NEURAL INFORMATION PROCESSING SYSTEMS

上海科技大学
ShanghaiTech University

CSOT: Curriculum and Structure-Aware Optimal Transport for Learning with Noisy Labels



Wanxing Chang[1] **Ye Shi[1, 2]** **Jingya Wang[1, 2*]**

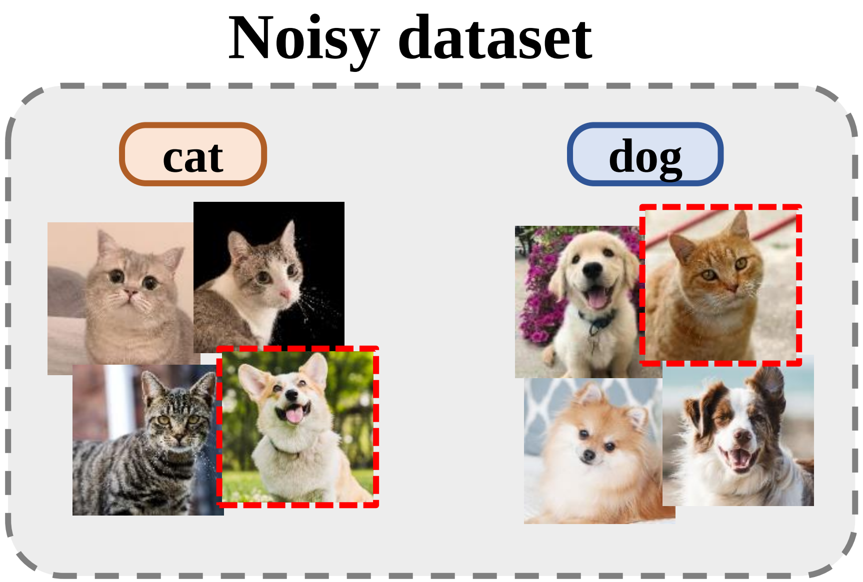
[1]ShanghaiTech University
[2]Engineering Research Center of Intelligent Vision and Imaging

NeurIPS 2023
Nov 12, 2023

Project page: https://changwxx.github.io/CSOT-webpage/

立志成才 报国裕民

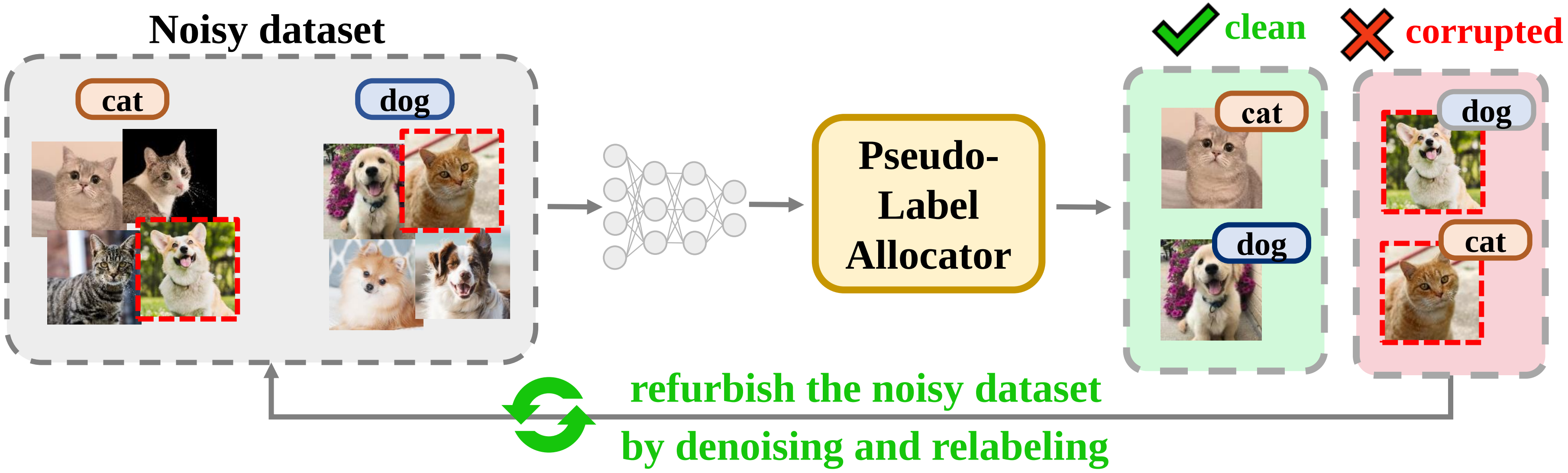
Learning with Noisy Labels



Learning with Noisy Labels (LNL) aims to train a ***classification*** *network that is* ***robust to corrupted labels*** *and achieves high accuracy on a clean test set.*

Denoising and Relabeling for Noisy Dataset

*A straight-forward strategy is to **identify clean labels** and **correct corrupted labels** in the noisy dataset, then **refurbish them to be clean**.*



We need a robust pseudo-label allocator!

