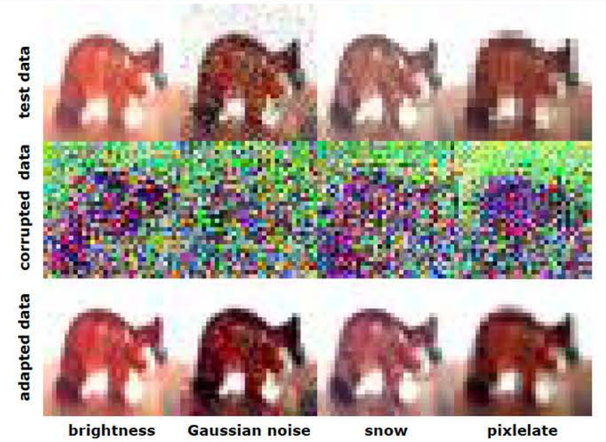
SODA: ROBUST TRAINING OF TEST-TIME DATA ADAPTORS

Zige Wang[1, 2], Yonggang Zhang[2], Zhen Fang[3], Long Lan[4],
Wenjing Yang[4,*], Bo Han[2]

[1] School of Computer Science, Peking University, [2] Hong Kong Baptist University
[3] University of Technology Sydney, [4] National University of Defense Technology

INTRODUCTION

- Motivation:
  - Deep neural networks suffer performance degradation due to distribution discrepancies between training and test data.
  - In practice, the parameters of deployed models may be unmodifiable and inaccessible in many applications due to intellectual property protection, misuse prevention, or privacy concerns in healthcare and finance.
  - Unreliable predicted labels will lead to unreliable gradient estimations in ZOO, which makes data features corrupted rather than adapted to deployed models.



Test-Time Adaptation (TTA)

Test-Time Data Adaptation + Zeroth-Order Optimization (ZOO)

Pseudo-Label-Robust Training Strategy

Pseudo-Label-Robust Data Adaptation (SODA)

METHOD

- Problem setting:
  - C-way image classification task with a distribution shift between the training and test data.
  - Given: Deployed model M with inaccessible parameters, data adaptor G, unlabeled test data $X = \{x_1, x_2, \dots, x_n\}$.
  - Restrictions: only the output probabilities are available from M.
- Goal: Adapt X to M without access to the parameters of M using G.

METHOD

- **ZOO in test-time data adaptation**:
  - Assume the true label of x_i is y_i, the directional derivative approximation of KL divergence loss is:

$$\widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_i = \frac{1}{\mu q} \sum_{j=1}^{q} \left[\left(\mathcal{L}(\mathbf{y}_i, \mathbf{M} \circ \mathbf{G}(\mathbf{x}_i; \boldsymbol{\theta} + \mu \mathbf{u}_j)) - \mathcal{L}(\mathbf{y}_i, \mathbf{M} \circ \mathbf{G}(\mathbf{x}_i; \boldsymbol{\theta})) \right) \mathbf{u}_j \right]$$

  - Let σ_i denote the distrubance of pseudo-label $\hat{y}_i$, i.e. $\hat{y}_i = y_i + \sigma_i$, and $\hat{p}_i^{\theta} = M \circ G(x_i; \theta)$, the KL divergence loss is:

$$\mathcal{L}_i = -H(\mathbf{y}_i + \boldsymbol{\sigma}_i) + \mathcal{L}_{\text{ce}}(\mathbf{y}_i, \hat{\mathbf{p}}_i^{\boldsymbol{\theta}}) - \boldsymbol{\sigma}_i \log \hat{\mathbf{p}}_i^{\boldsymbol{\theta}}$$

  - Then, replacing y_i with $\hat{y}_i$, the directional derivative approximation becomes:

