A Probabilistic Framework for Lifelong Test-Time Adaptation

Dhanajit Brahma and Piyush Rai

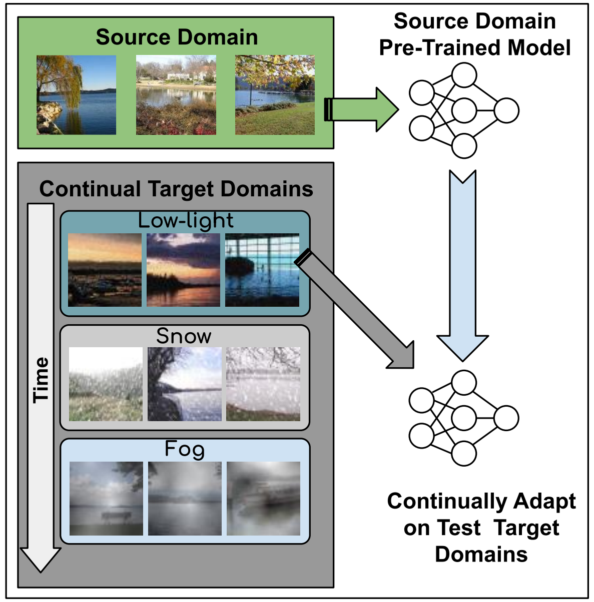
Indian Institute of Technology Kanpur, India

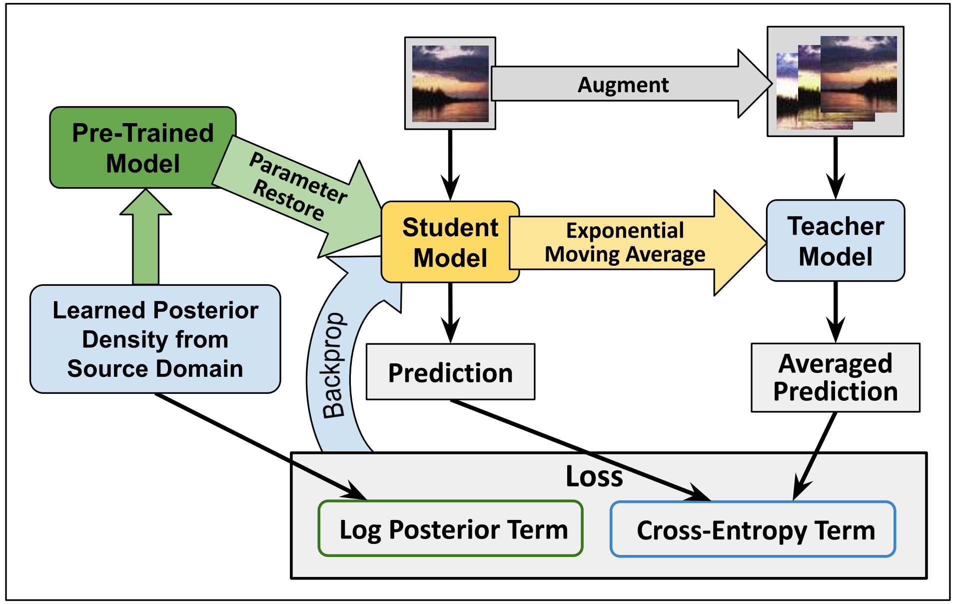
Poster: TUE-AM-341

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Overview

- We propose PETAL (Probabilistic lifElong Test-time Adaptation with seLf-training prior)
- PETAL is a probabilistic framework for lifelong TTA using a partly data-driven prior
- Probabilistic formulation → Student-teacher cross-entropy loss with a regularizer term corresponding to posterior of source domain data
- Further, we propose data-driven parameter restoration
- PETAL achieves SoTA on various lifelong TTA benchmarks





Introduction

- Domain shift between source training data and target test data
- Source data not available during inference: privacy concerns or legal constraints
- Deep neural networks make inaccurate predictions, unreliable uncertainty estimates
- One way to robustify DNNs: Test-time Adaptation
- **Test-time adaptation (TTA)**: Adapt source pre-trained model by learning from unlabeled test data
- Real-world machine systems work in non-stationary and continually changing environment
- **Lifelong/Continual TTA**: Target test domain distribution can change over time