How to Construct Robust Pseudo-Labeling?

上海科技大学 ShanghaiTech University

Existing Prediction-based PL

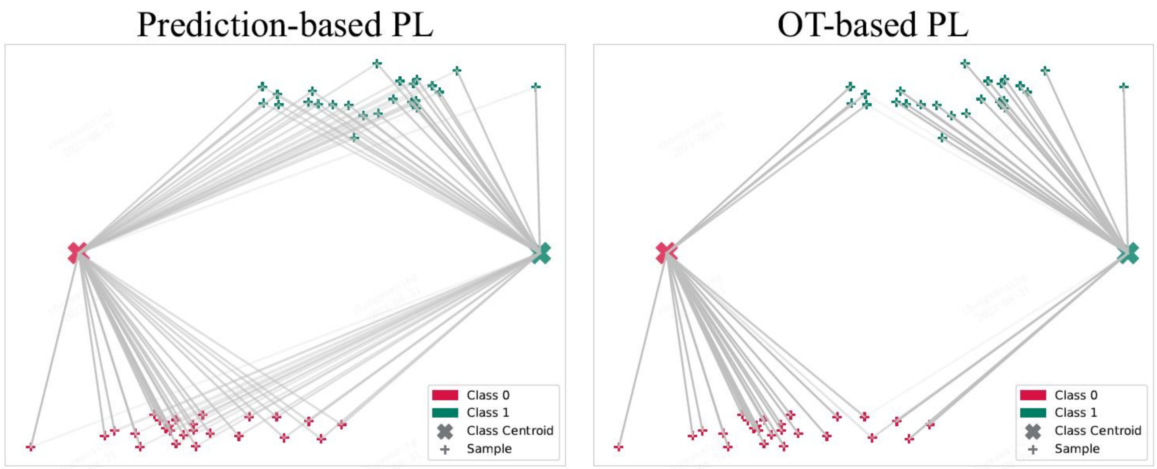
× evaluate **each sample independently**

× heavily rely on the **unreliable model**'s prediction

Optimal Transport (OT) -based PL

− further consider the **inter-distribution** structure of the samples and categories distribution

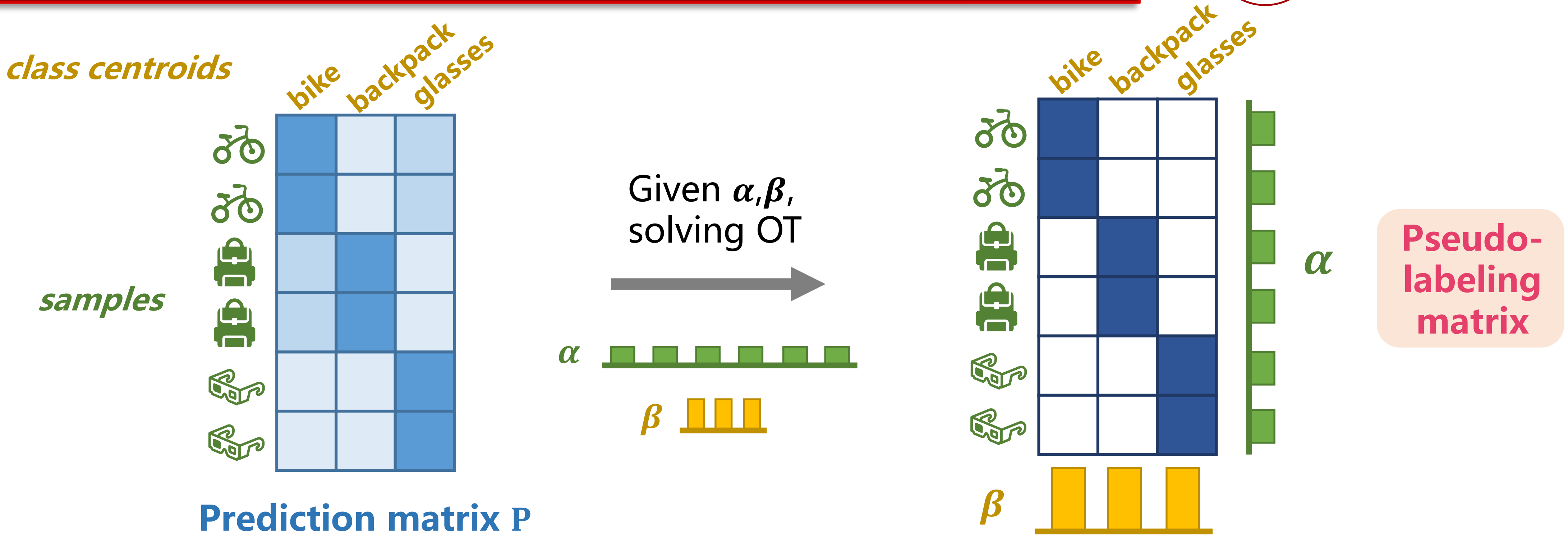
− produce pseudo labels with **global discriminability**



