# Twin Contrastive Learning with Noisy Labels

Zhizhong Huang[1] Junping Zhang[1] Hongming Shan[2,3,*]

[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China
[2]Institute of Science and Technology for Brain-inspired Intelligence and MOE Frontiers Center for Brain Science
Fudan University, Shanghai 200433, China
[3]Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 201210, China

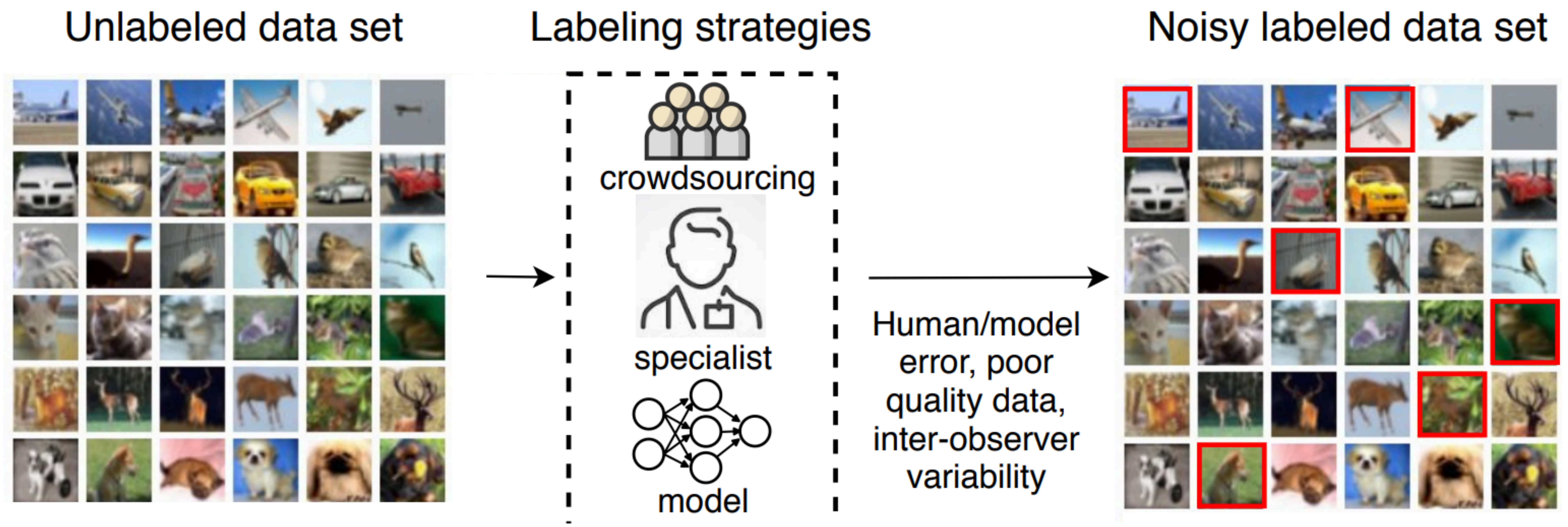
Code is available at https://github.com/Hzzone/TCL

