CVPR
JUNE 17-21, 2024
SEATTLE, WA

TEA: Test-time Energy Adaptation

Yige Yuan[1,2], Bingbing Xu[1], Liang Hou[3], Fei Sun[1], Huawei Shen[1,2], Xueqi Cheng[1,2]

[1]CAS Key Laboratory of AI Security, Institute of Computing Technology, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences

[3]Kuaishou Technology

Paper　　Code

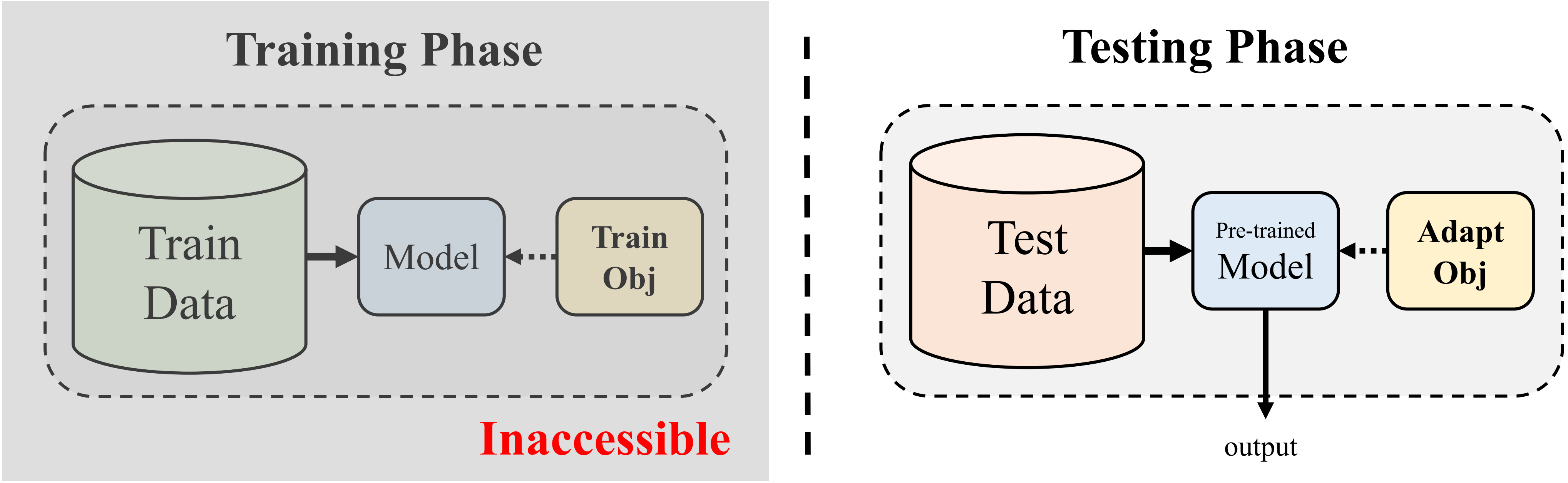
Presenter: Yige Yuan （袁一歌）
Email: yuanyige20z@ict.ac.cn

- **Motivation**
- Method
- Experiments

MOTIVATION

Test-time Adaptation

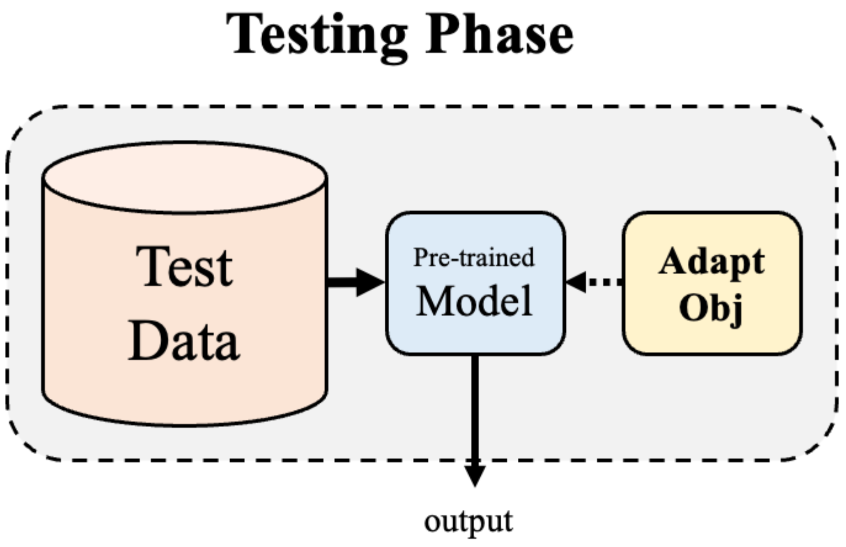
- Inaccessible training data and training procedure.
- Adapting a pre-trained model with test data.