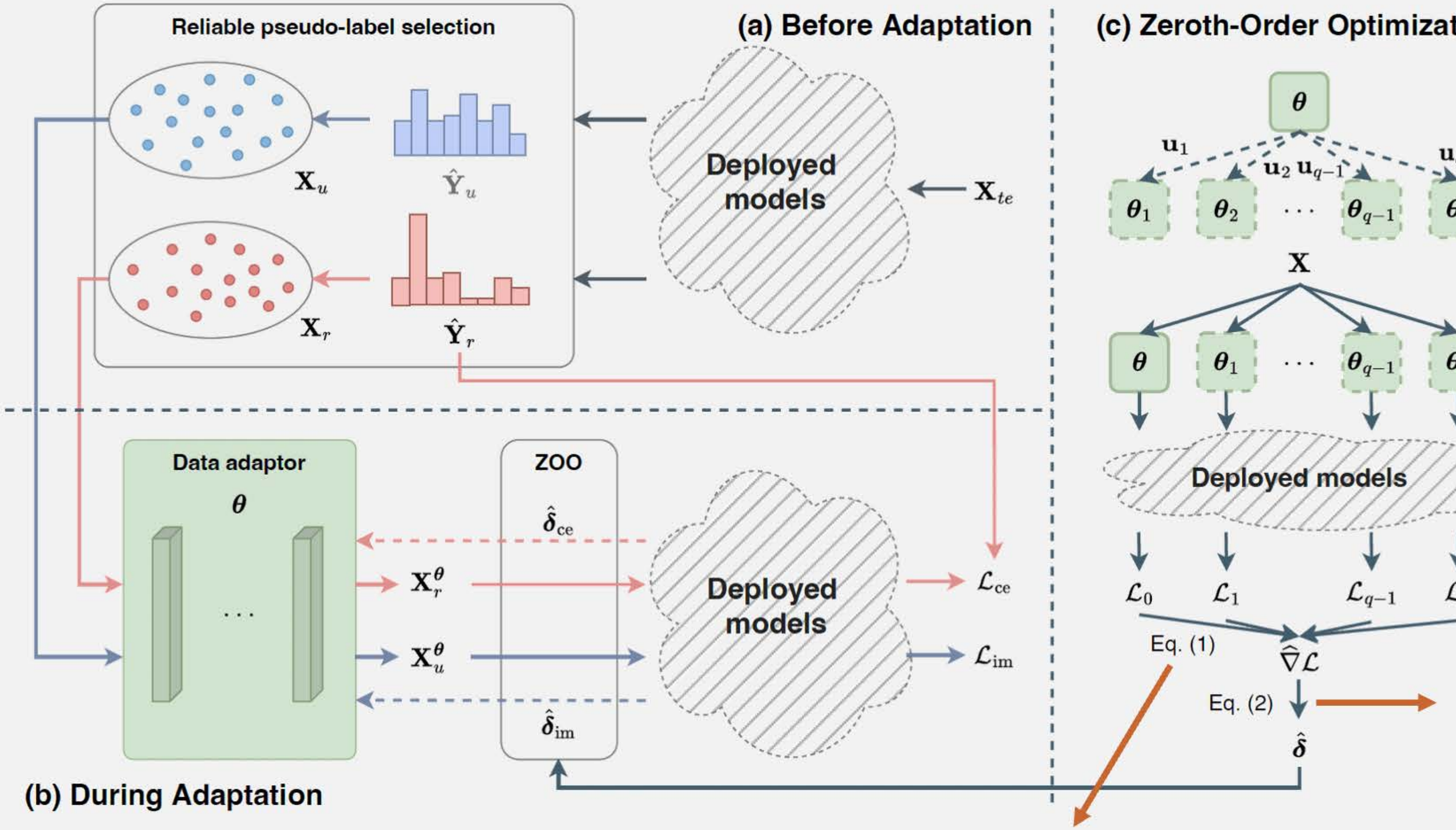
$$\widehat{\nabla}_{\boldsymbol{\theta}} \check{\mathcal{L}}_i = \widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\text{ce}} + \boxed{\frac{\boldsymbol{\sigma}_i}{\mu q} \sum_{j=1}^{q} \log \frac{\hat{\mathbf{p}}_i^{\boldsymbol{\theta}}}{\hat{\mathbf{p}}_i^{\boldsymbol{\theta} + \mu \mathbf{u}_j}} \mathbf{u}_j}$$

  - Where $\widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\text{ce}}$ is the ideal directional derivative approximation.
- **Pseudo-label-robust training**:
  - Select reliable pseudo-labels with small σ_i: pseudo-labels with confidence higher than τ; the number of selected pseudo-labels for each class less than $(1-\rho)n/C$.
  - Data with unreliable pseudo-labels: mutual information maximization

$$\mathcal{L}_{\text{im}}(\mathbf{X}_u^{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x}_i^{\boldsymbol{\theta}} \in \mathbf{X}_u^{\boldsymbol{\theta}}} [\sum_{k=1}^{C} \hat{\mathbf{p}}_{ik} \log \hat{\mathbf{p}}_{ik}] - \sum_{k=1}^{C} \mathbb{E}_{\mathbf{x}_i^{\boldsymbol{\theta}} \in \mathbf{X}_u^{\boldsymbol{\theta}}} \hat{\mathbf{p}}_{ik} \log \mathbb{E}_{\mathbf{x}_i^{\boldsymbol{\theta}} \in \mathbf{X}_u^{\boldsymbol{\theta}}} \hat{\mathbf{p}}_{ik}$$

