Robust Curriculum Learning: from clean-data detection to noisy-label self-correction

Tianyi Zhou*, Shengjie Wang*, Jeff A. Bilmes
University *of* Washington, Seattle

MELODI
MachinE Learning, Optimization, & Data Interpretation @ UW

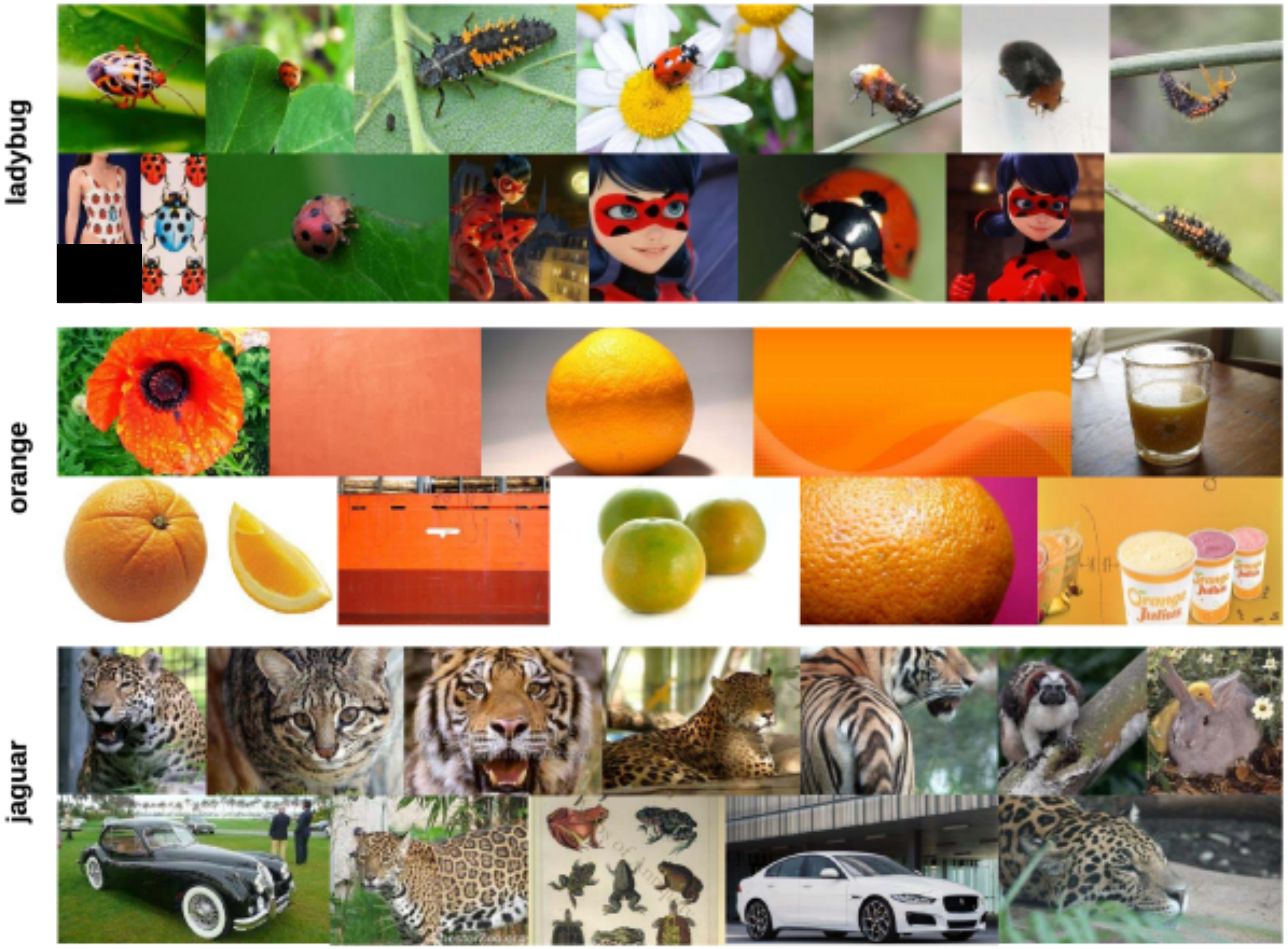
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY *of* WASHINGTON

