Noisy-Pair Robust Representation Alignment for Positive-Unlabeled Learning

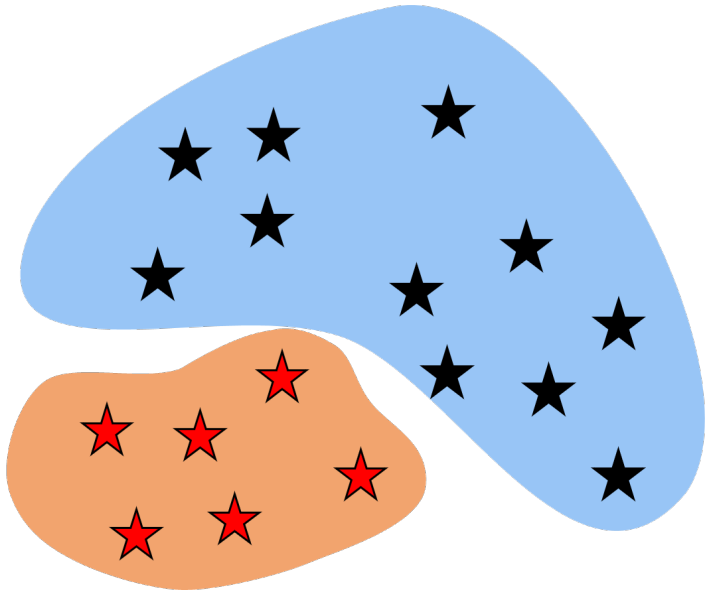
Hengwei Zhao* Zhengzhong Tu Zhuo Zheng Wei Wang Junjue Wang

Rusty Feagin Wenzhe Jiao*

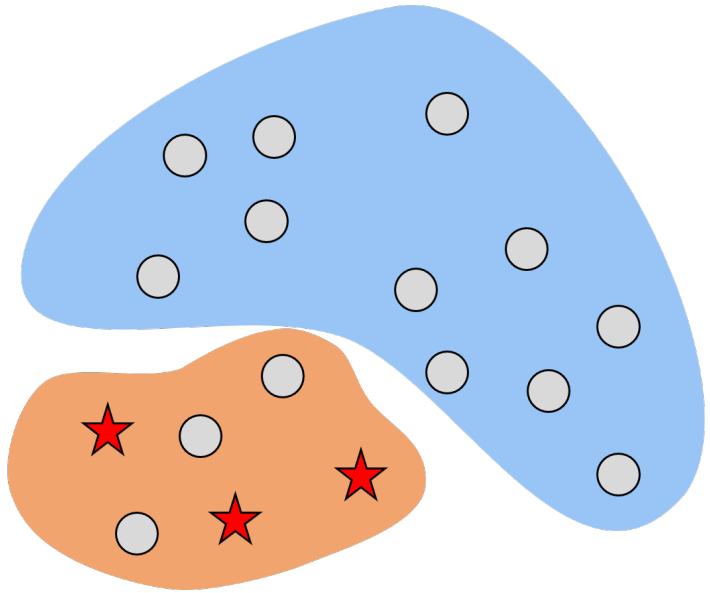
TEXAS A&M UNIVERSITY

What is Positive-Unlabeled (PU) Learning?

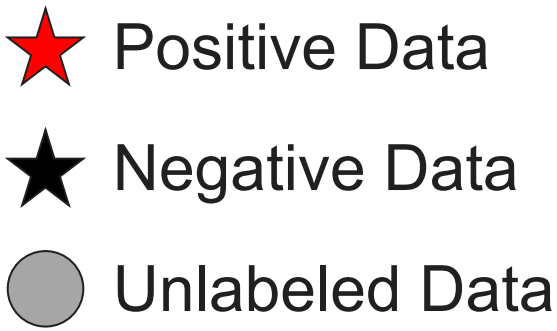
- PU learning aims to learn a binary classifier with limited positive data and a large pool of unlabeled data



Supervised Learning



PU Learning



Why is PU Learning Important?