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Contents

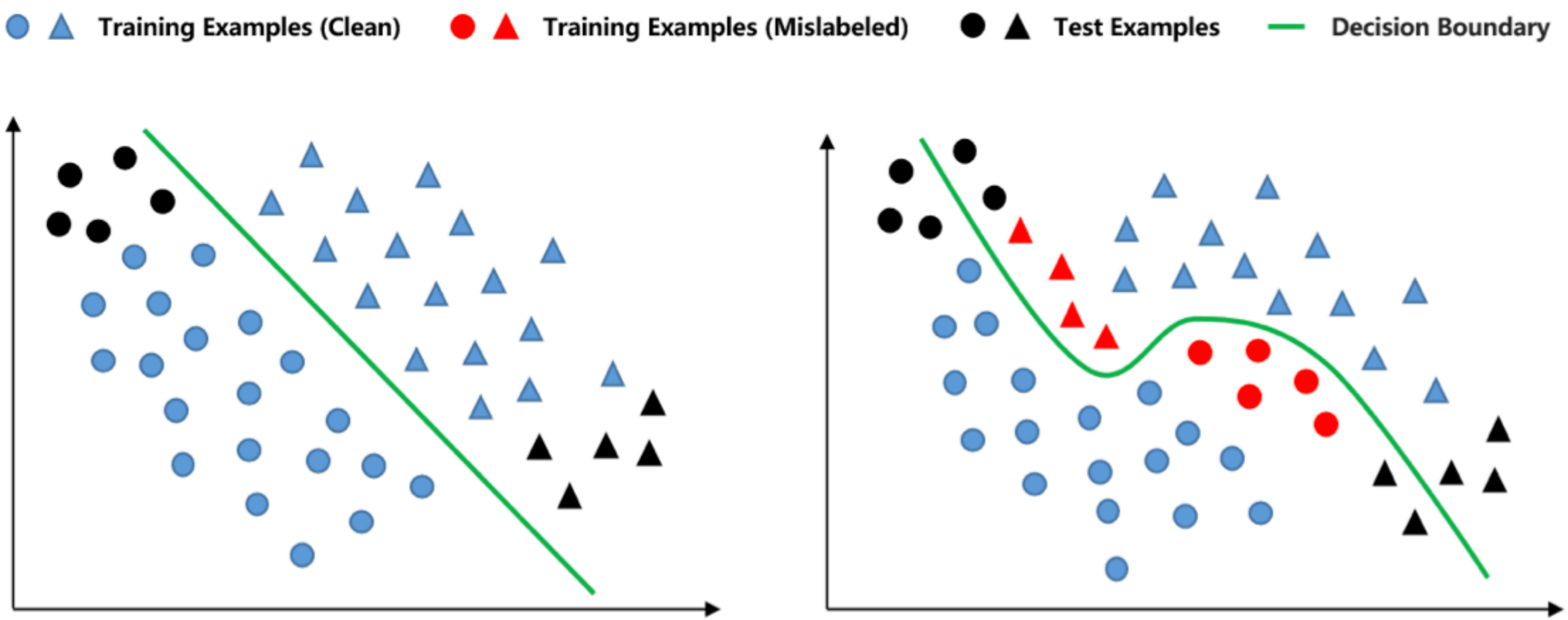
1. Background
2. The Proposed TCL
    - Modeling Data Distribution
    - Out-Of-Distribution Label Noise Detection
    - Cross-supervision with Entropy Regularization
    - Learning Robust Representations
3. Experiments & Results

# Background




- The label noise is produced during labeling process.