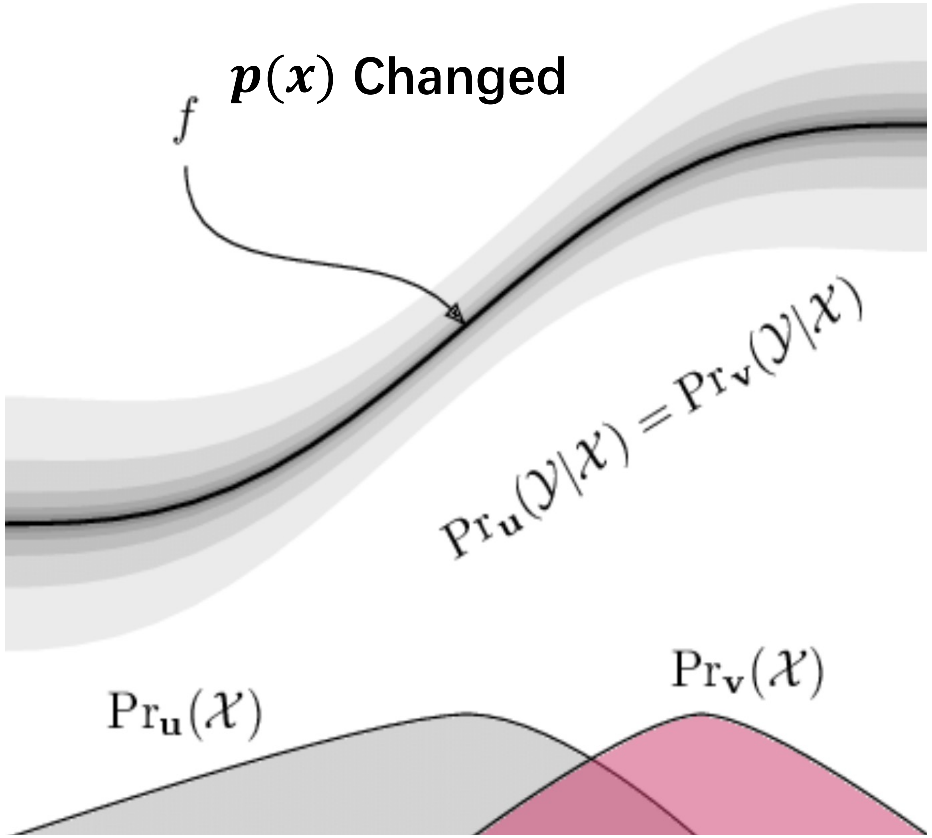
MOTIVATION

Related Work

- Entropy-based TTA
  - Minimize the prediction entropy.
  - TENT, ETA, EATA, SAR
- Pseudo-labeling-based TTA
  - Utilize test-time generated labels for updates.
  - PL, SHOT
- Consistency-based TTA
  - Constraint consistency across augmented samples.
  - MEMO, AdaContrast
- ….



MOTIVATION



Covariate Shift

- the decrease in generalization ability on test data with distribution shift can be attributed to the model's reliance on the marginal distribution of the training data
- They do not address the marginal distribution shift $p(x)$, imparing model calibration and introducing confirmation bias.

How to perceive marginal distribution $p(x)$?

MOTIVATION

Energy Based Model

- **What is? A Non-normalized Probabilistic Model**

the energy function maps each sample into an energy that can be considered as an unnormalized probability, with lower scores indicating higher likelihoods

$$\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{x}) : R^D \rightarrow R$$

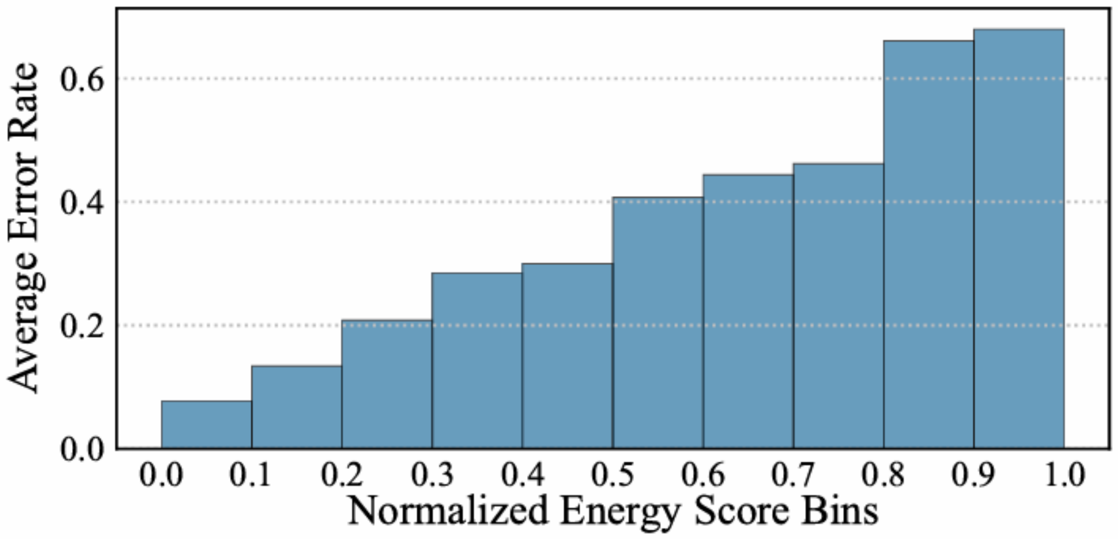
$$p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z_\theta}$$