- In numerous real-world applications, positive data are readily obtainable, whereas negative labels are scarce or even infeasible to acquire

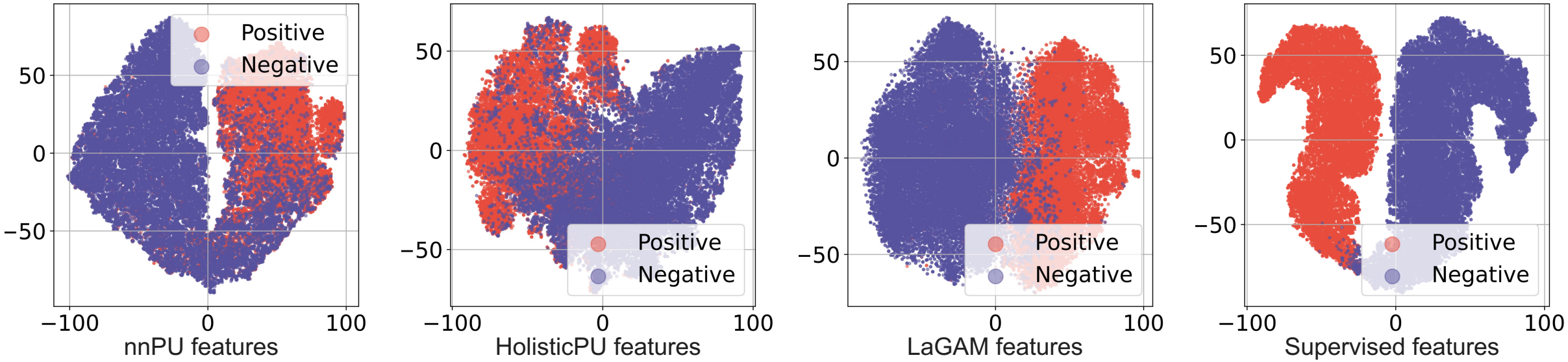
Pre-disaster image

Post-disaster image

Example: After a disaster, training a classification model using only positive (damaged buildings) data can significantly shorten the time required for building damage mapping

What Limits PU Classifier vs. Supervised Methods?

- Inaccurate annotations impede learning discriminative representations

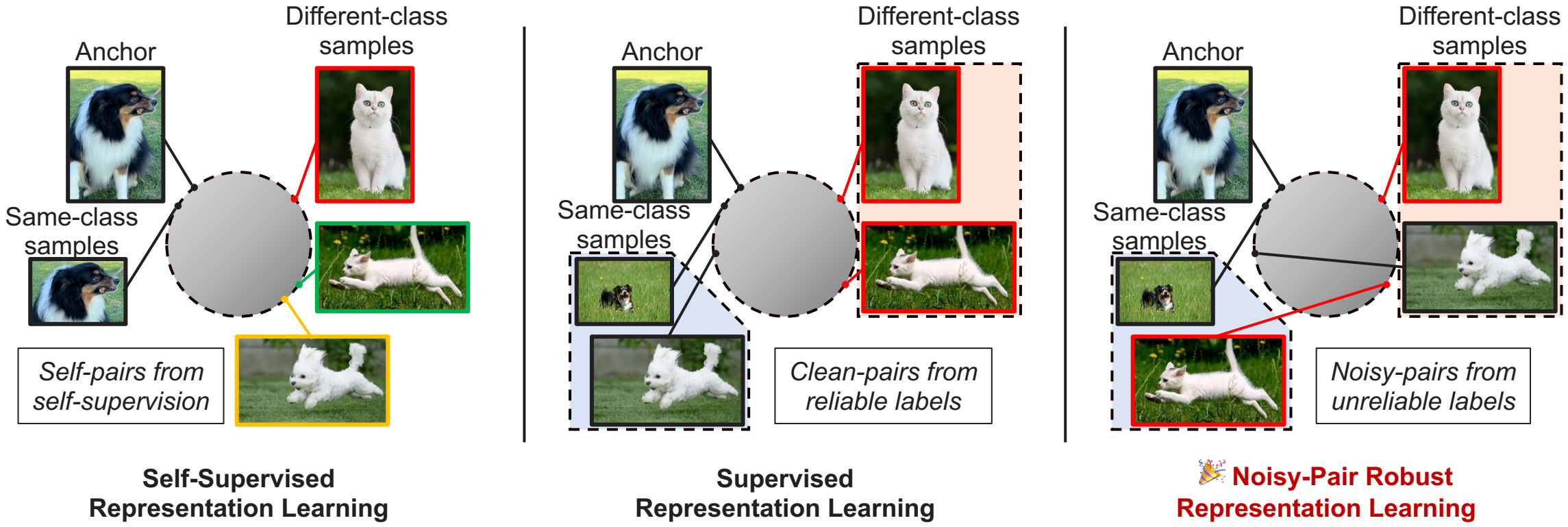


t-SNE visualizations of the representations on CIFAR-10 dataset

Challenge: How can discriminative representations be learned under unreliable supervision?