✓ **global discriminability**

立志成才 报国裕民

Optimal Transport (OT) -based PL



Prediction matrix $\mathbf{P}$

Coupling matrix $\mathbf{Q}^*$

Optimal Transport-based PL

$$\min_{\mathbf{Q}\in\Pi(\frac{1}{B}\mathbb{1}_B,\frac{1}{C}\mathbb{1}_C)} \langle -\log\mathbf{P}\,,\mathbf{Q}\rangle + \varepsilon\langle \mathbf{Q},\log\mathbf{Q}\rangle$$

$$\boldsymbol{\Pi}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \left\{\mathbf{Q}\in\mathbb{R}_+^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|}\,\middle|\,\mathbf{Q}\mathbb{1}_{|\boldsymbol{\beta}|}=\boldsymbol{\alpha},\ \mathbf{Q}^{\mathrm{T}}\mathbb{1}_{|\boldsymbol{\alpha}|}=\boldsymbol{\beta}\right\}$$

OT-based PL is not good enough

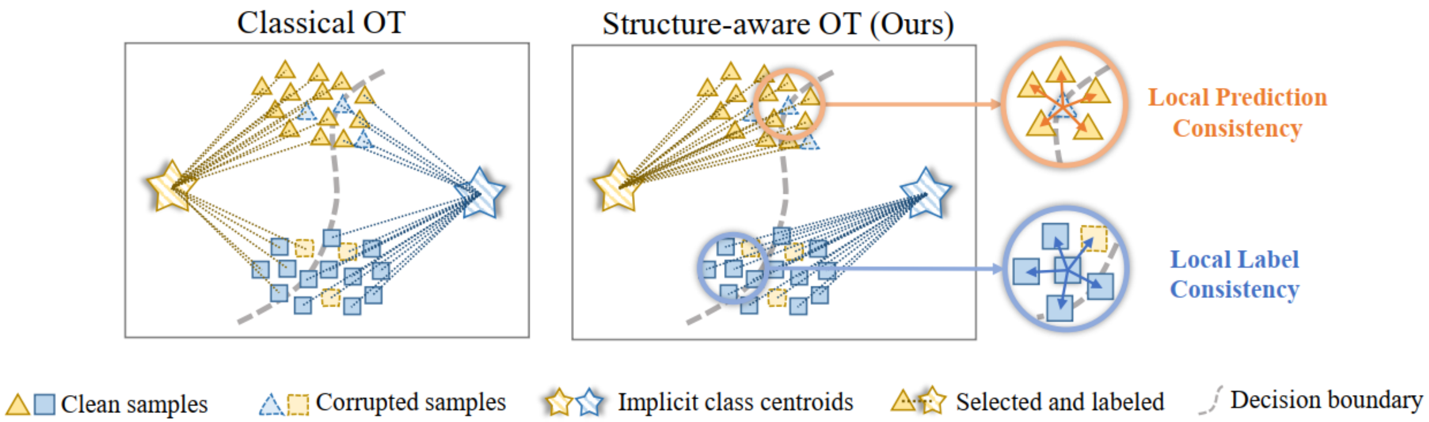
When the decision boundary is not accurate enough…