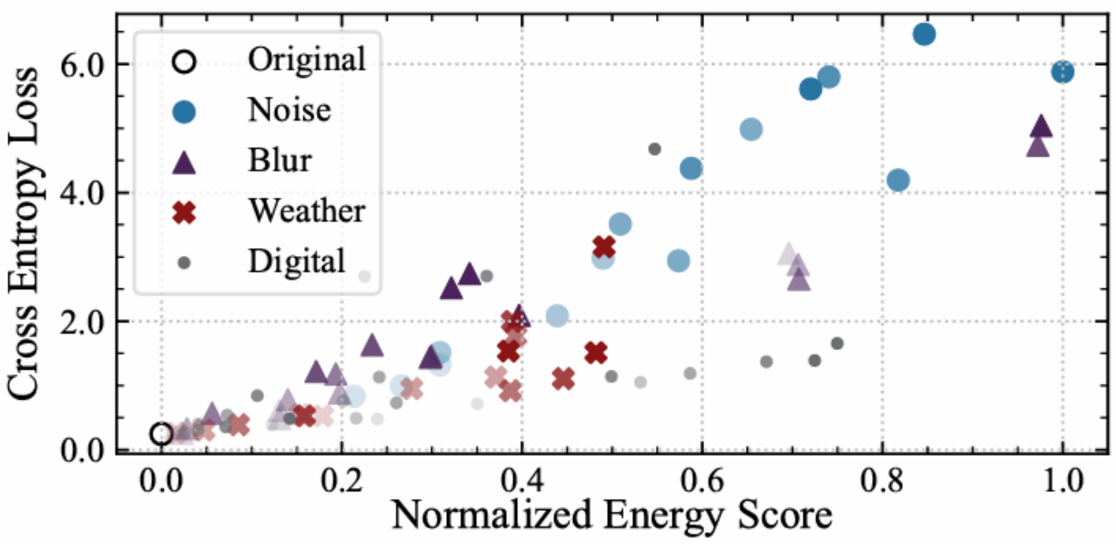
$$Z_\theta = \int \sum_y \exp\left(f_\theta(\mathbf{x})[y]\right) \mathrm{d}\mathbf{x}$$

MOTIVATION

How Energy related to Generalization

- Our Observation
  - Low energy = high probability, high performance
  - High enery = low probability, low performance





- Motivation
- **Method**
- Experiments

METHOD

Overall idea

Enhancing the model's perception of test distribution from an energy-based perspective, involving two key steps:

- treating the Pretrained Classifier as the energy-based model
- optimizing it to preceive test data by decreasing the energy

Notation

Labeled training data $(x, y) \sim \mathrm{P}_{\mathrm{train}}(x, y)$	Inaccessible
Unlabeled test data $x \sim \mathrm{P}_{\mathrm{test}}(x)$	Accessible
Pre-trained Classifier $f_\theta : X \rightarrow Y$	Accessible

METHOD

Treating Classifier as EBM

Pretrained Classifier

$$p_\theta(y \mid \mathbf{x}) = \frac{\exp(f_\theta(\mathbf{x})[y])}{\sum_{y'} \exp(f_\theta(\mathbf{x})[y'])}$$

introduce unknown normalizing constant Z_θ

$$p_\theta(\mathbf{x}, y) = \frac{\exp(f_\theta(\mathbf{x})[y])}{Z_\theta}$$

marginalize out y

$$p_\theta(\mathbf{x}) = \sum_y p_\theta(\mathbf{x}, y) = \frac{\sum_y \exp\left(f_\theta(\mathbf{x})[y]\right)}{Z_\theta}$$

substitute E_θ

$$p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z_\theta}$$

Energy Based Model

$$E_\theta(\mathbf{x}) = -\log \sum_y \exp\left(f_\theta(\mathbf{x})[y]\right)$$

A Pre-Trained Classifier can be reinterpret as an Energy Based Model

METHOD

Inject Test Distribution into Classifier

Our Objective:

Test Data

$$\max \boxed{p_\theta}(\boxed{\mathbf{x}_{\text{test}}})$$

Classifier's Inherent Distribution

Modeling Test Distribution under Pre-trained Classifier

$$p_\theta(\mathbf{x}_{\text{test}}) = \frac{\exp(-E_\theta(\mathbf{x}_{\text{test}}))}{Z_\theta}$$

$$Z_\theta = \int \sum_y \exp\left(f_\theta(\mathbf{x})[y]\right) \mathrm{d}\mathbf{x}$$

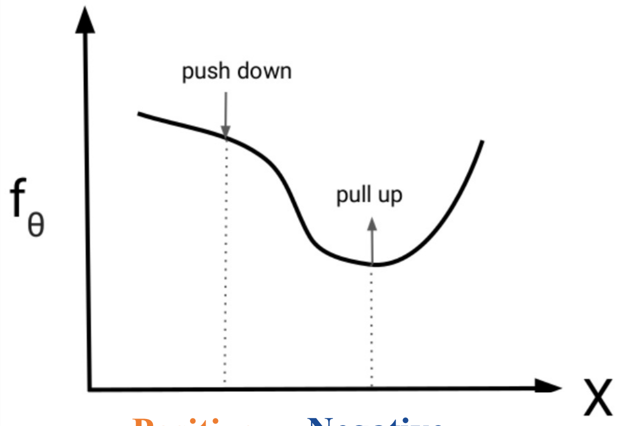
Computing Z requires integrating all x, How to Optimize?