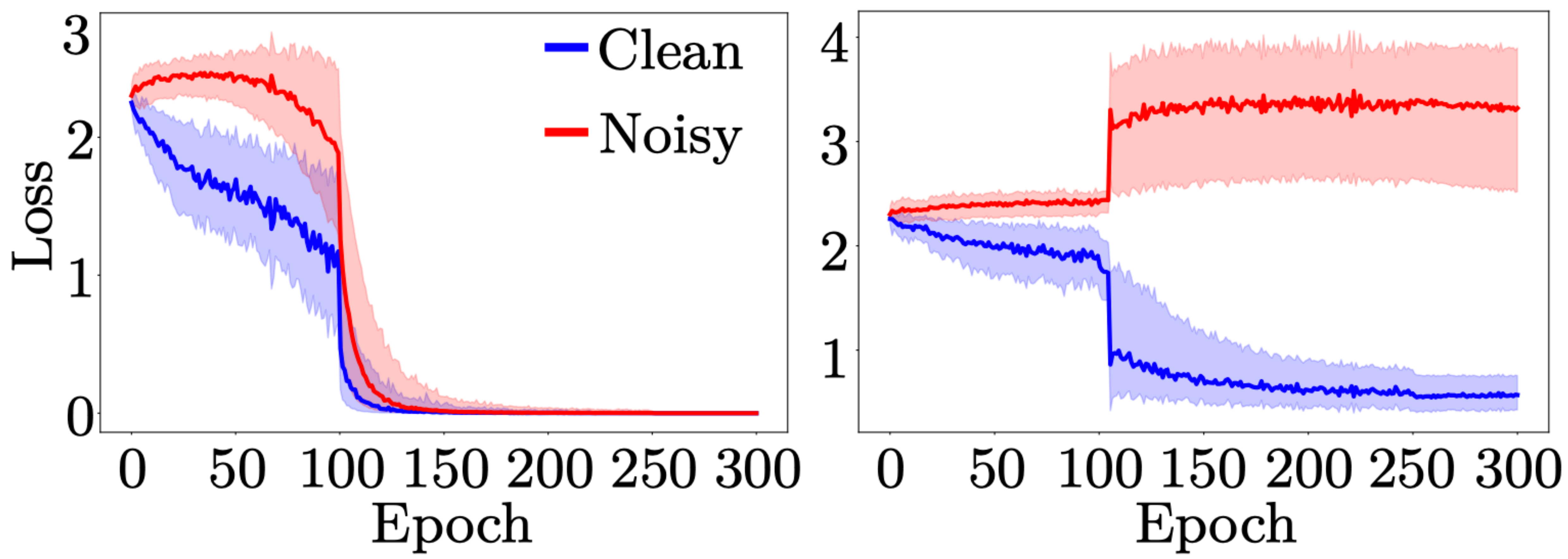
# Background

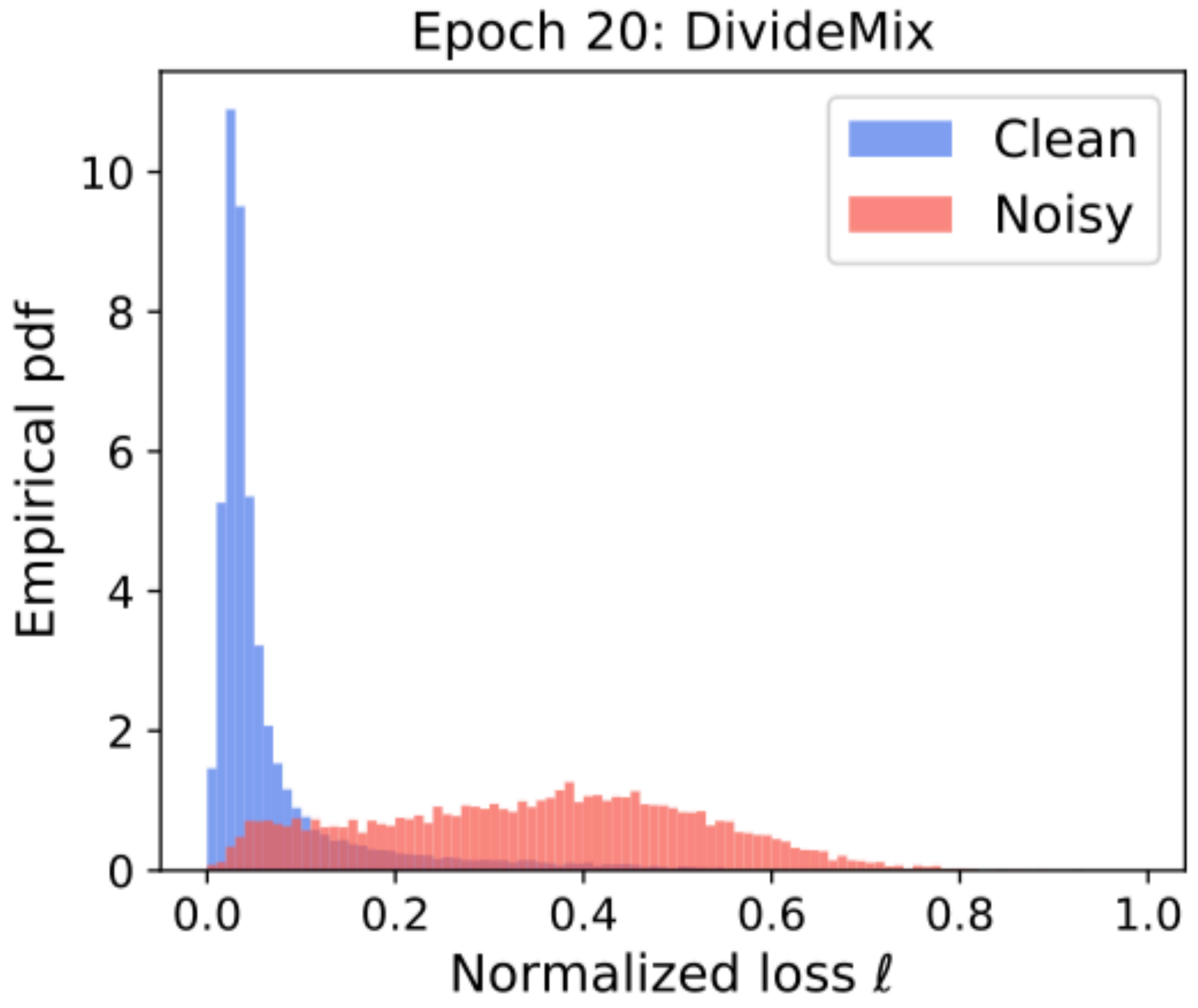



- Training with mislabeled examples would lead to WRONG decision Boundary

Pictures from Selective-Supervised Contrastive Learning with Noisy Labels (CVPR 2022)