Two Main Challenges in Noisy-label Learning

- **Noisy labels are not uncommon** in data collected in practice, e.g., web-tags.
- **Challenge 1: clean data detection + supervised learning**.
- **Challenge 2: noisy label correction (pseudo-labels) + self-supervised learning.**

Noisy Training Examples in the Webvision Dataset

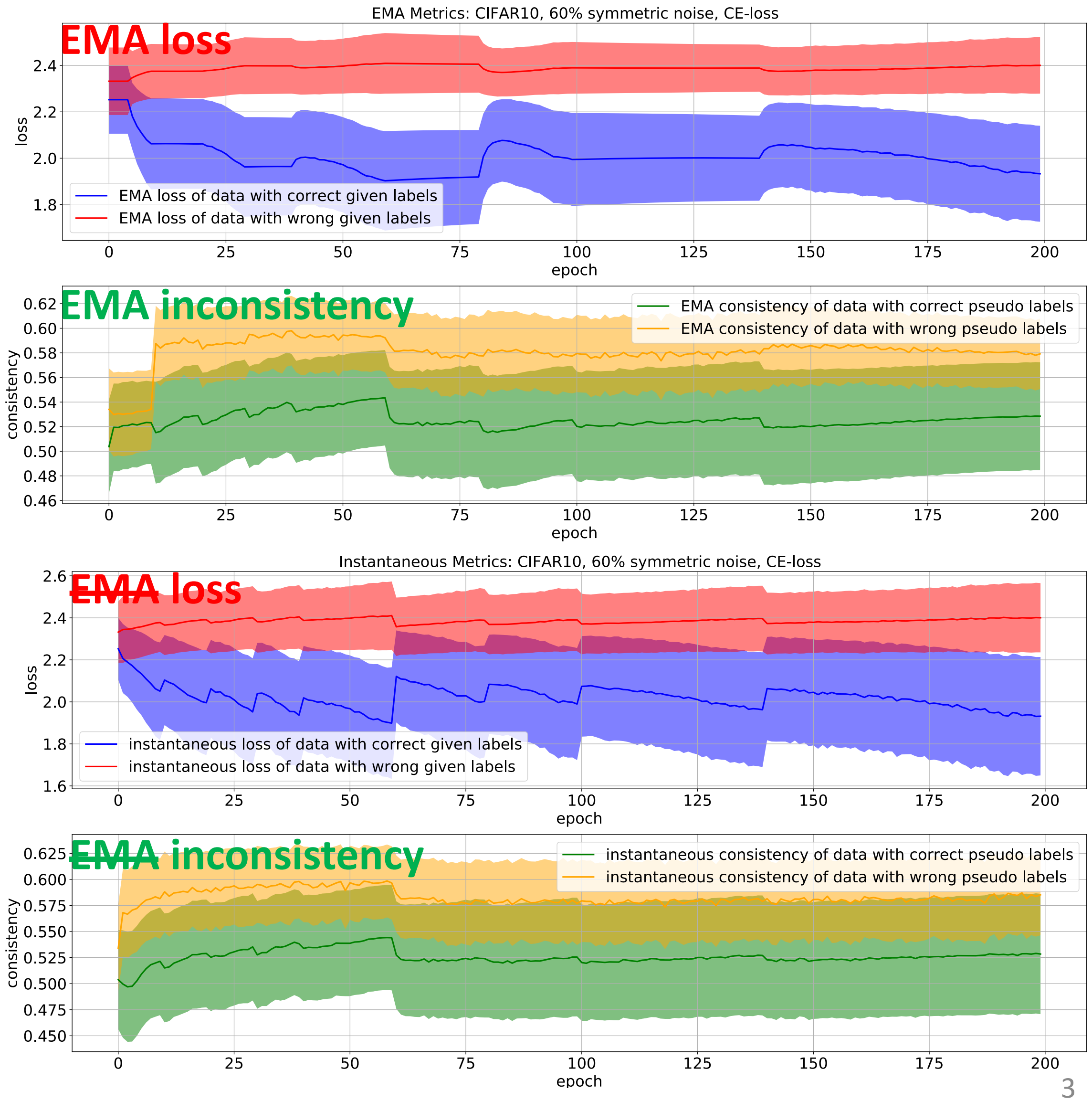


Training Dynamics identify correct given/pseudo-labels

- Clean Data Detection
  - We use **EMA (exponential moving average over time) loss** to select data with correct given labels.
  - **Supervised learning** on them.