Problem Set-Up

$X = \{x_n, y_n\}_{n=1}^{N}$: source training data

θ_0: pre-trained model trained on X

$U_d = \{x_m\}_{m=1}^{M_d}$: Unlabeled target (test) domain data

Test-Time Adaptation

Aim: Adapt θ_0 for each target domain data from U_d separately

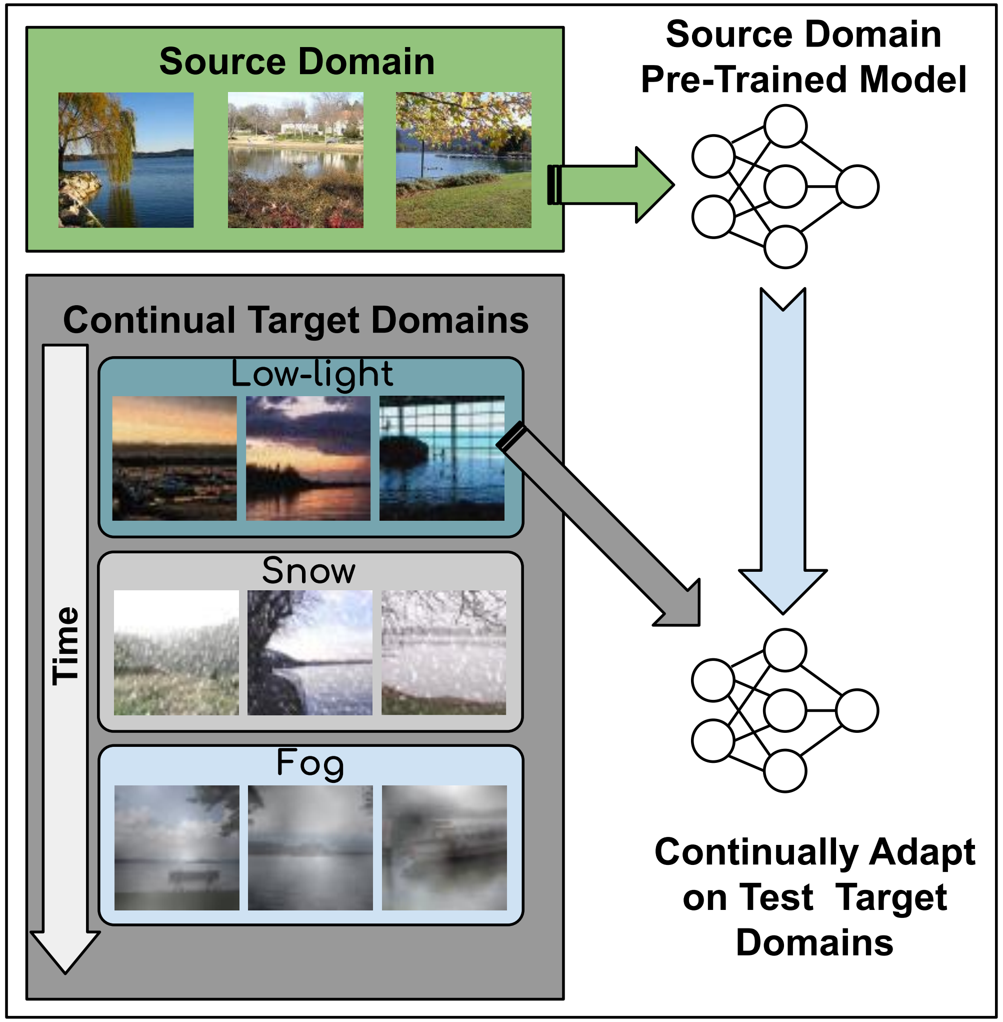
$$\theta_0 \rightarrow \theta_d$$

Lifelong Test-Time Adaptation

Aim: Initially start from θ_0. At time step t, continually adapt:

$$\theta_t \rightarrow \theta_{t+1}$$

Note: No information about change in domain



Related Prior Work

TENT and BACS

- Test entropy minimization (TENT) [Wang et al., 2021]
  - Adapts pre-trained model to test data
  - Updates trainable parameters in BN layers using entropy minimization
- Bayesian Adaptation for Covariate Shift (BACS) [Zhou et al., 2021]
  - Bayesian perspective for TTA naturally gives rise to a regularizer
    - BACS = TENT + regularizer
  - Computes approximate posterior of source model during training time