METHOD

Inject Test Distribution into Classifier

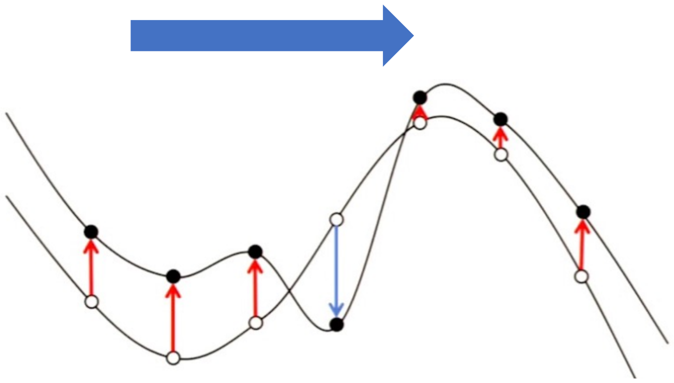
Contrastive Divergence

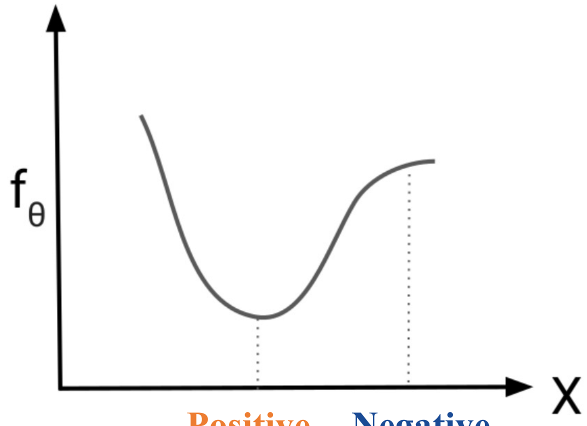
$$\frac{\partial \log p_\theta(\mathbf{x}_{\text{test}})}{\partial \theta} = \mathbb{E}_{\tilde{\mathbf{x}} \sim p_\theta}\left[\frac{\partial E_\theta(\tilde{\mathbf{x}})}{\partial \theta}\right] - \frac{\partial E_\theta(\mathbf{x}_{\text{test}})}{\partial \theta}$$



Positive Sample (Test)
Negative Sample (Train, Noise)

Energy-Based Training





Positive Sample
Negative Sample

METHOD

Inject Test Distribution into Classifier

Test Distribution Injection: $$\max \underbrace{\boxed{p_\theta}}_{\text{Classifier's Inherent Distribution}}(\overbrace{\boxed{\mathbf{x}_{\text{test}}}}^{\text{Test Data}})$$

Contrastive Divergence

$$\frac{\partial \log p_\theta(\mathbf{x}_{\text{test}})}{\partial \theta} = \mathbb{E}_{\boxed{\tilde{\mathbf{x}} \sim p_\theta}}\left[\frac{\partial E_\theta(\tilde{\mathbf{x}})}{\partial \theta}\right] - \frac{\partial E_\theta(\mathbf{x}_{\text{test}})}{\partial \theta}$$

