.7	75.2	44.5

Table 4: Ablation study results in terms of test accuracy (%) on CIFAR-10 and CIFAR-100 with symmetric label noise.

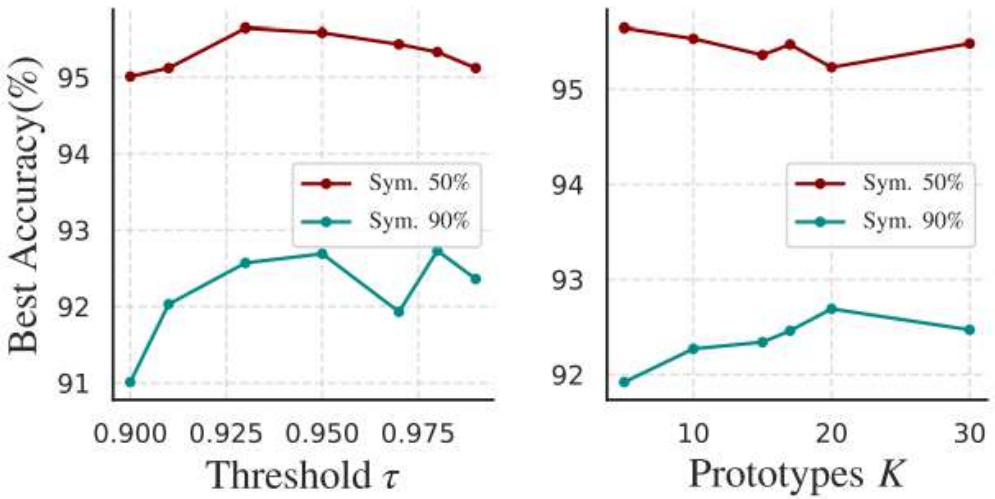
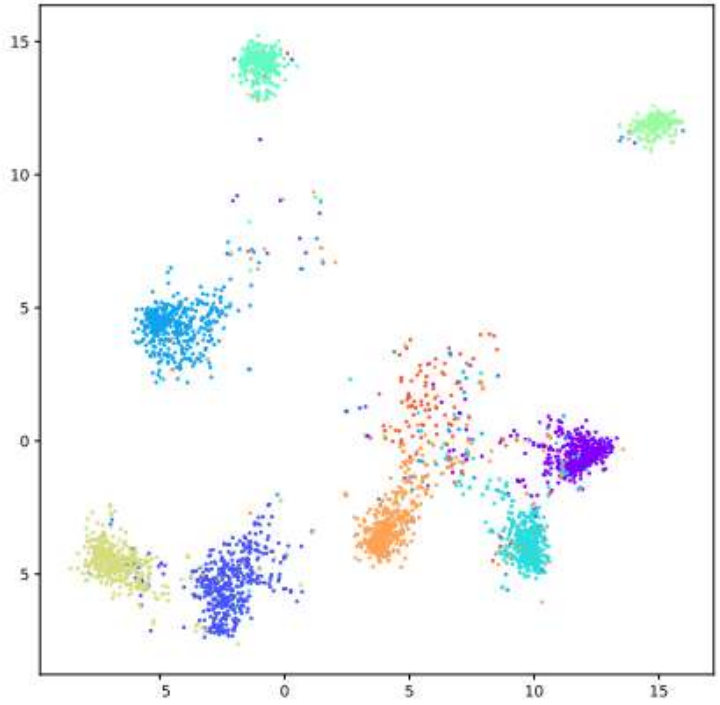
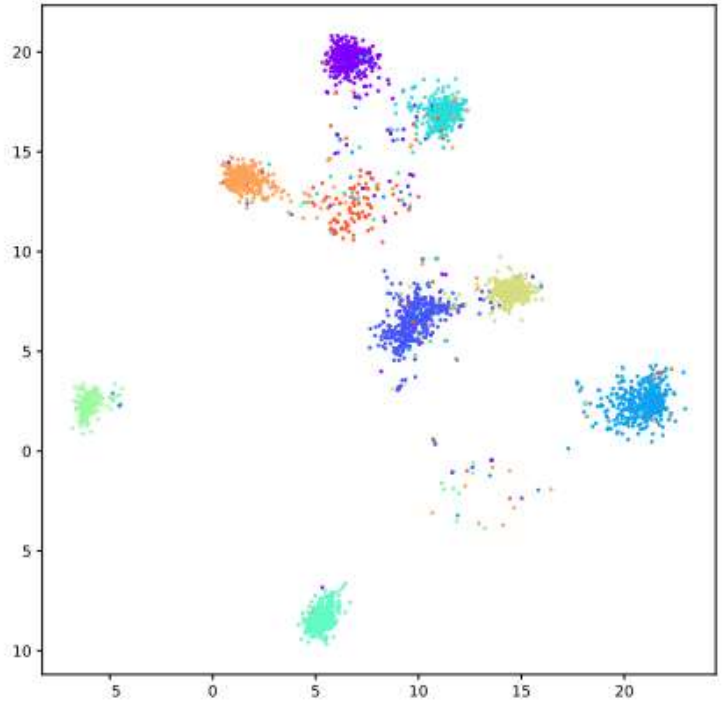


Figure 6: Sensitivity to the variance of hyperparameters. Experiments are conducted on CIFAR-10 under 50% and 90% symmetric noises.

Experiment



(a) RankMatch without RCL



(b) RankMatch with RCL

Experiment

Table 1: List of RankMatch hyperparameters for CIFAR

Hyperparameter	CIFAR-10				CIFAR-100			
	20%	50%	80%	90%	20%	50%	80%	90%
λ_n	0.2	1	10	10	0.5	2	5	8

THANKS