Wang et al., "Tent: Fully test-time adaptation by entropy minimization." *ICLR 2021*
Zhou et al., "Training on test data with Bayesian adaptation for covariate shift." *NeurIPS 2021*

CoTTA

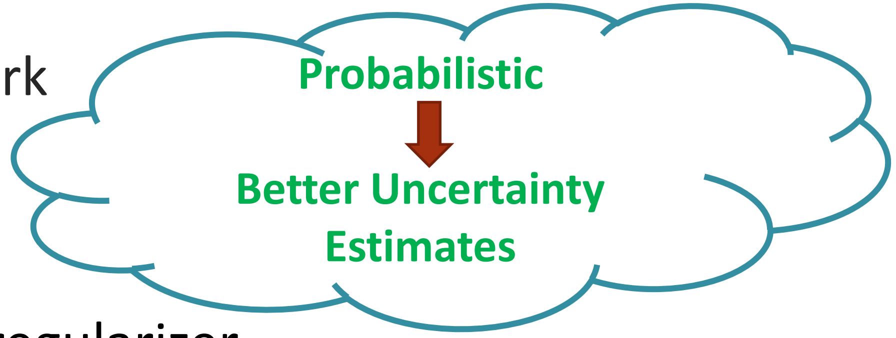
- Continual Test-Time Adaptation (CoTTA) [Wang et al., 2022]
  - Continually adapts pre-trained model to various target domain test data
  - Self-training framework that maintains weight-averaged teacher model
  - Augmentation-averaged prediction to improve quality of pseudo-labels
  - Stochastic restore to avoid long term performance deterioration