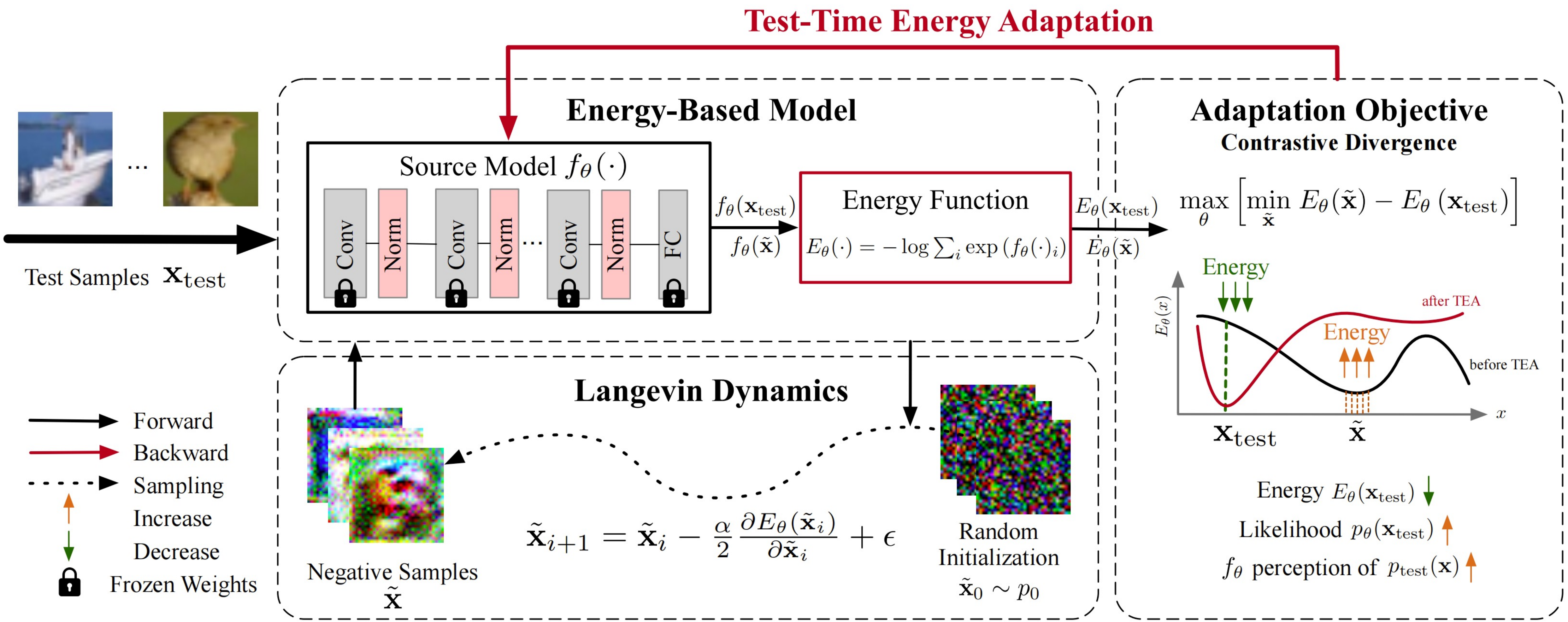
Stochastic Gradient Langevin Dynamics

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{x}_t - \frac{\alpha}{2}\frac{\partial E_\theta(\tilde{\mathbf{x}}_t)}{\partial \tilde{\mathbf{x}}_t} + \sqrt{\alpha}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

Overall Objective

$$\max p_\theta(\mathbf{x}_{\text{test}}) = \max_\theta \left[\min_{\tilde{\mathbf{x}}} E_\theta(\tilde{\mathbf{x}}) - E_\theta(\mathbf{x}_{\text{test}})\right]$$

METHOD



Test-Time Energy Adaptation
Energy-Based Model
Source Model $f_\theta(\cdot)$
Conv
Norm
Conv
Norm
Conv
Norm
FC
Test Samples $\mathbf{x}_{\text{test}}$
$f_\theta(\mathbf{x}_{\text{test}})$
$f_\theta(\tilde{\mathbf{x}})$
Energy Function
$E_\theta(\cdot) = -\log \sum_i \exp(f_\theta(\cdot)_i)$
$E_\theta(\mathbf{x}_{\text{test}})$
$E_\theta(\tilde{\mathbf{x}})$
Adaptation Objective
Contrastive Divergence
$\max_\theta \left[\min_{\tilde{\mathbf{x}}} E_\theta(\tilde{\mathbf{x}}) - E_\theta(\mathbf{x}_{\text{test}}) \right]$
Energy
after TEA
Energy
before TEA
$E_\theta(x)$
x
$\mathbf{x}_{\text{test}}$
$\tilde{\mathbf{x}}$
Energy $E_\theta(\mathbf{x}_{\text{test}})$
Likelihood $p_\theta(\mathbf{x}_{\text{test}})$
f_θ perception of $p_{\text{test}}(\mathbf{x})$
Langevin Dynamics
Negative Samples
$\tilde{\mathbf{x}}$
$\tilde{\mathbf{x}}_{i+1} = \tilde{\mathbf{x}}_i - \frac{\alpha}{2} \frac{\partial E_\theta(\tilde{\mathbf{x}}_i)}{\partial \tilde{\mathbf{x}}_i} + \epsilon$
Random
Initialization
$\tilde{\mathbf{x}}_0 \sim p_0$
Forward
Backward
Sampling
Increase
Decrease
Frozen Weights

- Motivation
- Method
- **Experiments**

EXPERIMENTS

Experimental Settings

- **Datasets**
  - 15 shift corruption distributions by **CIFAR-10(C), CIFAR-100(C), and Tiny-ImageNet(C)**.
  - 4 domains in the PCAS dataset: **Photo, Art, Cartoon, and Sketch**.
- **Baselines**
  - Without TTA: **Source**
  - Norm Based TTA: **BN DUA**
  - Pseudo Label Based TTA: **PL, SHOT**
  - Entropy Based TTA: **TENT, ETA, EATA, SAR**
- **Metrics**
  - **Accuracy** on clean data.
  - **Average Accuracy** and **mCE** across all severity levels and at the severest level.

EXPERIMENTS

Image Corruption Scenario Performance

Table 1. Comparisons of TEA and baselines for image corruption on CIFAR-10(C), CIFAR-100(C), and Tiny-ImageNet(C) using WRN-28-10 with BatchNorm. Accuracy and mCE are evaluated at the most severe level and across all levels with asterisk (*) indicating the results are taken from the original paper [56]. The best adaptation results are highlighted in **boldface**.