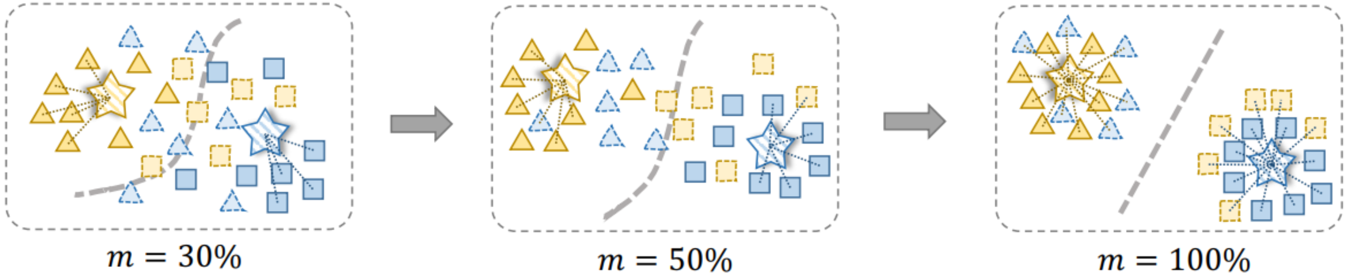
OT-based PL

- × tends to **mismatch two nearby samples to two far-away class** centroids
- × **fully** assign **inaccurate** pseudo labels

Ours (CSOT-based PL)

- ✓ generates **local consensus** assignments for each sample
- ✓ **partially** assign **top-reliable** labels controlled by **budget factor** m





How to Construct Robust Pseudo-Labeling?

Existing Prediction-based PL

× evaluate **each sample independently**

× heavily on the **unreliable model**'s prediction

Optimal Transport (OT) -based PL

− further consider the **inter-distribution** structure of the samples and categories distribution

− produce pseudo labels with **global discriminability**

Ours (CSOT-based PL)

✓ fully consider both **inter-** and **intra-distribution** structure of the samples

✓ produce robust pseudo labels with both **global discriminability** and **local coherence**

✓ **incrementally assigns** reliable labels to **a fraction of** the samples with the highest confidence

Curriculum and Structure-Aware OT

上海科技大学
ShanghaiTech University

(๑•̀ㅂ•́)و✧

Structure-Aware OT

- ✓ inter-distribution
- ✓ intra-distribution
- ✓ global discriminability
- ✓ local coherence

Curriculum and Structure-Aware OT

- ✓ prioritize samples with better global and local properties for robust label assignment!

Curriculum OT

- ✓ incremental assignment of reliable labels
- ✓ enable a curriculum pseudo-label allocator

CSOT

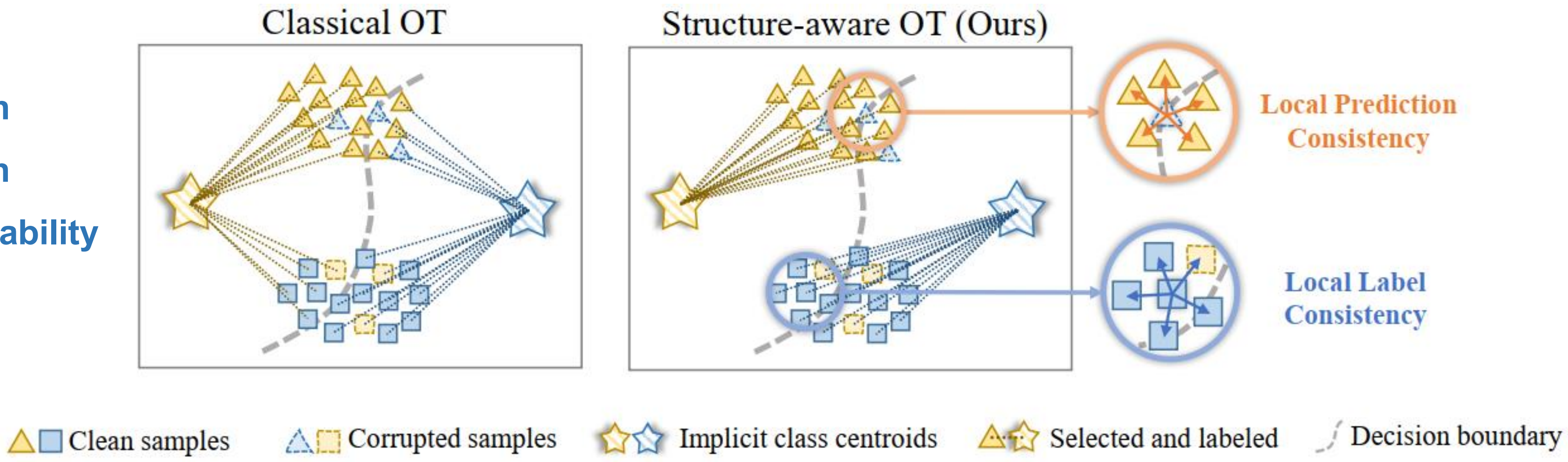
$$\min_{\mathbf{Q}\in\boxed{\Pi^{C}(\boldsymbol{\alpha},\boldsymbol{\beta})}} \langle \mathbf{C}, \mathbf{Q}\rangle + \kappa\boxed{\Omega(\mathbf{Q})} + \varepsilon\langle \mathbf{Q}, \log \mathbf{Q}\rangle \quad , \text{ where } \mathbf{C} = -\log \mathbf{P}$$