Kiryo R, Niu G, Du Plessis M C, et al. Positive-unlabeled learning with non-negative risk estimator[J]. Advances in neural information processing systems, 2017.
Xinrui W, Wan W, Geng C, et al. Beyond myopia: Learning from positive and unlabeled data through holistic predictive trends[J]. Advances in Neural Information Processing Systems, 2023.
Long L, Wang H, Jiang Z, et al. Positive-unlabeled learning by latent group-aware meta disambiguation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Noisy-Pair Robust Representation Learning



Self-Supervised Representation Learning

Supervised Representation Learning

Noisy-Pair Robust Representation Learning

Representation learning can acquire discriminative representations either by pulling same-class samples closer to the anchor and pushing different-class samples apart (contrastive representation learning), or by pulling same-class samples closer to the anchor (non-contrastive representation learning):

- Self-supervised representation learning: same-class pairs from augmented anchor
- Supervised representation learning: same-class pairs from reliable labels
- **Noisy-pair representation learning: same-class pairs from unreliable labels**

Analysis of Gradients

- Gradient magnitudes from noisy pairs overwhelm those from clean pairs in $\mathcal{L}_r$

❑ **In noisy-pair robust representation learning, the representations may have:**

Noisy pairs: $(\boldsymbol{x}_i, \boldsymbol{x}_m) \rightarrow \widetilde{\boldsymbol{q}}_i^{\mathrm{T}} \widetilde{\boldsymbol{q}}_m \approx 0$ Clean pairs: $(\boldsymbol{x}_i, \boldsymbol{x}_j) \rightarrow \widetilde{\boldsymbol{q}}_i^{\mathrm{T}} \widetilde{\boldsymbol{q}}_j \rightarrow 1$

➢ **Supervised Non-Contrastive Loss (*SNCL* $\mathcal{L}_r$)**

$$\mathcal{L}_r(\boldsymbol{x}_i, \boldsymbol{x}_j) = 2\left(1 - \langle \widetilde{\boldsymbol{q}}_i, \widetilde{\boldsymbol{k}}_j \rangle\right) \mathbb{1}\{y_i = y_j\} \Rightarrow \left\| \frac{\partial \mathcal{L}_r(\boldsymbol{x}_i, \boldsymbol{x}_m)}{\partial \boldsymbol{q}_i} \right\|_2^2 = \frac{4}{\|\boldsymbol{q}_i\|_2^2}\left(1 - (\widetilde{\boldsymbol{q}}_i^{\mathrm{T}} \widetilde{\boldsymbol{q}}_m)^2\right) > \frac{4}{\|\boldsymbol{q}_i\|_2^2}\left(1 - (\widetilde{\boldsymbol{q}}_i^{\mathrm{T}} \widetilde{\boldsymbol{q}}_j)^2\right) = \left\| \frac{\partial \mathcal{L}_r(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \boldsymbol{q}_i} \right\|_2^2$$