WRN-28-10 BatchNorm		CIFAR-10(C)					CIFAR-100(C)					Tiny-ImageNet(C)				
		Clean	Corr Severity 5		Corr Severity 1-5		Clean	Corr Severity 5		Corr Severity 1-5		Clean	Corr Severity 5		Corr Severity 1-5	
		Acc (↑)	Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)	Acc (↑)	Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)	Acc (↑)	Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)
Source		94.77	56.47	100.00	73.45	100.00	81.79	35.39	100.00	52.12	100.00	63.19	21.21	100.00	34.13	100.00
Norm	BN [52]	93.97	79.56	52.65	85.63	60.00	80.83	60.06	63.54	68.11	69.42	45.04	27.74	93.42	34.27	100.96
	DUA* [41]	-	80.10	50.78	-	-	-	-	-	-	-	-	-	-	-	-
Pseudo	PL [34]	93.75	51.42	106.98	72.62	99.37	80.52	53.40	72.12	64.53	75.29	47.84	28.26	91.22	39.83	91.67
	SHOT [36]	93.25	74.77	63.19	82.35	72.61	80.52	56.53	68.01	66.00	73.28	47.95	29.14	90.16	**40.01**	**91.41**
Entropy	TENT [60]	93.66	81.41	48.13	86.75	56.17	80.14	63.09	59.42	69.47	67.80	39.54	26.31	95.52	32.03	104.49
	ETA [45]	93.96	79.58	52.64	85.63	59.99	80.65	59.82	64.52	67.17	72.40	43.20	27.28	94.12	33.46	102.25
	EATA [45]	93.96	79.59	52.62	85.64	59.98	80.68	60.24	63.75	67.48	71.66	43.42	27.28	94.09	33.47	102.24
	SAR [46]	93.97	79.77	51.94	85.83	58.97	80.84	62.95	59.37	70.01	65.99	41.58	28.21	92.82	34.60	100.47
Energy	TEA	**94.09**	**83.34**	**43.69**	**87.88**	**52.00**	**80.88**	**65.10**	**56.18**	**71.22**	**63.74**	**51.65**	**31.67**	**87.99**	39.96	92.12

Figure: Adaptation performance under image corruption scenarios using BatchNorm

EXPERIMENTS

Image Corruption Scenario Performance

Table 2. Comparisons for image corruption on CIFAR-10(C), CIFAR-100(C), and Tiny-ImageNet(C) using ResNet-50 with GroupNorm across all severity levels. Best results are in **boldface**.

ResNet50 GroupNorm		CIFAR-10(C)		CIFAR-100(C)		Tiny-ImageNet(C)	
		Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)	Acc (↑)	mCE (↓)
Source		78.71	100.00	54.98	100.00	26.64	100.00
Pseudo	PL	79.43	94.76	56.68	96.02	26.60	99.92
	SHOT	81.98	86.65	58.31	93.45	29.11	96.73
Entropy	TENT	77.29	102.88	56.34	96.88	26.65	99.94
	ETA	78.68	100.09	56.72	96.37	29.25	96.42
	EATA	78.70	100.02	56.76	96.28	29.25	96.42
	SAR	78.78	99.65	55.28	99.33	27.05	99.41
Energy	TEA	**83.05**	**79.09**	**59.67**	**89.32**	**30.41**	**94.81**

Figure: Adaptation performance under image corruption scenarios using GroupNorm

EXPERIMENTS

Domain Generalization Scenario Performance

Table 3. Single source domain generalization comparisons on PACS datasets using ResNet-18 with BatchNorm in terms of Accuracy. The best adaptation results are highlighted in **boldface**.

Source Domain	Method	Target Domain: Photo	Art	Cartoon	Sketch	Avg
Photo	Source	-	26.76	22.40	16.62	21.93
	BN	-	26.66	27.94	15.96	23.52
	TENT	-	26.95	29.86	17.54	24.78
	EATA	-	26.66	28.11	15.98	23.59
	SAR	-	26.71	28.41	15.98	23.70
	SHOT	-	26.61	29.86	20.92	25.80
	TEA	-	**28.81**	**33.62**	**20.49**	**27.64**
Art	Source	49.04	-	36.43	24.48	36.65
	BN	46.65	-	28.28	22.73	32.55
	TENT	50.78	-	30.12	24.61	35.17
	EATA	46.83	-	29.31	23.42	33.19
	SAR	47.90	-	33.02	26.27	35.73
	SHOT	50.24	-	34.30	**29.37**	37.97
	TEA	**56.29**	-	**38.57**	28.71	**41.19**

Source Domain	Method	Target Domain: Photo	Art	Cartoon	Sketch	Avg
Cartoon	Source	42.69	29.79	-	29.47	33.98
	BN	28.68	25.15	-	20.87	24.90
	TENT	30.96	23.34	-	22.65	25.65
	EATA	28.80	25.10	-	**25.04**	26.31
	SAR	29.70	25.78	-	21.51	25.66
	SHOT	**37.72**	22.66	-	23.14	27.84
	TEA	36.05	**31.44**	-	22.88	**30.12**
Sketch	Source	19.94	18.70	32.21	-	23.62
	BN	13.47	17.14	29.86	-	20.16
	TENT	13.53	17.38	29.52	-	20.14
	EATA	13.17	17.33	30.08	-	20.19
	SAR	13.29	18.80	29.95	-	20.68
	SHOT	**19.76**	18.75	30.46	-	22.99
	TEA	19.64	**21.24**	**33.19**	-	**24.69**