# Background



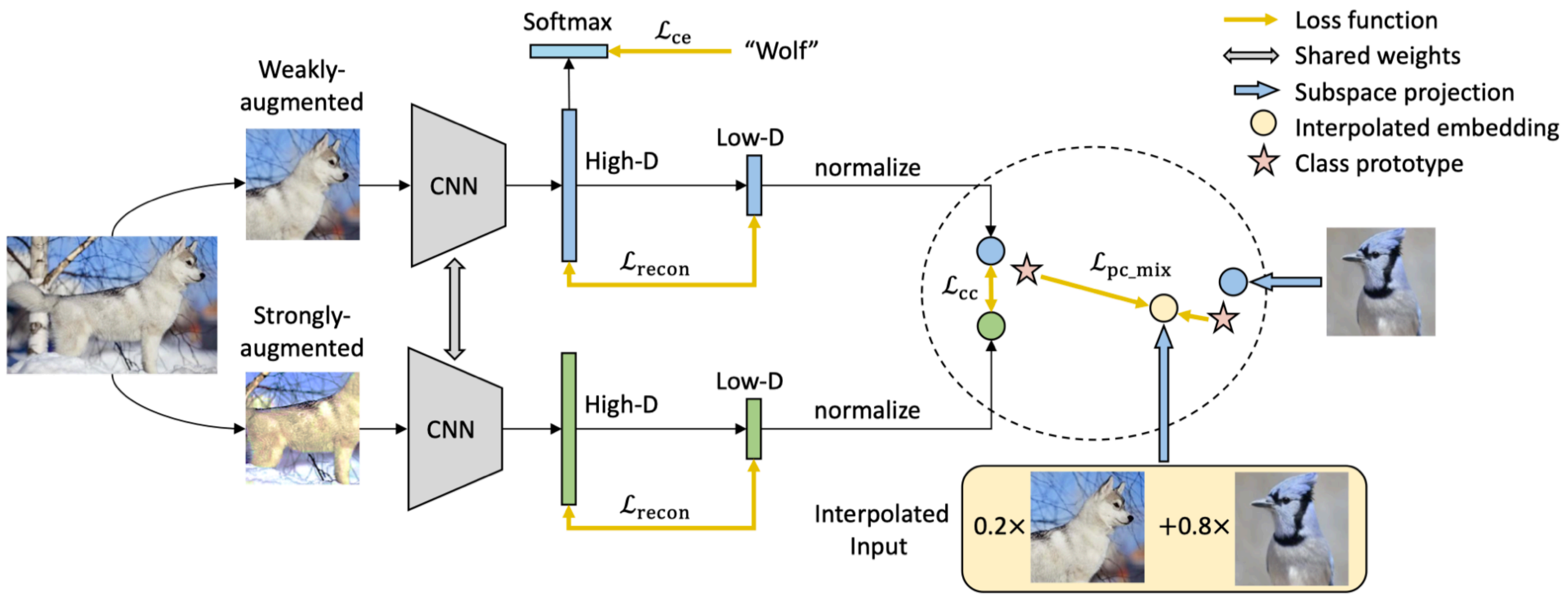




(c)

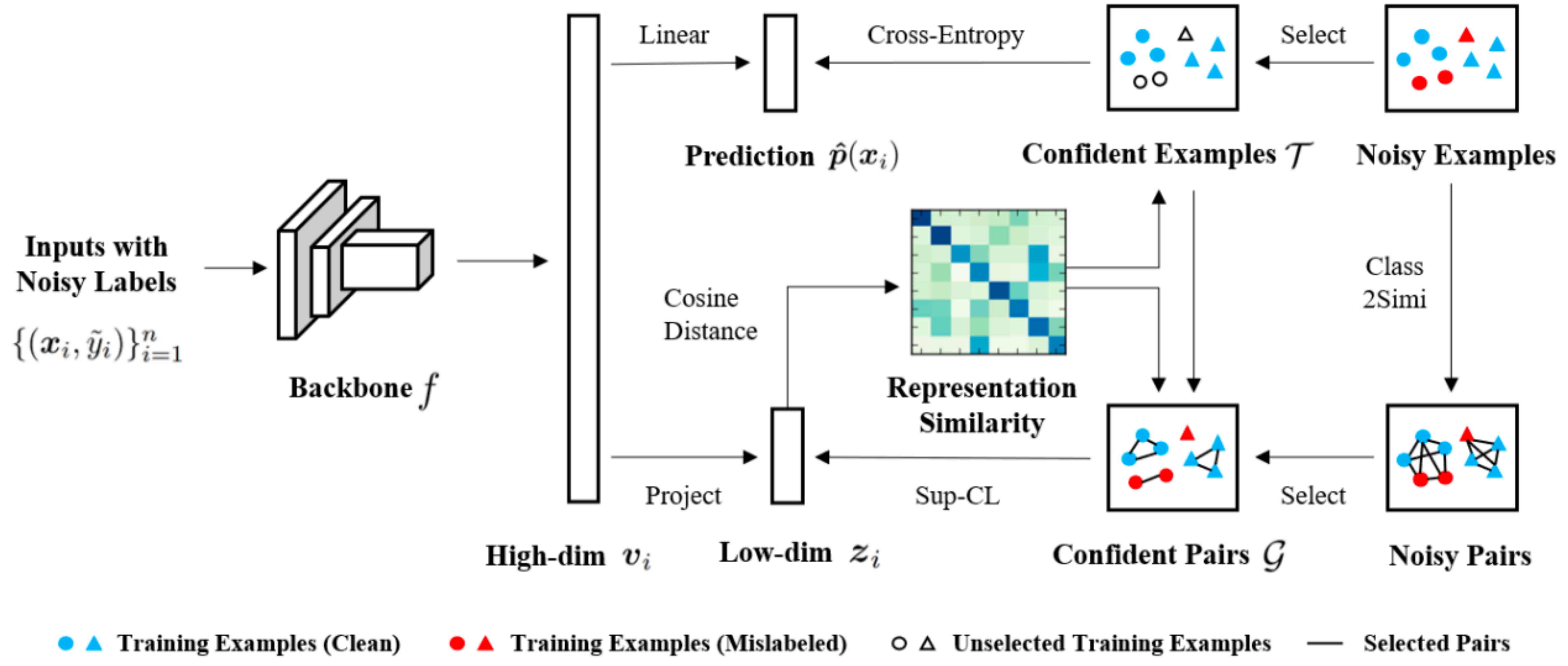
- Previous work usually select the clean samples with small loss trick.
- Small loss trick: the neural network tend to fit the clean samples which has small losses.