Wang et al., “Continual Test-Time Domain Adaptation.”, *CVPR 2022*

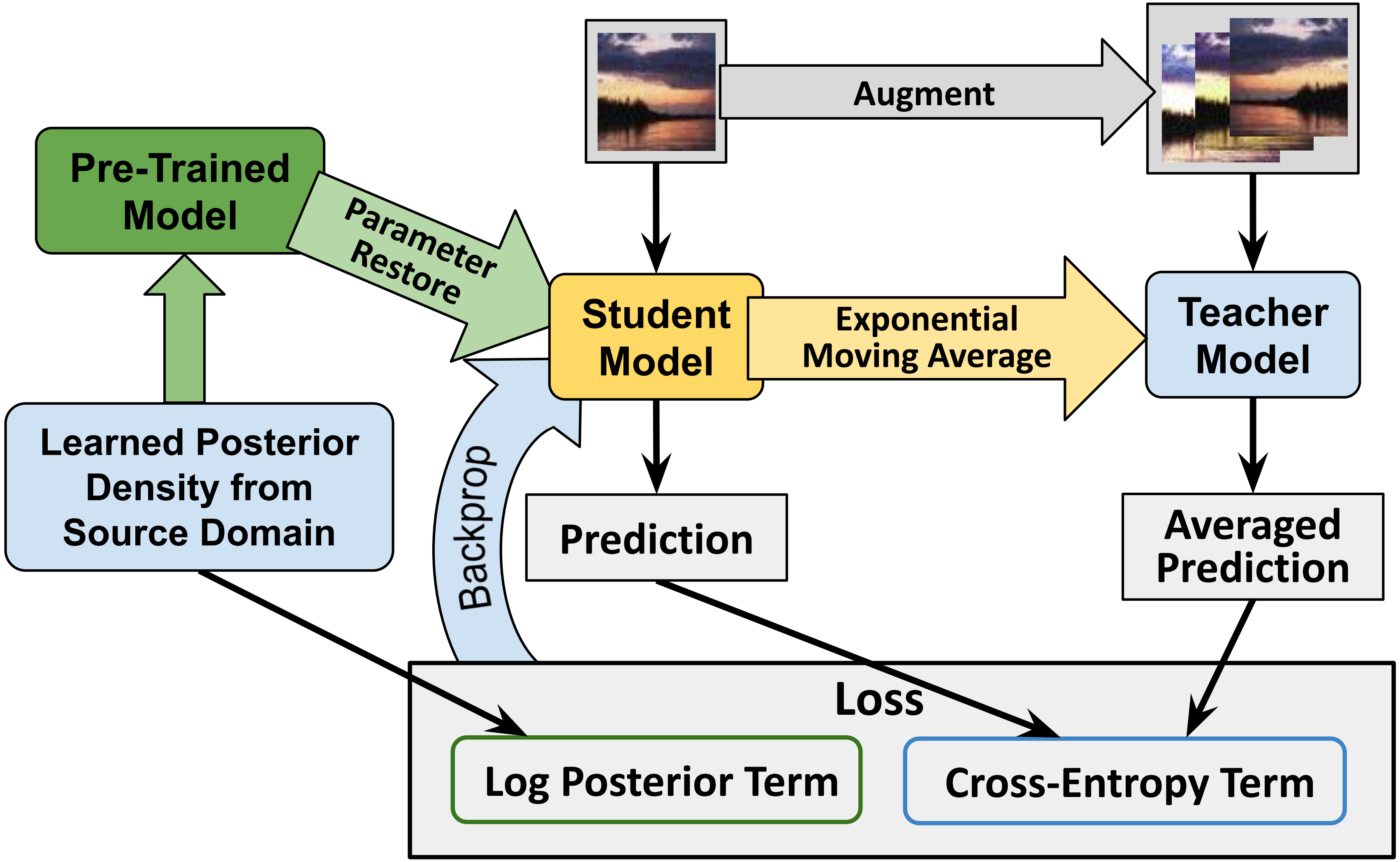
Proposed Approach

PETAL: Probabilistic lifElong Test-time Adaptation with seLf-training prior

- Probabilistic perspective for Lifelong Test-Time Adaptation
- CoTTA arises as a special case of our probabilistic framework
- Posterior for source training data → regularizer
- Naturally gives student-teacher self-training framework + regularizer
- Data driven Fisher Information Matrix (FIM) based restoration
- Principled use of approximate training posterior surpasses prior heuristic approaches



PETAL



Augment
Pre-Trained Model
Parameter Restore
Student Model
Exponential Moving Average
Teacher Model
Learned Posterior Density from Source Domain
Backprop
Prediction
Averaged Prediction
Loss
Log Posterior Term
Cross-Entropy Term

Proposed Self-Training for Bayesian SSL

Bayesian SSL with Student-Teacher Model

Dataset: $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}, \mathcal{U} = \{\tilde{\mathbf{x}}_m\}_{m=1}^{M}$

Posterior: $p(\theta|\mathcal{D}) \propto p(\theta|\psi) \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \theta)$

$$p(\theta|\psi) \propto p(\theta) \exp(-\lambda \mathrm{H}^{\mathrm{xe}}_{\theta,\psi}(y', y|\mathbf{x}))$$
$$= p(\theta) \exp(\lambda \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\psi), y'\sim p(y|\mathbf{x},\theta')}[\log p(y|\mathbf{x}, \theta)])$$

Cross-Entropy

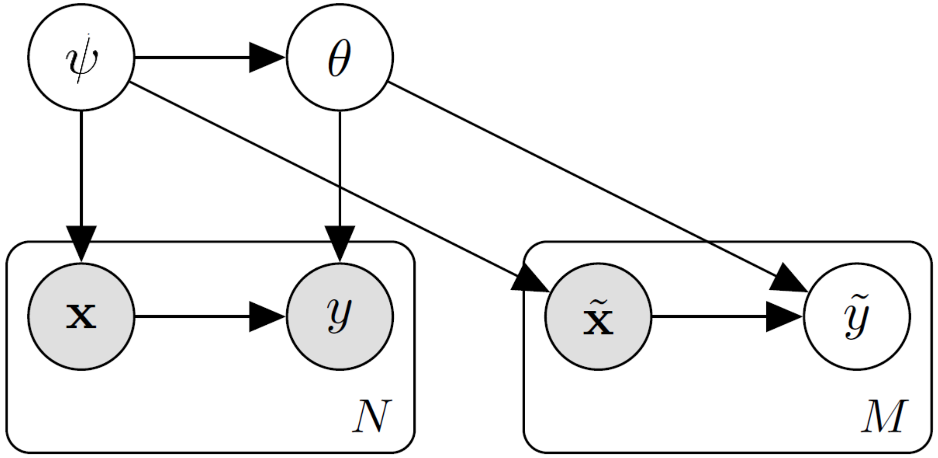
Prior is partly defined from the data

H^{xe}: conditional cross entropy of labels conditioned on inputs

θ': exponential moving average of student model θ

$$\theta'_{t+1} = \delta\theta'_t + (1 - \delta)\theta_{t+1}$$

y': pseudo-labels corresponding to input x obtained from teacher model θ'