立志成才报国裕民

Structure-Aware OT

上海科技大学
ShanghaiTech University

- ✓ inter-distribution
- ✓ intra-distribution
- ✓ global discriminability
- ✓ local coherence



Structure-Aware OT

$$\min_{\mathbf{Q}\in\Pi(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle \mathbf{C}, \mathbf{Q}\rangle + \kappa\Omega(\mathbf{Q}) + \varepsilon\langle \mathbf{Q}, \log \mathbf{Q}\rangle, \quad \text{where } \mathbf{C} = -\log \mathbf{P}$$

Structure-Aware Regularization Terms

$$\Omega^{\mathrm{P}}(\mathbf{Q}) = -\sum_{i,j} \mathbf{S}_{ij} \sum_{k} \mathbf{P}_{ik}\mathbf{P}_{jk}\mathbf{Q}_{ik}\mathbf{Q}_{jk} = -\left\langle \mathbf{S}, (\mathbf{P}\odot\mathbf{Q})(\mathbf{P}\odot\mathbf{Q})^{\mathrm{T}} \right\rangle$$

$$\Omega^{\mathrm{L}}(\mathbf{Q}) = -\sum_{i,j} \mathbf{S}_{ij} \sum_{k} \mathbf{L}_{ik}\mathbf{L}_{jk}\mathbf{Q}_{ik}\mathbf{Q}_{jk} = -\left\langle \mathbf{S}, (\mathbf{L}\odot\mathbf{Q})(\mathbf{L}\odot\mathbf{Q})^{\mathrm{T}} \right\rangle$$

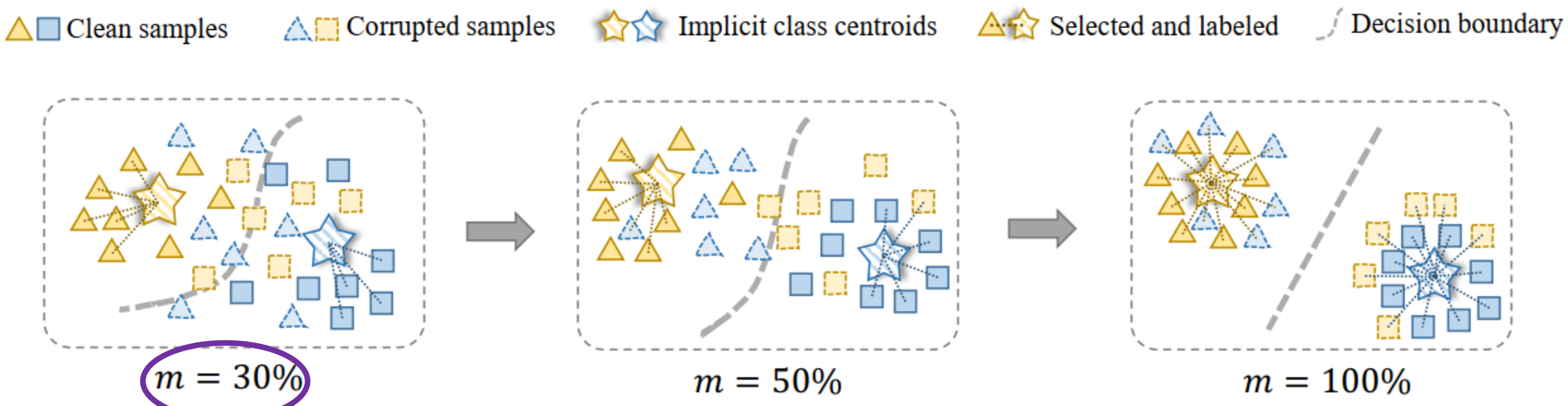
***Notations**

- $\mathbf{S} \in \mathbb{R}^{B\times B}$ is samples similarity matrix
- $\mathbf{P} \in \mathbb{R}_{+}^{B\times C}$ is softmax prediction matrix
- $\mathbf{L} \in \mathbb{R}_{+}^{B\times C}$ is one-hot label matrix

立志成才 报国裕民

Curriculum OT

- ✓ incremental assignment of reliable labels
- ✓ enable a curriculum pseudo-label allocator



top 30% reliable labels are selected!