Figure: Adaptation performance under domain generalization scenarios

EXPERIMENTS

Energy Reduction & Generalizability Enhancement

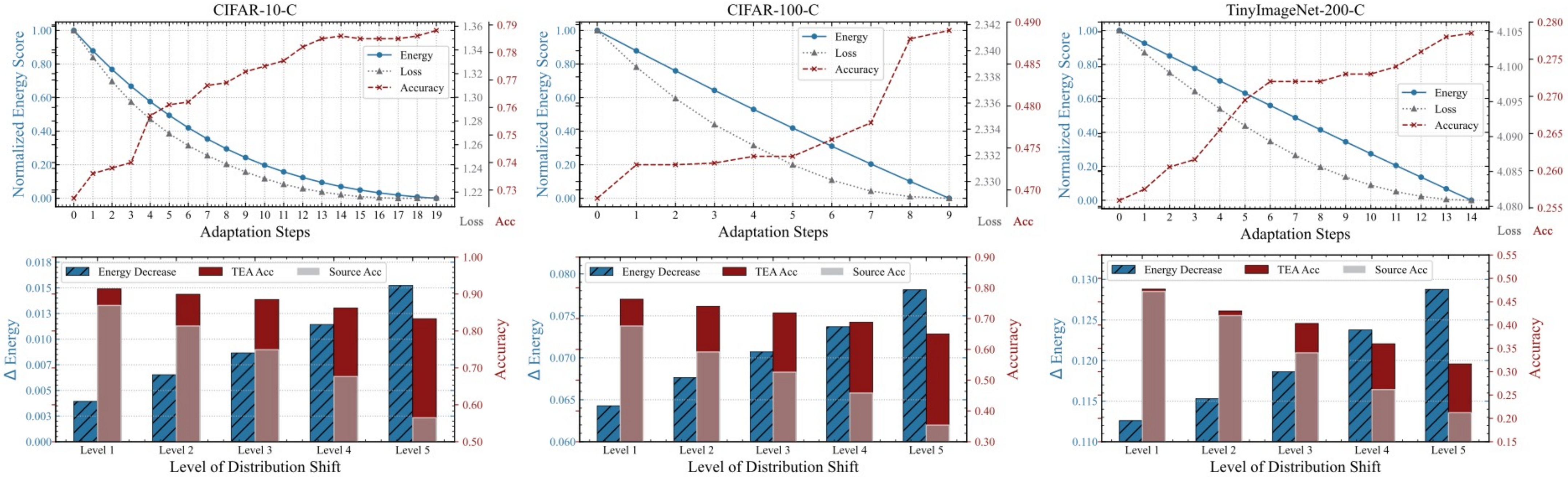


Figure 3. This illustration captures the energy reduction and generalizability enhancement achieved by TEA across CIFAR-10-C, CIFAR-100-C, and TinyImageNet-200-C, displayed from left to right. The **upper** set of graphs trace the evolution of energy score, corresponding loss and accuracy in response to incrementally increasing TEA adaptation steps. The **lower** set uncovers the extent of energy reduction and the consequent performance improvement before and after executing TEA adaptation, under different levels of distribution shift.

Figure: Relation between energy reduction and generalizability enhancement

EXPERIMENTS

Distribution Perception and Generation

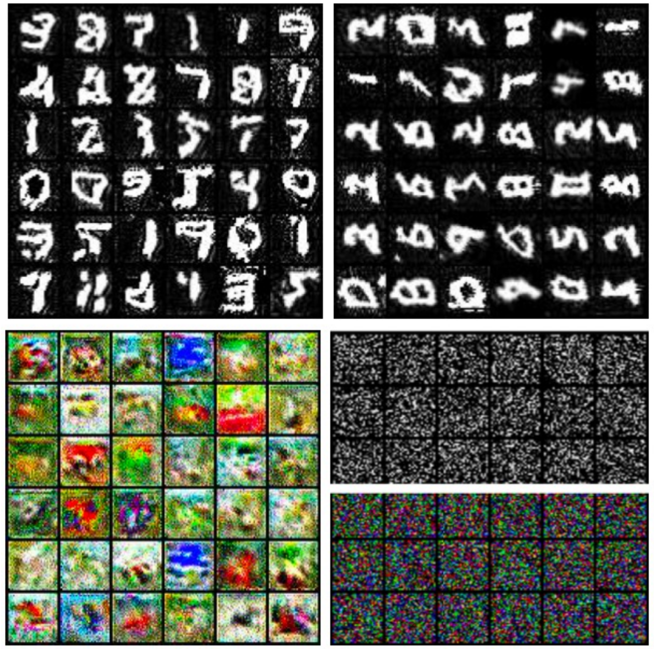


Figure 4. Test distribution perception visualization for identical training and testing distributions on MNIST and CIFAR-10.