- Wrong-label Correction
  - We use **EMA inconsistency** of model output over time to select data with correct pseudo labels.
  - **Self-supervised learning** on them.

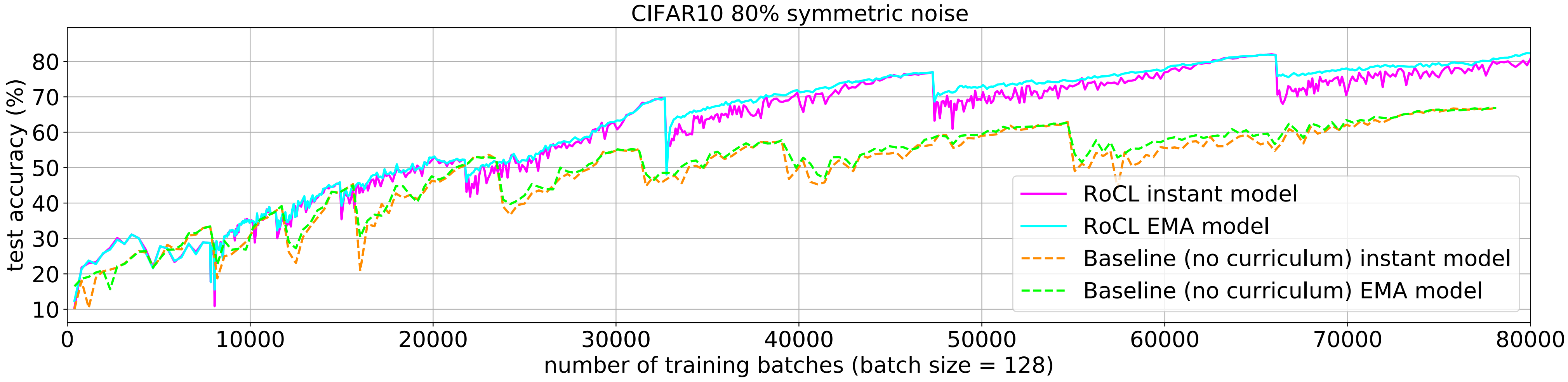
- **EMA loss** & **EMA inconsistency** work together to select:
  - **Clean data** with **wrong pseudo labels**
  - **Noisy data** with **correct pseudo labels**



Robust Curriculum Learning (RoCL) [Zhou et al., *ICLR* 2021]

- **Earlier: Supervised learning** on data with correct given label but wrong pseudo-label [small EMA loss & large EMA time inconsistency]

- **Later: Self-supervised learning** on data with wrong given label but correct pseudo-label [large EMA loss & small EMA time inconsistency]



Robust Curriculum Learning (RoCL) [Zhou et al., *ICLR* 2021]

- **Earlier: Supervised learning** on data with correct given label but wrong pseudo-label [small EMA loss & large EMA time inconsistency]
- **Later: Self-supervised learning** on data with wrong given label but correct pseudo-label [large EMA loss & small EMA time inconsistency]
- **Curriculum** τ_1: - → +, τ_2: + → -, λ: 1 → 0

$$\min_{\theta} F(\theta) \triangleq \boxed{\frac{\lambda}{\tau_1}\log\left(\frac{1}{n}\sum_{i=1}^{n}\exp[\tau_1\ell(f(x_i;\theta),y_i)]\right)} + \boxed{\frac{1-\lambda}{\tau_2}\log\left(\frac{1}{n}\sum_{i=1}^{n}\exp[\tau_2\zeta(i)]\right)}$$

Supervised loss:
LogSumExp loss with temperature τ_1 and weight λ

Self-supervised loss:
LogSumExp consistency loss with temp τ_2 and weight $1-\lambda$