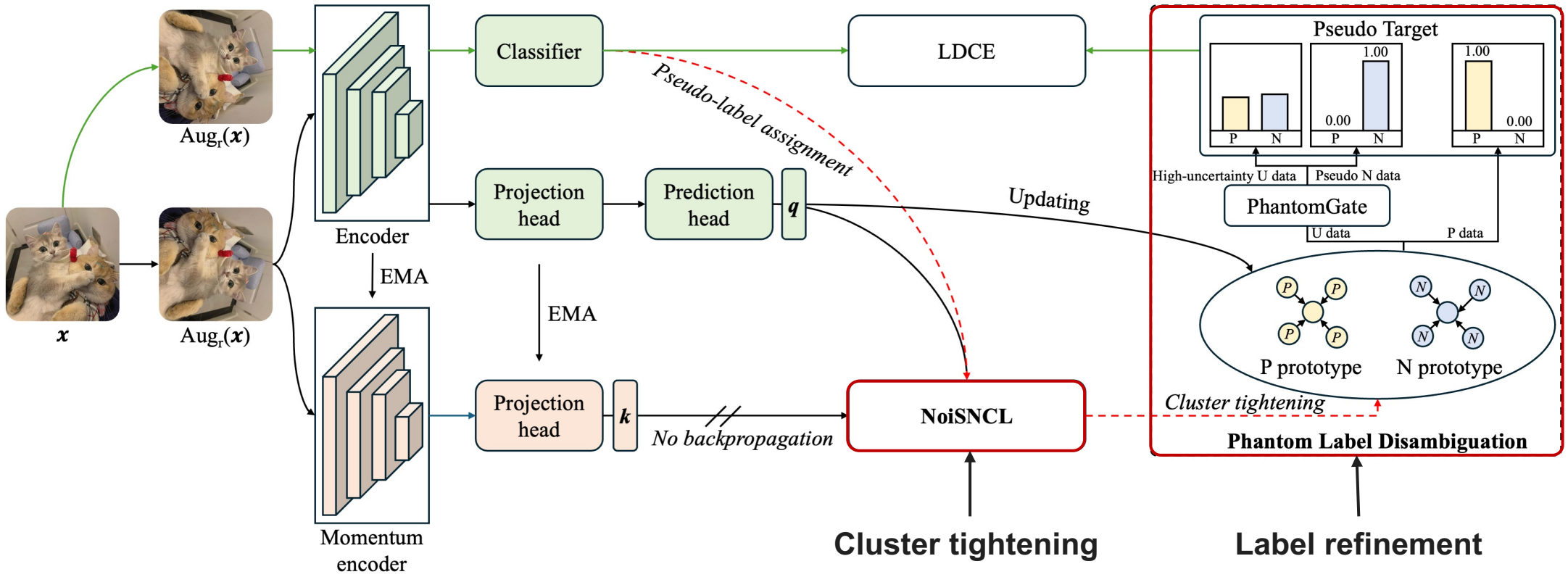
➢ **Noisy-Pair Robust Supervised Non-Contrastive Loss (*NoiSNCL* $\tilde{\mathcal{L}}_r$)**

$$\tilde{\mathcal{L}}_r(\boldsymbol{x}_i, \boldsymbol{x}_j) = 2\sqrt{1 - \langle \widetilde{\boldsymbol{q}}_i, \widetilde{\boldsymbol{k}}_j \rangle} \mathbb{1}\{y_i = y_j\} \Rightarrow \left\| \frac{\partial \tilde{\mathcal{L}}_r(\boldsymbol{x}_i, \boldsymbol{x}_m)}{\partial \boldsymbol{q}_i} \right\|_2^2 = \frac{1}{\|\boldsymbol{q}_i\|_2^2}\left(1 + (\widetilde{\boldsymbol{q}}_i^{\mathrm{T}} \widetilde{\boldsymbol{q}}_m)\right) < \frac{1}{\|\boldsymbol{q}_i\|_2^2}\left(1 + (\widetilde{\boldsymbol{q}}_i^{\mathrm{T}} \widetilde{\boldsymbol{q}}_j)\right) = \left\| \frac{\partial \tilde{\mathcal{L}}_r(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \boldsymbol{q}_i} \right\|_2^2$$

In $\mathcal{L}_{\mathrm{r}}$, the gradient magnitudes contributed by noisy pairs dominate those from clean pairs; this issue is effectively mitigated by the proposed $\tilde{\mathcal{L}}_{\mathrm{r}}$

Noisy-Pair Robust Non-Contrastive PU Learning

- Noisy-Pair Robust Representation Alignment Framework (*NcPU*)



- NoiSNCL: Learning discriminative representations under unreliable supervision
- Phantom Label Disambiguation (PLD): Refines supervision by regret-based label updating

Theoretical Analysis

- The NoiSNCL and PLD modules can be theoretically justified to iteratively benefit each other from the perspective of the EM framework

➢ **E-Step** (Label refinement). At the E-step, each unlabeled example is assigned to one specific cluster. Given a network $g(\cdot)$ parameterized by $\boldsymbol{\theta}$, the objective is to find $\boldsymbol{\theta}^*$ that maximizes the log-likelihood function:

$$\boldsymbol{\theta}^* = arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x}\in\mathcal{U}} \sum_{y\in\mathcal{Z}} \mathbb{1}(\tilde{y} =) \, y \log p(x, y|\boldsymbol{\theta})$$