Pictures from Unsupervised Label Noise Modeling and Loss Correction (NIPS 2019) and DivideMix: Learning with Noisy Labels as Semi-supervised Learning (ICLR 2020)

# Background




Cleaning datasets with nearest neighbors
Learning from Noisy Data with Robust Representation Learning (ICCV 2021)

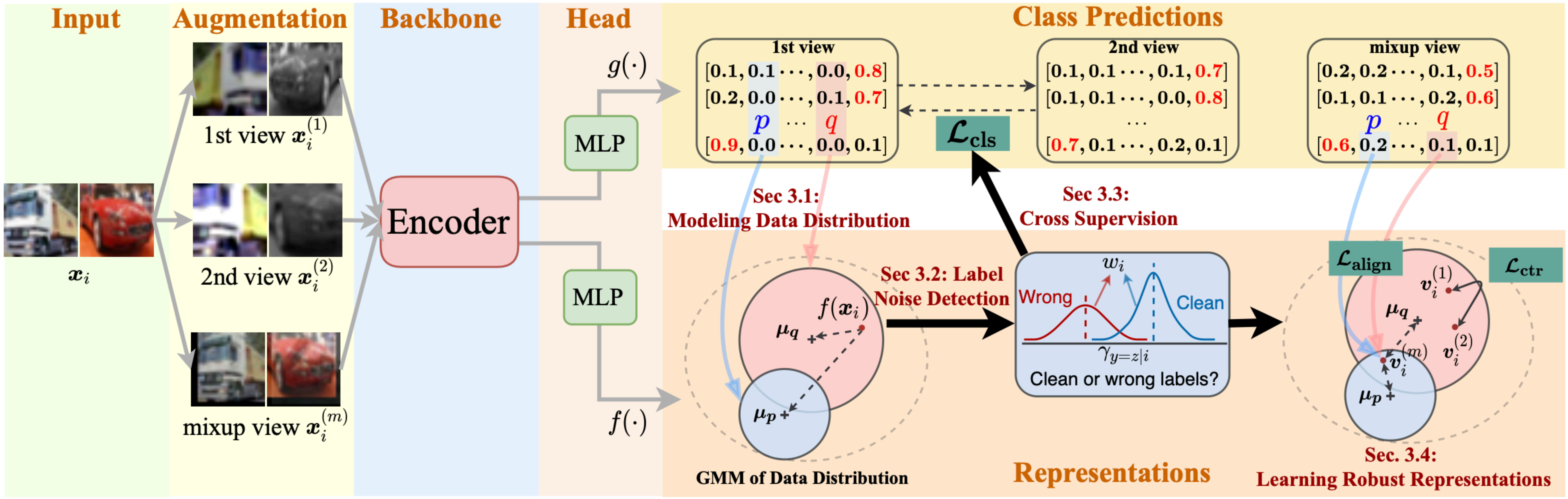



Construct confident positive pairs for supervised contrastive learning
Selective-Supervised Contrastive Learning with Noisy Labels (CVPR 2022)

Contrastive Learning enables Noisy Label Learning by the Unsupervised Noise-Robust Representations

They cannot handle extremely noisy scenario when Nearest Neighbors are All mislabeled!

# The Proposed TCL




The proposed TCL

(1) leverages contrastive learning for learning robust representations,

(2) models the data distribution via a GMM, and

(3) detects the examples with wrong labels as out-of-distribution examples.