METHOD

- Framework overview:



$$\hat{\boldsymbol{\delta}} = \frac{1}{l}\sum_{i=1}^{l} \widehat{\nabla}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}; \mathbf{x}_i)$$

$$\mathcal{L}_{\text{all}}(\mathbf{X}, \hat{\mathbf{Y}}_r) = -\mathcal{L}_{\text{im}}(\mathbf{X}_u) + \alpha \mathcal{L}_{\text{ce}}(\mathbf{X}_r, \hat{\mathbf{Y}}_r)$$

$$\widehat{\nabla}_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) := \frac{1}{\mu q}\sum_{i=1}^{q} \left[(f(\boldsymbol{\theta} + \mu \mathbf{u}_i) - f(\boldsymbol{\theta}))\mathbf{u}_i \right]$$

THEORETICAL ANALYSIS

- For simplicity, we consider the special case where directional derivative approximation equals to gradient estimation with the mini-batch size = 1.
- The **expected estimation error** between the true gradient and the estimated gradient w.r.t. to the whole test dataset is:

$$\mathcal{R}_{\mathbf{X}} = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}[\| \hat{\nabla}_{\boldsymbol{\theta}} \check{\mathcal{L}}_i - \nabla_{\boldsymbol{\theta}} \mathcal{L}_i \|_2]\right]$$

- **Before applying pseudo-label-robust training**: denote $h(x_i) = -\sigma_i \log \hat{p}_i^{\theta}$,

$$\mathcal{R}_{\mathbf{X}} \leq \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}[\| \widehat{\nabla}_{\boldsymbol{\theta}} \check{\mathcal{L}}_{\mathrm{ce}} - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{ce}} \|_2] + \mathbb{E}[\| \widehat{\nabla}_{\boldsymbol{\theta}} h - \nabla_{\boldsymbol{\theta}} h \|_2]\right].$$

- **After applying pseudo-label-robust training** : according to previous study[1], minimizing cross-entropy loss is equivalent to maximizing mutual information, then:

$$\widetilde{\mathcal{R}}_{\mathbf{X}} \leq \mathbb{E}_{\mathbf{X}_r}\left[\mathbb{E}[\| \hat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{ce}} - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{ce}} \|_2] + \mathbb{E}[\| \hat{\nabla}_{\boldsymbol{\theta}} h - \nabla_{\boldsymbol{\theta}} h \|_2]\right] + \mathbb{E}_{\mathbf{X}_u}\left[\mathbb{E}[\| \hat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{ce}} - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{ce}} \|_2]\right]$$

- **The upper bound of expected estimation error is tightened after applying our pseudo-label-robust training strategy.**

[1] Boudiaf, Malik, et al. "A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses." ECCV, 2020.

EXPERIMENTS

- Experiments on common OOD benchmarks, CIFAR-10-C, CIFAR-100-C and ImageNet-C, the reported accuracies (%) are averaged over 19 corruptions:

Categories	Methods	FO Grad.	Model Mod.	C10-C	C100-C	IN-C
-	Deployed	-	-	72.39	41.41	31.36
Distill.	DINE	✓	✗	73.86	40.52	-
	BETA	✓	✗	75.71	39.62	-
DA	DA-PGD	✗	✗	24.63	4.15	14.39
	DA-ZOO-Input	✗	✗	68.70	31.53	17.57
	DA-Direct	✗	✗	70.48	37.67	29.37
	DA-PL	✗	✗	72.93	41.44	31.91
	SODA (Ours)	✗	✗	**82.55**	**52.41**	**42.14**
	SODA-R (Ours)	✓	✗	**88.39**	60.31	48.70
MA	MA-SO	✓	✓	86.54	**62.02**	**56.90**

- More extensive experiments and discussions can be found in paper.