➢ **M-Step** (Cluster tightening). At the M-step, minimizing $\tilde{\mathcal{L}}_{\boldsymbol{r}}$ encourages embeddings to concentrate around their cluster centers. Under some mild assumptions, minimizing $\tilde{\mathcal{R}}_r(\boldsymbol{x})$ is equivalent to maximizing a lower bound of the likelihood:

$$\tilde{\mathcal{R}}_r(\boldsymbol{x}) = \frac{1}{n_u} \sum_{\boldsymbol{x}\in\mathcal{U}} \frac{1}{|Q|} \sum_{k_+\in Q} \tilde{\mathcal{L}}_{\boldsymbol{r}}$$

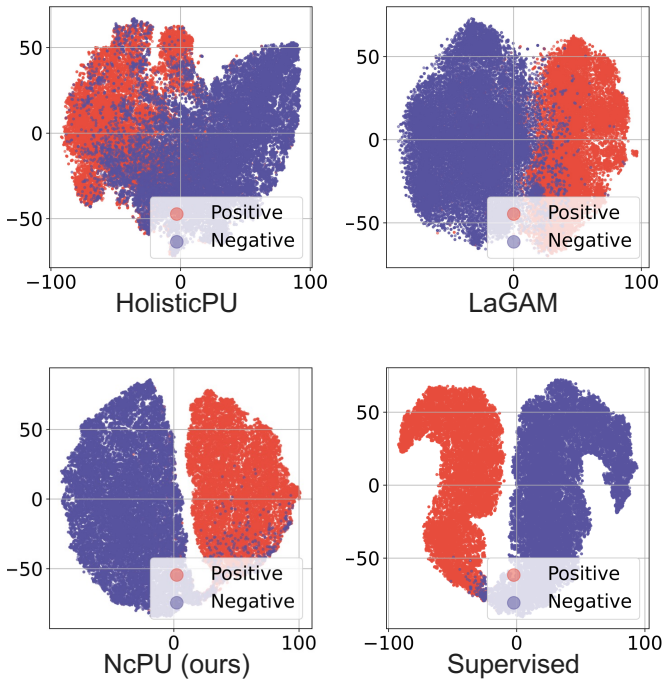
❖ **Theorem 1** Assume the distribution of each class in the representation space follows a d-variate von Mises-Fisher (vMF) distribution, which leads to: $h(\boldsymbol{x}|\tilde{\boldsymbol{\nu}}_{\boldsymbol{c}}, \kappa) = c_d(\kappa) e^{\kappa \tilde{\boldsymbol{\nu}}_{\boldsymbol{c}}^{\mathrm{T}} \tilde{g}(\boldsymbol{x})}$, where $\tilde{\boldsymbol{\nu}}_{\boldsymbol{c}} = \boldsymbol{\nu}_{\boldsymbol{c}}/\|\boldsymbol{\nu}_{\boldsymbol{c}}\|$, κ is the concentration parameter, and $c_d(\kappa)$ is the normalization constant. Under the assumption of a uniform class prior, optimizing $\tilde{\mathcal{R}}_r(\boldsymbol{x})$ and log-likelihood function is equivalent to maximizing L_1 and L_2 below, respectively.

$$L_1 = \sum_{\mathcal{S}_c\in\mathcal{U}} \frac{|\mathcal{S}_c|}{n_u} \|\boldsymbol{\nu}_{\boldsymbol{c}}\|^2 \leq \sum_{\mathcal{S}_c\in\mathcal{U}} \frac{|\mathcal{S}_c|}{n_u} \|\boldsymbol{\nu}_{\boldsymbol{c}}\| = L_2$$