RoCL achieves SoTA on Noisy-Label Benchmarks

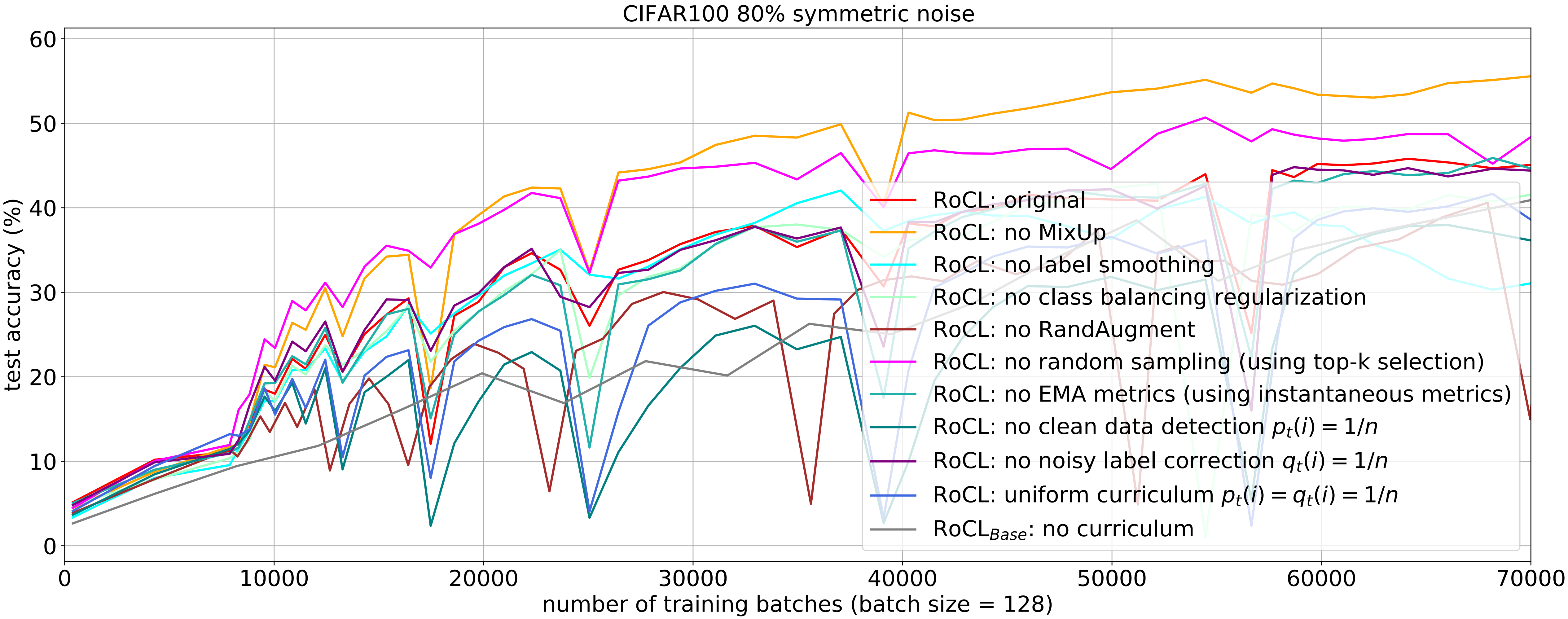
- **RoCL achieves state-of-the-art performance** on most benchmarks, including the ones with symmetric noises, asymmetric noises, and real-world web-label noises.
- RoCL significantly improves the **robustness to noise, test accuracy and efficiency**.

Table 1: Accuracy (%) evaluated on WebVision and ILSVRC2012 validation sets for DNNs trained by noisy-label learning methods on mini-WebVision training set (first 50 classes), which contains **real-world web-label noises**.