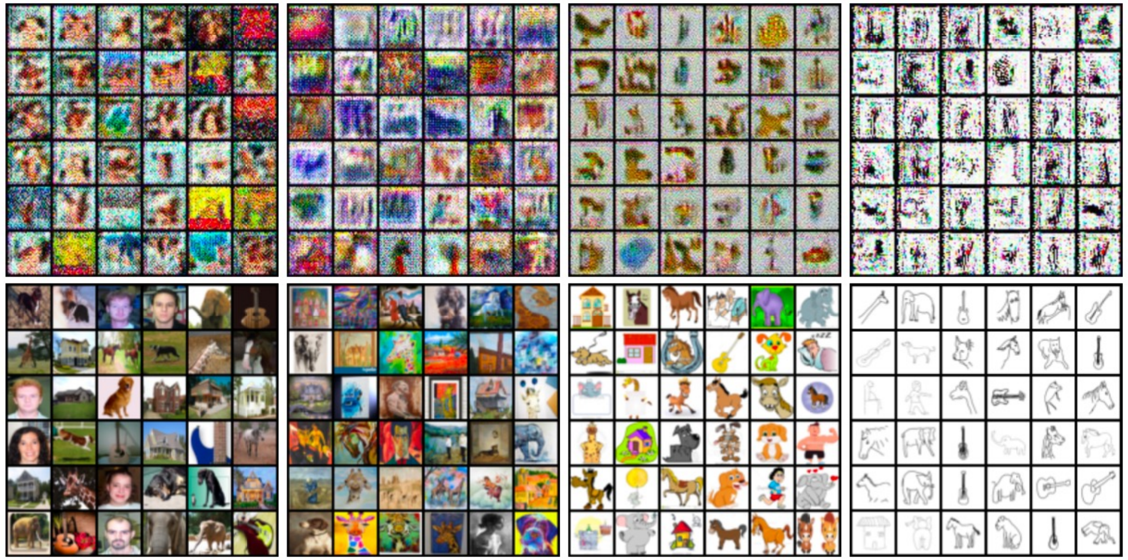


Figure 5. Test distribution perception visualization (**upper**) and real samples (**lower**) on shifted distribution: A model trained on PACS-A dataset then individually tested with TEA adaptation across PACS-P, PACS-A, PACS-C, PACS-S datasets.

Figure: TEA's Distribution Perception and Generation

EXPERIMENTS

Distribution Perception and Generation

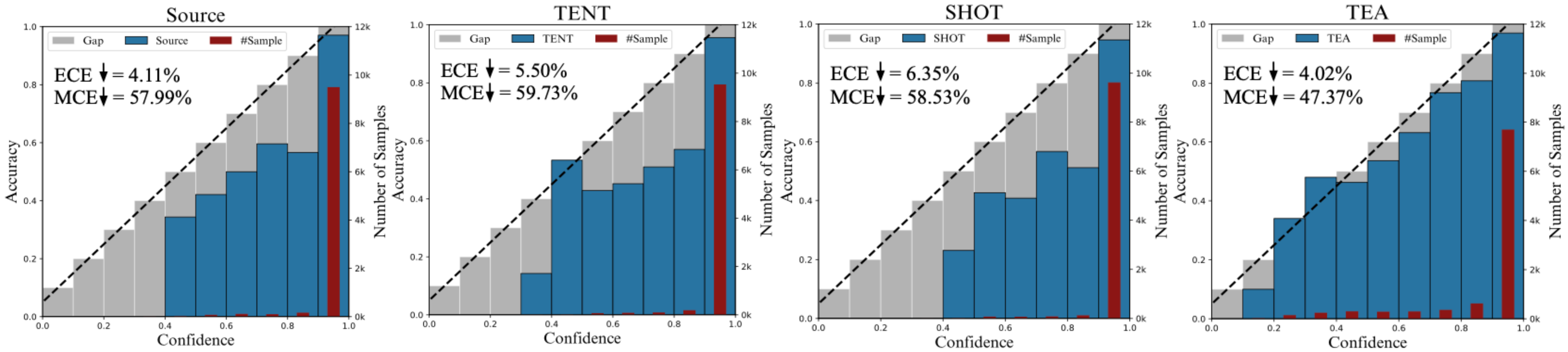


Figure 6. Calibration comparison between TEA and baselines on CIFAR-10 dataset. In an ideal scenario for optimal calibration, blue bars should align with the diagonal line, and a smaller grey gap area is preferred. Quantitative measures are provided via ECE and MCE metrics, where lower values indicate better calibration.

Figure: TEA’s Improvements in Confidence Calibration

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Thanks for all the authors of this paper:

Yige Yuan
- Generalization
- Trustworthy AI

Bingbing Xu
- Graph Neural Networks
- Network Embedding

Liang Hou
- Generative Adversarial Nets
- Generative Models

Fei Sun
- Recommender Systems
- Natural Language Processing

Huawei Shen
- Network Data Mining
- Social Network Analysis
- Graph Neural Networks

Xueqi Cheng
- Network Data Science
- Social Computing
- Information Retrieval

Thank you for your attention!

Paper Code Homepage

Email: yuanyige20z@ict.ac.cn