Experiments

Main results

Method	Additional N Data	CIFAR-10		CIFAR-100		STL-10		ABCD		xBD	
		OA	F1	OA	F1	OA	F1	OA	F1	OA	F1
CE		$60.45^{\pm0.1}$	$2.42^{\pm0.4}$	$50.36^{\pm0.0}$	$1.86^{\pm0.2}$	$50.30^{\pm0.0}$	$1.19^{\pm0.2}$	$55.70^{\pm0.2}$	$20.93^{\pm0.4}$	$84.08^{\pm0.2}$	$25.70^{\pm2.4}$
uPU		$65.52^{\pm0.2}$	$26.82^{\pm0.9}$	$61.44^{\pm0.9}$	$43.12^{\pm2.1}$	$57.08^{\pm0.4}$	$25.88^{\pm1.3}$	$83.76^{\pm2.1}$	$81.47^{\pm2.9}$	$86.82^{\pm0.3}$	$55.43^{\pm2.8}$
nnPU		$87.29^{\pm0.5}$	$83.71^{\pm0.6}$	$72.00^{\pm0.8}$	$74.93^{\pm0.4}$	$80.62^{\pm0.1}$	$79.28^{\pm0.2}$	$87.73^{\pm0.4}$	$88.36^{\pm0.3}$	$82.60^{\pm0.6}$	$59.66^{\pm0.7}$
vPU		$85.94^{\pm0.6}$	$82.98^{\pm0.9}$	$69.01^{\pm1.2}$	$70.78^{\pm0.2}$	$75.76^{\pm5.5}$	$70.52^{\pm10.4}$	$84.06^{\pm3.0}$	$84.13^{\pm3.4}$	$73.60^{\pm1.8}$	$50.30^{\pm1.2}$
ImbPU		$87.29^{\pm0.4}$	$83.80^{\pm0.4}$	$72.07^{\pm0.7}$	$75.05^{\pm0.6}$	$80.68^{\pm0.6}$	$79.41^{\pm0.6}$	$88.14^{\pm0.6}$	$88.69^{\pm0.5}$	$82.51^{\pm0.5}$	$59.72^{\pm0.5}$
TEDn		$86.29^{\pm2.4}$	$80.70^{\pm4.6}$	$69.85^{\pm0.9}$	$61.73^{\pm1.9}$	$66.26^{\pm4.9}$	$49.90^{\pm10.7}$	$88.90^{\pm0.9}$	$89.10^{\pm0.9}$	$85.40^{\pm0.8}$	$52.65^{\pm4.6}$
PUET		$78.51^{\pm0.4}$	$73.85^{\pm0.5}$	$62.81^{\pm0.2}$	$71.09^{\pm0.1}$	$75.36^{\pm0.2}$	$73.56^{\pm0.1}$	$78.09^{\pm2.9}$	$66.52^{\pm24.9}$	$74.92^{\pm0.1}$	$38.38^{\pm0.6}$
HolisticPU		$84.20^{\pm2.1}$	$78.10^{\pm3.9}$	$64.01^{\pm6.5}$	$51.94^{\pm15.1}$	$72.81^{\pm6.4}$	$66.06^{\pm14.9}$	$65.49^{\pm1.5}$	$51.60^{\pm1.5}$	$81.98^{\pm4.1}$	$53.35^{\pm2.4}$
DistPU		$85.29^{\pm2.6}$	$83.96^{\pm2.2}$	$67.63^{\pm0.8}$	$73.68^{\pm0.8}$	$85.62^{\pm1.5}$	$85.41^{\pm0.9}$	$86.25^{\pm1.7}$	$87.36^{\pm1.2}$	$82.94^{\pm0.8}$	$57.58^{\pm0.2}$
PiCO		$89.72^{\pm0.1}$	$87.40^{\pm0.0}$	$69.98^{\pm0.4}$	$72.71^{\pm0.3}$	$60.71^{\pm0.6}$	$71.04^{\pm0.3}$	$74.07^{\pm2.2}$	$79.27^{\pm1.3}$	$49.36^{\pm0.5}$	$39.52^{\pm0.2}$
LaGAM	✓	$95.78^{\pm0.5}$	$94.90^{\pm0.6}$	$84.82^{\pm0.1}$	$84.42^{\pm0.2}$	$88.64^{\pm0.0}$	$88.50^{\pm0.1}$	$75.90^{\pm0.4}$	$75.38^{\pm0.6}$	$79.14^{\pm1.5}$	$58.78^{\pm1.7}$
WSC		$90.55^{\pm0.3}$	$87.92^{\pm0.8}$	$75.39^{\pm2.1}$	$73.76^{\pm4.0}$	$79.06^{\pm4.5}$	$74.16^{\pm7.0}$	$80.10^{\pm2.8}$	$76.12^{\pm4.3}$	$84.89^{\pm0.8}$	$62.17^{\pm1.3}$
NcPU(ours)		$97.36^{\pm0.1}$	$96.67^{\pm0.2}$	$88.28^{\pm0.6}$	$88.14^{\pm0.9}$	$91.40^{\pm0.4}$	$90.82^{\pm0.6}$	$91.10^{\pm0.6}$	$91.21^{\pm0.5}$	$87.60^{\pm1.0}$	$64.84^{\pm1.0}$
Supervised	✓	$96.96^{\pm0.2}$	$96.24^{\pm0.2}$	$89.65^{\pm0.3}$	$89.78^{\pm0.4}$	—	—	$92.00^{\pm0.2}$	$91.96^{\pm0.2}$	$88.47^{\pm0.3}$	$73.32^{\pm0.4}$



t-SNE visualizations of the representations on CIFAR-10 dataset

NcPU outperforms SOTA methods, achieving performance comparable to its supervised counterpart