Curriculum OT

$$\min_{\mathbf{Q}\in\Pi^{c}(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle \mathbf{C}, \mathbf{Q}\rangle + \varepsilon\langle \mathbf{Q}, \log \mathbf{Q}\rangle \;, \;\text{ where } \mathbf{C} = -\log \mathbf{P}$$

Curriculum Constraints

$$\boldsymbol{\Pi}^{c}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \left\{\mathbf{Q}\in\mathbb{R}_{+}^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|}\middle|\mathbf{Q}\mathbb{1}_{|\boldsymbol{\beta}|}\le\boldsymbol{\alpha},\; \mathbf{Q}^{\mathrm{T}}\mathbb{1}_{|\boldsymbol{\alpha}|}=\boldsymbol{\beta}\right\},$$

where $\boldsymbol{\alpha} = \frac{1}{B}\mathbb{1}_{B},\quad \boldsymbol{\beta} = \frac{m}{C}\mathbb{1}_{C},\quad m\in[0,1]$ is a curriculum budget factor

A Lightspeed Solver for CSOT

上海科技大学
ShanghaiTech University

Algorithm 1 Efficient scaling iteration for entropic regularized Curriculum OT

Input: Cost matrix $\mathbf{C}$, marginal constraints vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, entropic regularization weight ε
Initialize: $\mathbf{K} \leftarrow e^{-\mathbf{C}/\varepsilon}$, $\boldsymbol{v}^{(0)} \leftarrow \mathbb{1}_{|\boldsymbol{\beta}|}$
Compute: $\mathbf{K}_{\boldsymbol{\alpha}} \leftarrow \frac{\mathbf{K}}{\mathrm{diag}(\boldsymbol{\alpha})\mathbb{1}_{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|}}$, $\mathbf{K}_{\boldsymbol{\beta}}^{\top} \leftarrow \frac{\mathbf{K}^{\top}}{\mathrm{diag}(\boldsymbol{\beta})\mathbb{1}_{|\boldsymbol{\beta}|\times|\boldsymbol{\alpha}|}}$ `// Saving computation`
for $n = 1, 2, 3, \ldots$ **do**
  $\boldsymbol{u}^{(n)} \leftarrow \min\left(\frac{\mathbb{1}_{|\boldsymbol{\alpha}|}}{\mathbf{K}_{\boldsymbol{\alpha}}\boldsymbol{v}^{(n-1)}}, \mathbb{1}_{|\boldsymbol{\alpha}|}\right)$
  $\boldsymbol{v}^{(n)} \leftarrow \frac{\mathbb{1}_{|\boldsymbol{\beta}|}}{\mathbf{K}_{\boldsymbol{\beta}}^{\top}\boldsymbol{u}^{(n)}}$
end for
Return: $\mathrm{diag}(\boldsymbol{u}^{(n)})\mathbf{K}\mathrm{diag}(\boldsymbol{v}^{(n)})$

Algorithm 2 Generalized conditional gradient algorithm for entropic regularized CSOT

Input: Cost matrix $\mathbf{C}$, marginal constraints vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, entropic regularization weight ε, local coherent regularization weight κ, local coherent regularization function $\Omega : \mathbb{R}^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|} \rightarrow \mathbb{R}$, and its gradient function $\nabla\Omega : \mathbb{R}^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{|\boldsymbol{\alpha}|\times|\boldsymbol{\beta}|}$
Initialize: $\mathbf{Q}^{(0)} \leftarrow \boldsymbol{\alpha}\boldsymbol{\beta}^{T}$
for $i = 1, 2, 3, \ldots$ **do**
  $\mathbf{G}^{(i)} \leftarrow \mathbf{Q}^{(i)} + \kappa\nabla\Omega(\mathbf{Q}^{(i)})$ `// Gradient computation`
  $\widetilde{\mathbf{Q}}^{(i)} \leftarrow \mathrm{argmin}_{\mathbf{Q}\in\mathbf{\Pi}^{c}(\boldsymbol{\alpha},\boldsymbol{\beta})} \left\langle \mathbf{Q}, \mathbf{G}^{(i)} \right\rangle + \varepsilon \left\langle \mathbf{Q}, \log \mathbf{Q} \right\rangle$
  `// Linearization, solved efficiently by Algorithm`
  Choose $\eta^{(i)} \in [0, 1]$ so that it satisfies the Armijo rule `// Backtracking line-search`
  $\mathbf{Q}^{(i+1)} \leftarrow \left(1 - \eta^{(i)}\right) \mathbf{Q}^{(i)} + \eta^{(i)}\widetilde{\mathbf{Q}}^{(i)}$ `// Update`
end for
Return: $\mathbf{Q}^{(i)}$

Table 3: **Time cost (s) for solving CSOT optimization problem of different input sizes.** VDA indicates vanilla Dykstras algorithm-based CSOT solver, while ESI indicates the efficient scaling iteration-based solver.