# The Proposed TCL

How to model the data distribution:

Given the representation $v = f(x)$ and discrete latent variables $z \in \{1,2,\ldots,K\}$, the unsupervised GMM can be defined as

$$
\begin{aligned}
p(\boldsymbol{v}) &= \sum\nolimits_{k=1}^{K} p(\boldsymbol{v}, z=k) \\
&= \sum\nolimits_{k=1}^{K} p(z=k)\mathcal{N}(\boldsymbol{v}|\boldsymbol{\mu}_k, \sigma_k).
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\mu}_k &= \operatorname{norm}\left(\frac{\sum_i p_\theta(y_i = k|\boldsymbol{x}_i)\boldsymbol{v}_i}{\sum_i p_\theta(y_i = k|\boldsymbol{x}_i)}\right), \\
\sigma_k &= \frac{\sum_i p_\theta(y_i = k|\boldsymbol{x}_i)(\boldsymbol{v}_i - \boldsymbol{\mu}_k)(\boldsymbol{v}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_i p_\theta(y_i = k|\boldsymbol{x}_i)}
\end{aligned}
$$

# The Proposed TCL

Out-Of-Distribution Label Noise Detection

The posterior probability can be defined as:

$$\gamma_{ik} = \frac{\exp\left(-(\boldsymbol{v}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}(\boldsymbol{v}_i - \mu_k)/2\sigma_k\right)}{\sum_k \exp\left(-(\boldsymbol{v}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}(\boldsymbol{v}_i - \boldsymbol{\mu}_k)/2\sigma_k\right)}$$

$$\begin{aligned}\gamma_{ik} &= p(z_i = k|\boldsymbol{x}_i) \\ &= \exp(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\mu}_k/\sigma_k) \Big/ \sum\nolimits_k \exp(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\mu}_k/\sigma_k)\end{aligned}$$

Then, we can introduce the noisy label $y$ and define the following conditional probability to measure the probability of one sample with clean label:

$$\begin{aligned}\gamma_{y=z|i} &= p(y_i = z_i|\boldsymbol{x}_i) \\ &= \exp(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\mu}_{z_i}/\sigma_{z_i}) \Big/ \sum\nolimits_k \exp(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\mu}_k/\sigma_k)\end{aligned}$$

# The Proposed TCL

## Out-Of-Distribution Label Noise Detection

Another two-component GMM is employed to automatically classify the clean/wrong labels:

$$p(\gamma_{y=z|i}) = \sum_{c=0}^{1} p(\gamma_{y=z|i}, c) = \sum_{c=0}^{1} p(c)p(\gamma_{y=z|i}|c)$$

where $c$ is the new introduced latent variable: $c = 1$ indicates the cluster of clean labels with higher mean value and vice versus $c = 0$.

# The Proposed TCL

Cross-supervision with Entropy Regularization

The true targets are the convex combination of its noisy labels and the predictions from the model itself:

$$\begin{cases} \boldsymbol{t}_i^{(1)} = w_i \boldsymbol{y}_i + (1 - w_i) g(\boldsymbol{x}_i^{(1)}) \\ \boldsymbol{t}_i^{(2)} = w_i \boldsymbol{y}_i + (1 - w_i) g(\boldsymbol{x}_i^{(2)}) \end{cases}$$

where $w_i \in [0, 1]$ $= p(c = 1|\gamma_{y=z|i})$, $\boldsymbol{y}_i$ the noisy one-hot label. $g(\boldsymbol{x}_i^{(1)})$ and $g(\boldsymbol{x}_i^{(2)})$ are the predictions.

# The Proposed TCL

Cross-supervision with Entropy Regularization

The loss can be defined as:

$$\mathcal{L}_{\text{cross}} = \ell\left(g(\boldsymbol{x}_i^{(1)}), \boldsymbol{t}_i^{(2)}\right) + \ell\left(g(\boldsymbol{x}_i^{(2)}), \boldsymbol{t}_i^{(1)}\right)$$

$$\mathcal{L}_{\text{reg}} = -\mathbb{H}\left(\frac{1}{|\mathcal{B}|}\sum_{\boldsymbol{x}\in\mathcal{B}} g(\boldsymbol{x})\right) + \frac{1}{|\mathcal{B}|}\sum_{\boldsymbol{x}\in\mathcal{B}} \mathbb{H}\left(g(\boldsymbol{x})\right)$$

where $\mathbb{H}(\cdot)$ is the entropy of predictions to avoid the predictions collapsing into a single class and encourage the model to have high confidence for predictions.