Experiments

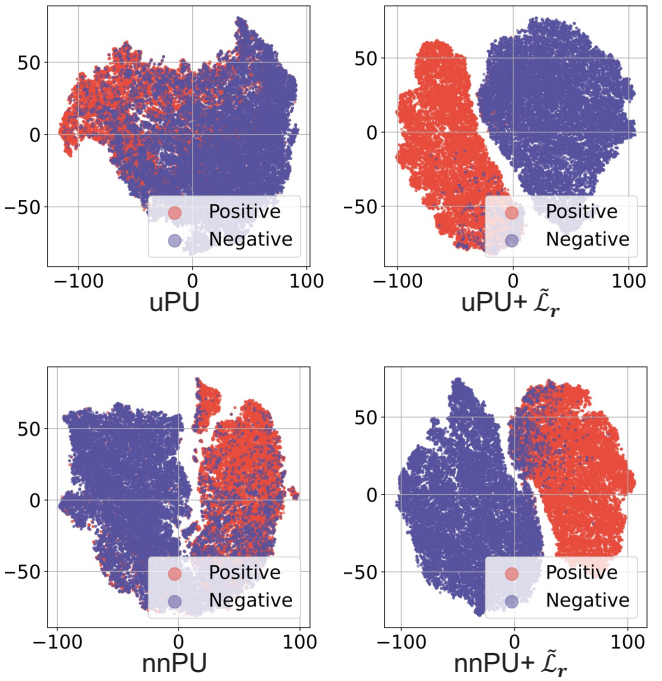
■ Ablation studies and analyses

❖ Ablation and comparative analyses on CIFAR-100 dataset

NCL	LD	OA	F1	P	R
	s	$61.54^{\pm 7.8}$	$40.58^{\pm 22.9}$	$84.36^{\pm 7.1}$	$30.69^{\pm 25.0}$
$\tilde{\mathcal{L}}_{\mathrm{r}}$		$50.27^{\pm 0.1}$	$1.09^{\pm 0.4}$	$97.97^{\pm 1.8}$	$0.55^{\pm 0.2}$
$\mathcal{L}_{\text{self-r}}$	s	$73.22^{\pm 1.7}$	$72.75^{\pm 1.6}$	$74.10^{\pm 2.2}$	$71.47^{\pm 1.9}$
$\mathcal{L}_{\mathrm{r}}$	s	$84.58^{\pm 0.8}$	$85.90^{\pm 0.6}$	$79.12^{\pm 1.1}$	$93.96^{\pm 0.3}$
$\tilde{\mathcal{L}}_{\mathrm{r}}$	s'	$75.14^{\pm 2.7}$	$79.91^{\pm 1.7}$	$67.15^{\pm 2.6}$	$98.73^{\pm 0.5}$
$\tilde{\mathcal{L}}_{\mathrm{r}}$	s'+SAT	$50.25^{\pm 0.0}$	$1.01^{\pm 0.1}$	$97.85^{\pm 3.7}$	$0.51^{\pm 0.1}$
$\tilde{\mathcal{L}}_{\mathrm{r}}$	s	$88.28^{\pm 0.6}$	$88.14^{\pm 0.9}$	$89.12^{\pm 1.7}$	$87.27^{\pm 3.2}$

❖ Performance analysis of $\tilde{\mathcal{L}}_r$ under risk estimation with the real π_p and supervised methods