$(\lvert\boldsymbol{\alpha}\rvert, \lvert\boldsymbol{\beta}\rvert)$	VDA-based	ESI-based (Ours)
(1024,10)	0.83	**0.82** ↓
(1024,50)	1.00	**0.80** ↓
(1024,100)	0.87	**0.80** ↓
(50,50)	0.82	**0.79** ↓
(100,100)	0.88	**0.80** ↓
(500,500)	0.88	**0.87** ↓
(1000,1000)	0.94	**0.81** ↓
(2000,2000)	2.11	**0.98** ↓
(3000,3000)	3.74	**0.99** ↓

x3.7 faster

立志成才 报国裕民

Comparison with state-of-the-arts

Table 1: **Comparison with state-of-the-art methods in test accuracy (%) on CIFAR-10 and CIFAR-100.** The results are mainly copied from [45, 48]. We present the performance of our CSOT method using the "mean±variance" format, which is obtained from 3 trials with different seeds.

Dataset	CIFAR-10					CIFAR-100			
Noise type	Symmetric				Assymetric	Symmetric			
Method/Noise ratio	0.2	0.5	0.8	0.9	0.4	0.2	0.5	0.8	0.9
Cross-Entropy	86.8	79.4	62.9	42.7	85.0	62.0	46.7	19.9	10.1
F-correction [56]	86.8	79.8	63.3	42.9	87.2	61.5	46.6	19.9	10.2
Co-teaching+ [31]	89.5	85.7	67.4	47.9	-	65.6	51.8	27.9	13.7
PENCIL [76]	92.4	89.1	77.5	58.9	88.5	69.4	57.5	31.1	15.3
DivideMix [46]	96.1	94.6	93.2	76.0	93.4	77.3	74.6	60.2	31.5
ELR [50]	95.8	94.8	93.3	78.7	93.0	77.6	73.6	60.8	33.4
NGC [72]	95.9	94.5	91.6	80.5	90.6	79.3	75.9	62.7	29.8
RRL [48]	96.4	95.3	93.3	77.4	92.6	80.3	76.0	61.1	33.1
MOIT [55]	93.1	90.0	79.0	69.6	92.0	73.0	64.6	46.5	36.0
UniCon [41]	96.0	95.6	93.9	**90.8**	94.1	78.9	77.6	63.9	44.8
NCE [45]	96.2	95.3	93.9	88.4	94.5	**81.4**	76.3	64.7	41.1
OT Cleaner [73]	91.4	85.4	56.9	-	-	67.4	58.9	31.2	-
OT-Filter [23]	96.0	95.3	94.0	90.5	95.1	76.7	73.8	61.8	42.8
CSOT (Best)	**96.6±0.10**	**96.2±0.11**	**94.4±0.16**	90.7±0.33	**95.5±0.06**	80.5±0.28	**77.9±0.18**	**67.8±0.23**	**50.5±0.46**
CSOT (Last)	96.4±0.18	96.0±0.11	94.3±0.20	90.5±0.36	95.2±0.12	80.2±0.31	77.7±0.14	67.6±0.36	50.3±0.33

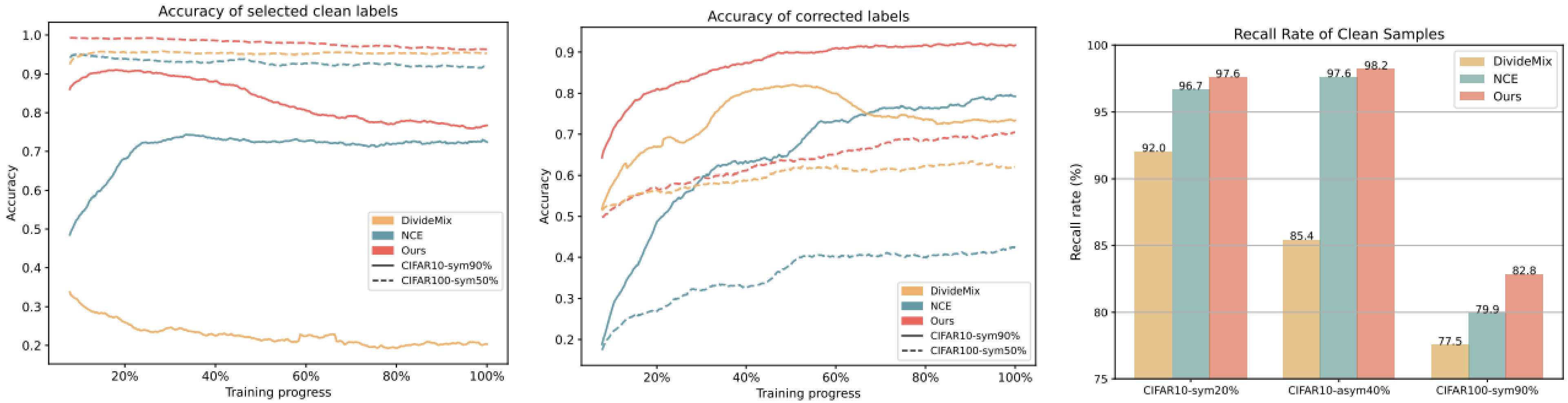
+9%!