Val. set	WebVision		ILSVRC2012	
Accuracy	Top-1	Top-5	Top-1	Top-5
F-correct +⋆	61.12	82.68	57.36	82.36
Decoupling ⋆⋆	62.54	84.74	58.26	82.26
Co-teaching ⋆	63.58	85.20	61.48	84.70
MentorNet ⋆⋆	63.00	81.40	57.80	79.92
MentorMix ⋆‡⋆	76.00	90.20	72.90	91.10
D2L ⋆	62.68	84.00	57.80	81.36
INCV ⋆	65.24	85.34	61.60	84.98
RoCL (ours) ‡⋆†≀	**78.80**	**92.52**	**75.72**	**92.20**

Dataset	CIFAR10			CIFAR100		
Noise Rate	40%	60%	80%	40%	60%	80%
MD-DYR-SH	92.3	86.1	74.1	70.1	59.5	39.5
MentorNet	91.2	74.2	60.0	68.5	61.2	35.5
MentorMix	94.2	91.3	81.0	71.3	64.6	41.2
O2U-net	90.3	-	43.4	69.2	-	39.4
RoG+D2L	87.0	78.0	-	64.9	40.6	-
PENCIL	-	-	-	69.12 ± 0.62	57.79 ± 3.86	fail
GCE	87.62 ± 0.26	82.70 ± 0.23	67.92 ± 0.60	62.64 ± 0.33	54.04 ± 0.56	29.60 ± 0.51
SCE	85.34 ± 0.07	80.07 ± 0.02	53.81 ± 0.27	53.69 ± 0.07	41.47 ± 0.04	15.00 ± 0.04
NFL+MAE	83.81 ± 0.06	76.36 ± 0.31	45.23 ± 0.52	58.18 ± 0.08	46.10 ± 0.50	24.78 ± 0.82
NFL+RCE	86.05 ± 0.12	79.78 ± 0.13	55.06 ± 1.08	58.20 ± 0.31	46.30 ± 0.45	25.16 ± 0.55
NCE+MAE	84.19 ± 0.43	77.61 ± 0.05	49.62 ± 0.72	59.22 ± 0.36	48.06 ± 0.34	25.50 ± 0.76
NCE+RCE	86.02 ± 0.09	79.78 ± 0.50	52.71 ± 1.90	59.48 ± 0.56	47.12 ± 0.62	25.80 ± 1.12
RoCL (ours) ‡⋆†≀	$\mathbf{94.55 \pm 0.12}$	$\mathbf{92.98 \pm 0.23}$	$\mathbf{88.18 \pm 0.26}$	$\mathbf{74.64 \pm 0.43}$	$\mathbf{69.72 \pm 0.58}$	$\mathbf{58.72 \pm 0.62}$

Ablation Study and Hyperparameters of RoCL



Ablation Study and Hyperparameters of RoCL

- **The proposed curriculum** brings the most improvements.
- **Mix-Up** is less necessary since mixing wrong and correct labels rarely happens in our curriculum.
- **Data augmentation** is important for accurate identification of correct given/pseudo-labels by EMA metrics.
- **Class-balance regularization** is only important under very high noise rates.

Table 5: **Ablation study:** Test accuracy (%) of RoCL variants with one part removed/changed when applied to CIFAR10/100 corrupted by **symmetric(uniform)** label noise.

Dataset	CIFAR10		CIFAR100	
Noise Rate	60%	80%	60%	80%
RoCL: no MixUp	92.98	**88.18**	**69.72**	**58.72**
RoCL: no LabelSmooth	91.94	85.05	62.92	42.95
RoCL: no ClassBalance	**93.08**	74.91	62.66	43.94
RoCL: no RandAugment	86.59	72.35	64.84	44.06
RoCL: no RandSampling	92.31	85.99	64.09	57.00
RoCL: no EMA metrics	92.84	87.79	65.99	53.10
RoCL: $p_t(i) = 1/n$	92.42	86.05	62.69	44.35
RoCL: $q_t(i) = 1/n$	92.59	86.93	64.71	50.79
RoCL: $p_t(i) = q_t(i) = 1/n$	92.07	85.77	64.18	47.88
RoCL$_{Base}$: no curriculum	87.83	66.93	61.84	41.92
MentorMix: +RandAugment	85.45	20.68	52.70	8.02
MentorMix: +RandAugment-MixUp	84.31	38.21	58.31	8.18
MentorMix: original version	91.30	81.00	64.60	41.20
RoCL: original version	92.82	88.00	66.79	54.22

Thank you!

Poster Session 3:
May 3rd (Monday) at 17:00-19:00 PDT