EXPERIMENTS

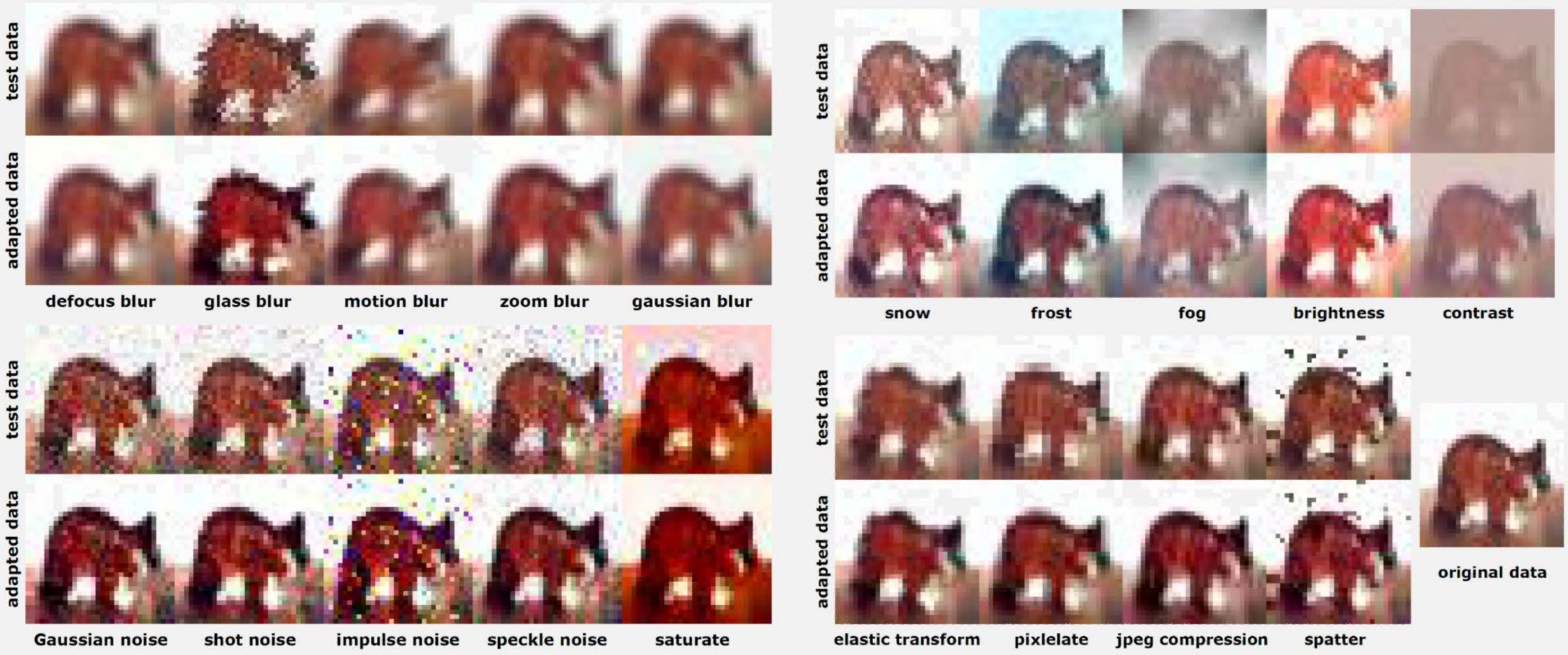
- Experiments in online setting where test data points arrive sequentially:
  - An ordered queue with queue size S is maintained during adaptation to store the selected reliable pseudo-labels and their corresponding data points.
  - The optimization in SODA-O is not repeated after reaching the entire test dataset but only repeats for the current test data batch and the cached queue
- The results on CIFAR-10-C and CIFAR-100-C:

Methods	Deployed	SODA-O						SODA
Epochs/Batch	-	5	10	30	50	100	150	150*
CIFAR-10-C	72.39	75.22	77.03	79.63	80.38	81.33	81.71	82.55
CIFAR-100-C	41.41	43.59	45.81	48.56	49.26	50.04	50.12	52.41

*SODA is trained over the entire test dataset for 150 epochs

EXPERIMENTS

- Visualization:



CONCLUSIONS

- Three challenges:
  - Unmodifiable model parameters: test-time data adaptation.
  - Infeasible gradients: zeroth-order optimization.
  - Unreliable pseudo-labels: pseudo-label-robust training.
- Revisiting ZOO in test-time data adaptation and pointing out that the unreliable pseudo-labels can cause biased gradient estimation in ZOO.
- Both experimental and theoretical analyses demonstrate the effectiveness of SODA.

THANKS FOR LISTENING!