Comparison with state-of-the-arts

Table 2: **Comparison with state-of-the-art methods in top-1 / 5 test accuracy (%) on the Webvision and ImageNet ILSVRC12 validation sets.** The models are trained on the training set of the Webvision dataset.

Method	Webvision top-1	Webvision top-5	ILSVRC12 top-1	ILSVRC12 top-5
F-correction [46]	61.12	82.68	57.36	82.36
Decoupling [42]	62.54	84.74	58.26	82.26
MentorNet [31]	63.00	81.40	57.80	79.92
Co-teaching [24]	63.58	85.20	61.48	84.70
DivideMix [37]	77.32	91.64	75.20	90.84
ELR [40]	76.26	91.26	68.71	87.84
ELR+ [40]	77.78	91.68	70.29	89.76
NGC [60]	79.20	91.80	74.40	91.00
RRL [38]	77.80	91.30	74.40	90.90
MOIT [45]	77.90	91.90	73.80	91.70
NCE [36]	79.50	**93.80**	76.30	**94.10**
CSOT	**79.67**	91.95	**76.64**	91.67

Comparison with state-of-the-arts



(a) Clean accuracy (b) Corrected accuracy (c) Clean recall rate

Figure 2: **Performance comparison for clean label identification and corrupted label correction.**

Ablation study

上海科技大学
ShanghaiTech University

Table 3: Ablation studies under multiple label noise ratios on CIFAR-10 and CIFAR-100. "repl." is an abbreviation for "replaced", and L^{ce} represents a cross-entropy loss. GMM refers to the selection of clean labels based on small-loss criterion [37]. CT (confidence thresholding [52]) is a relabeling scheme where we set the CT value to 0.95.

	Dataset	CIFAR-10				CIFAR-100			
	Noise type	Sym.			Asym.	Sym.			Avg
	Method/Noise ratio	0.5	0.8	0.9	0.4	0.5	0.8	0.9	
Denoise Relabeling Technique	(a) Classical OT	95.45	91.95	82.35	95.04	75.96	62.46	43.28	78.07
	(b) Structure-aware OT	95.86	91.87	83.29	95.06	76.20	63.73	44.57	78.65
	(c) CSOT w/o Ω^P and Ω^L	95.53	93.84	89.50	95.14	75.96	66.50	47.55	80.57
	(d) CSOT w/o Ω^P	95.77	94.08	89.97	95.35	76.09	66.79	48.13	80.88
	(e) CSOT w/o Ω^L	95.55	93.97	90.41	95.15	76.17	67.28	48.01	80.93
Learning Technique	(f) GMM + L^{sup}	92.48	80.37	31.76	90.80	69.52	48.49	20.86	62.04
	(g) CSOT repl. L^{sup} with L^{ce}	93.47	81.93	53.45	91.43	72.66	50.62	21.77	66.48
	(h) CSOT w/o L^{semi}	95.34	93.04	88.9	94.11	75.16	61.13	36.94	77.80
	(i) CSOT repl. correction with CT (0.95)	95.46	90.73	89.09	95.21	75.85	64.28	48.76	79.91
	(j) CSOT w/o $L^{simsiam}_{D_{corrupted}}$	95.92	94.17	89.31	95.16	76.38	66.17	45.56	80.38
	CSOT	96.20	94.39	90.65	95.50	77.94	67.78	50.50	81.85

立志成才 报国裕民

Take-home message

上海科技大学
ShanghaiTech University

Structure-Aware OT

- ✓ inter-distribution
- ✓ intra-distribution
- ✓ global discriminability
- ✓ local coherence

Curriculum and Structure-Aware OT

- ✓ prioritize samples with better global and local properties for robust label assignment!

Curriculum OT

- ✓ incremental assignment of reliable labels
- ✓ enable a curriculum pseudo-label allocator

scan QR code for more details (code, poster, slides...)