# The Proposed TCL

Learning Robust Representations

The contrastive loss and mixup augmentation are employed to learn robust representations.
Contrastive loss:

$$\mathcal{L}_{\text{ctr}} = -\log \frac{\exp\left(f(\boldsymbol{x}^{(1)})^{\mathrm{T}} f(\boldsymbol{x}^{(2)})/\tau\right)}{\sum_{\boldsymbol{x}\in\mathcal{S}} \exp\left(f(\boldsymbol{x}^{(1)})^{\mathrm{T}} f(\boldsymbol{x})/\tau\right)}$$

Mixup augmentation:

$$\begin{cases} \boldsymbol{x}_i^{(\mathrm{m})} = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j, \\ \bar{\boldsymbol{t}}_i^{(\mathrm{m})} = \lambda \bar{\boldsymbol{t}}_i + (1-\lambda)\bar{\boldsymbol{t}}_j, \end{cases}$$

$$\mathcal{L}_{\text{align}} = \ell\left(g(\boldsymbol{x}_i^{(\mathrm{m})}), \bar{\boldsymbol{t}}_i^{(\mathrm{m})}\right) + \ell(p(\boldsymbol{z}|\boldsymbol{x}_i^{(\mathrm{m})}), \bar{\boldsymbol{t}}_i^{(\mathrm{m})})$$

# Experiments & Results

| Noise type/rate | CIFAR-10 *Sym.* 20% | CIFAR-10 *Sym.* 50% | CIFAR-10 *Sym.* 80% | CIFAR-10 *Sym.* 90% | CIFAR-10 *Asym.* 40% | CIFAR-10 *Avg.* | CIFAR-100 *Sym.* 20% | CIFAR-100 *Sym.* 50% | CIFAR-100 *Sym.* 80% | CIFAR-100 *Sym.* 90% | CIFAR-100 *Avg.* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cross-Entropy | 82.7 | 57.9 | 26.1 | 16.8 | 76.0 | 51.9 | 61.8 | 37.3 | 8.8 | 3.5 | 27.8 |
| Mixup (17') [46] | 92.3 | 77.6 | 46.7 | 43.9 | 77.7 | 67.6 | 66.0 | 46.6 | 17.6 | 8.1 | 34.6 |
| P-correction (19') [43] | 92.0 | 88.7 | 76.5 | 58.2 | 91.6 | 81.4 | 68.1 | 56.4 | 20.7 | 8.8 | 38.5 |
| M-correction (19') [1] | 93.8 | 91.9 | 86.6 | 68.7 | 87.4 | 85.7 | 73.4 | 65.4 | 47.6 | 20.5 | 51.7 |
| ELR (20') [25] | 93.8 | 92.6 | 88.0 | 63.3 | 85.3 | 84.6 | 74.5 | 70.2 | 45.2 | 20.5 | 52.6 |
| DivideMix (20') [20] | 95.0 | 93.7 | 92.4 | 74.2 | 91.4 | 89.3 | 74.8 | 72.1 | 57.6 | 29.2 | 58.4 |
| MOIT (21') [29] | 93.1 | 90.0 | 79.0 | 69.6 | 92.0 | 84.7 | 73.0 | 64.6 | 46.5 | 36.0 | 55.0 |
| RRL (21') [21] | **95.8** | **94.3** | 92.4 | 75.0 | 91.9 | 89.8 | **79.1** | **74.8** | 57.7 | 29.3 | 60.2 |
| Sel-CL+ (22') [23] | 95.5 | 93.9 | 89.2 | 81.9 | **93.4** | 90.7 | 76.5 | 72.4 | 59.6 | 48.8 | 64.3 |
| TCL (**ours**) | 95.0 | 93.9 | **92.5** | **89.4** | 92.6 | **92.7** | 78.0 | 73.3 | **65.0** | **54.5** | **67.7** |
| | ±0.1 | ±0.1 | ±0.2 | ±0.2 | ±0.1 | | ±0.2 | ±0.2 | ±0.3 | ±0.5 | |

Results on CIFAR

# Experiments & Results

| | WebVision | | ILSVRC12 | |
|---|---|---|---|---|
| | top1 | top5 | top1 | top5 |
| Forward [30] | 61.1 | 82.6 | 57.3 | 82.3 |
| D2L [26] | 62.6 | 84.0 | 57.8 | 81.3 |
| Iterative-CV [2] | 65.2 | 85.3 | 61.6 | 84.9 |
| Decoupling [27] | 62.5 | 84.7 | 58.2 | 82.2 |
| MentorNet [16] | 63.0 | 81.4 | 57.8 | 79.9 |
| Co-teaching [11] | 63.5 | 85.2 | 61.4 | 84.7 |
| ELR [25] | 76.2 | 91.2 | 68.7 | 87.8 |
| DivideMix [20] | 77.3 | 91.6 | 75.2 | 90.8 |
| RRL [21] | 76.3 | 91.5 | 73.3 | 91.2 |
| NGC [22] | **79.1** | 91.8 | 74.4 | 91.0 |
| MOIT [29] | 77.9 | 91.9 | 73.8 | 91.7 |
| TCL (**ours**) | **79.1** | **92.3** | **75.4** | **92.4** |

Table 4. Results on WebVision (mini).

| Method | Acc (%) |
|---|---|
| Cross-Entropy | 69.2 |
| Label Correction [1] | 71.0 |
| Joint-Opt [34] | 72.2 |
| ELR [25] | 72.8 |
| SL [37] | 74.4 |
| DivideMix [20] | 74.4 |
| MentorMix [15] | 74.3 |
| RRL [21] | **74.8** |
| TCL (**ours**) | **74.8** |

Table 5. Results on Clothing1M.