Method	CIFAR-10		CIFAR-100	
	OA	F1	OA	F1
uPU	$69.43^{\pm 0.3}$	$41.32^{\pm 1.0}$	$61.68^{\pm 0.9}$	$44.18^{\pm 2.6}$
uPU+$\tilde{\mathcal{L}}_{\mathrm{r}}$	$97.35^{\pm 0.1}$	$96.66^{\pm 0.2}$	$83.71^{\pm 1.6}$	$81.40^{\pm 2.2}$
nnPU	$83.25^{\pm 0.2}$	$76.94^{\pm 0.5}$	$71.22^{\pm 0.5}$	$68.12^{\pm 1.0}$
nnPU+$\tilde{\mathcal{L}}_{\mathrm{r}}$	$97.03^{\pm 0.2}$	$96.37^{\pm 0.2}$	$87.81^{\pm 0.3}$	$87.23^{\pm 0.4}$
Supervised+$\mathcal{L}_{\mathrm{r}}$	$98.53^{\pm 0.0}$	$98.17^{\pm 0.1}$	$94.45^{\pm 0.1}$	$94.52^{\pm 0.1}$
Supervised+$\tilde{\mathcal{L}}_{\mathrm{r}}$	$98.75^{\pm 0.0}$	$98.43^{\pm 0.1}$	$94.56^{\pm 0.1}$	$94.64^{\pm 0.1}$

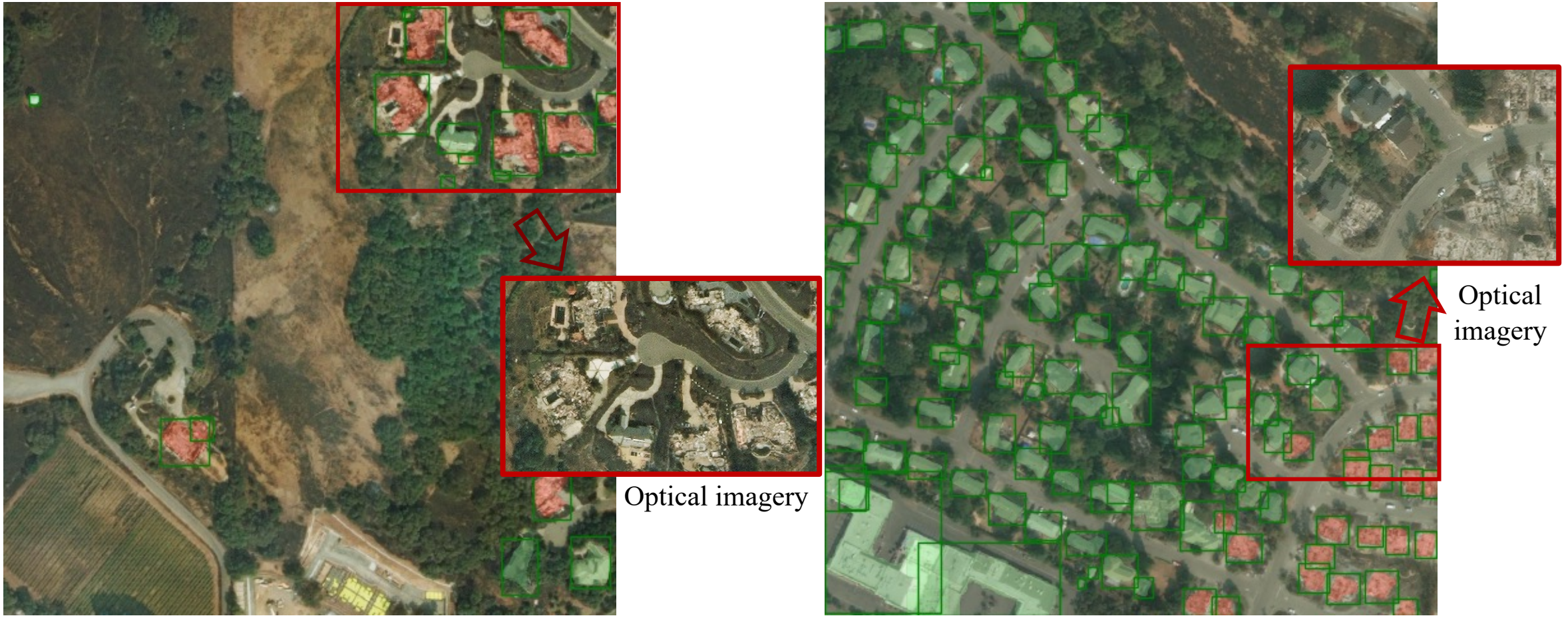


t-SNE visualizations of the representations on CIFAR-10 dataset

$\tilde{\mathcal{L}}_r$ enhances the quality of learned representations, enabling simple PU methods to achieve competitive performance

Application

- Building damage mapping from satellite imagery



Thank you for your attention!

Homepage

Code