Results on Real-world Datasets

# Experiments & Results

| Dataset | | CIFAR-10 | | | | CIFAR-100 | | |
|---|---|---|---|---|---|---|---|---|
| Noise type/rate | | *Sym.* | | *Asym.* | *Avg.* | *Sym.* | | *Avg.* |
| | | 50% | 90% | 40% | | 50% | 90% | |
| (i) | Baseline | 70.0 | 20.6 | 77.5 | 56.1 | 47.3 | 6.8 | 27.1 |
| (ii) | Loss [1, 20] | 92.5 | 75.9 | 73.2 | 80.6 | 71.2 | 16.0 | 43.6 |
| | $k$-NN [29] | 92.9 | 79.7 | 91.3 | 88.0 | 70.3 | 39.8 | 55.1 |
| | OOD (**ours**) | 93.1 | 82.1 | 92.0 | 89.1 | 70.7 | 45.9 | 58.3 |
| (iii) | *Ensem.* [25] | 91.3 | 72.7 | 89.8 | 84.6 | 68.2 | 36.9 | 52.6 |
| | $\mathcal{L}_{\text{cross}}$ (**ours**) | 93.9 | 89.4 | 92.6 | 92.0 | 73.3 | 54.5 | 63.9 |
| (iv) | w/o $\mathcal{L}_{\text{reg}}$ | 92.0 | 34.5 | 90.3 | 72.3 | 68.5 | 24.3 | 46.4 |
| (v) | w/o $\mathcal{L}_{\text{align}}$ | 91.8 | 84.6 | 89.7 | 88.7 | 69.4 | 48.4 | 58.9 |
| (vi) | MoCo | 94.4 | 90.7 | 93.1 | 92.7 | 74.0 | 57.3 | 65.6 |

Table 3. Ablation results of different components in TCL.

Ablation Study

# Experiments & Results

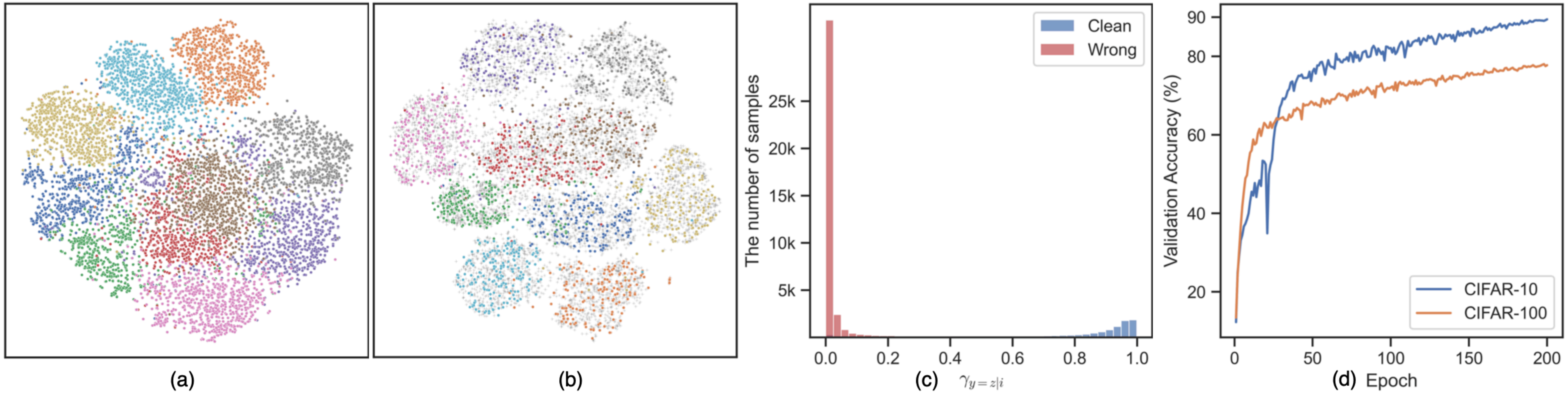



Figure 2. Qualitative results. For the model trained on CIFAR-10 with 90% *sym.* noise at 200th epoch, we show t-SNE visualizations for the learned representations of (a) testing set where different color denotes different class predicted by $g(\cdot)$ and (b) 10K samples from training set colored by the true labels; the gray '+' denotes the samples with noisy labels. (c) The histogram of $p(y = z|\boldsymbol{x})$ for full training set colored by the clean and noisy labels. (d) The validation accuracy across training of CIFAR-10 and CIFAR-100 on 90% *sym.* noise.

Visualization

# ***Thank You!***

*On behalf of all my co-authors*