✓ Error Accumulation
✓ Catastrophic Forgetting

Proposed Self-Training for Covariate Shift

Covariate Shift with Student-Teacher Model

Dataset: $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}, \mathcal{U} = \{\tilde{\mathbf{x}}_m\}_{m=1}^{M}$

Posterior: $p(\theta|\mathcal{D}) \propto p(\theta|\psi) \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \theta)$

$$p(\theta|\psi, \tilde{\psi}) \propto p(\theta) \exp(-\lambda \mathrm{H}^{\mathrm{xe}}_{\theta,\psi}(y', y|\mathbf{x})) \exp(-\bar{\lambda} \mathrm{H}^{\mathrm{xe}}_{\theta,\tilde{\psi}}(y', y|\tilde{\mathbf{x}}))$$

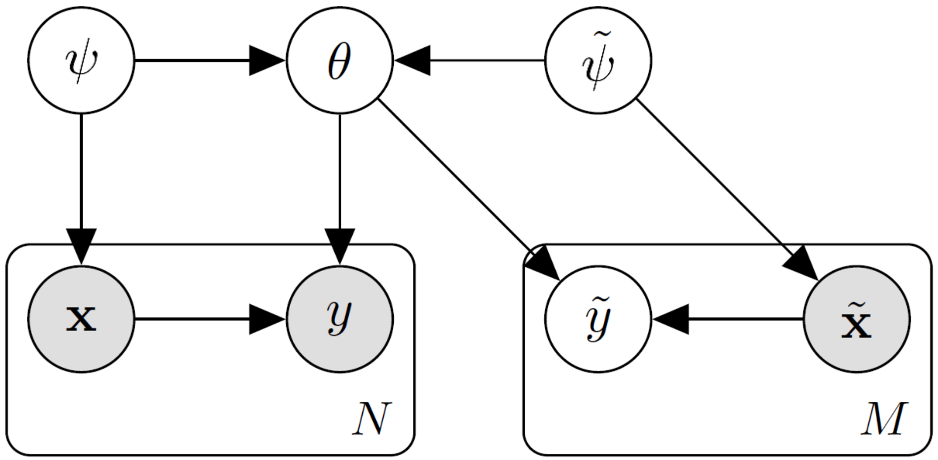
Additional Factor for Covariate Shift

H^{xe}: conditional cross entropy of labels conditioned on inputs

θ': exponential moving average of student model θ

$$\theta'_{t+1} = \delta\theta'_t + (1-\delta)\theta_{t+1}$$

y': pseudo-labels corresponding to input x obtained from teacher model θ'



Learning Objective

- Apply plug-in approximation and further simplify to get

$$\log p(\theta|\mathcal{D},\mathcal{U}) = \log q(\theta) - \frac{\bar{\lambda}}{M}\sum_{m=1}^{M} \mathrm{H}^{\mathrm{xe}}_{\theta,\bar{\psi}}(y', y|\bar{\mathbf{x}})$$

- Here, $q(\theta) \approx p(\theta|\mathcal{D})$ is an approximate posterior learned during training time itself
- Obtain adapted parameters by maximizing the equation above
- Like BACS (Zhou and Levine 2021), use SWAG diagonal for approximate posterior
  - SWAG-diag: Posterior approximated with Gaussian with diag. cov. [Maddox et al., 2019]

Zhou and Levine, "Training on test data with Bayesian adaptation for covariate shift." *NeurIPS 2021*
Maddox et al., "A Simple Baseline for Bayesian Uncertainty in Deep Learning." *NeurIPS 2019*

Fisher Information Based Restoration

- We propose a data driven parameter restoration in order to improve upon random restoration
- Fisher information matrix (FIM) of student model parameterized by θ
- For a given time step of L many input data, we consider following diagonal approximation of FIM

$$F = \mathrm{Diag}\left(\frac{1}{L}\sum_{l=1}^{L} \nabla \log p(\theta|\mathcal{D},\mathcal{U}) \nabla \log p(\theta|\mathcal{D},\mathcal{U})^{\mathrm{T}}\right)$$

- Using this, parameter restoration mask becomes

$$\gamma = \mathrm{quantile}(F, \delta)$$

$$\mathbf{m}_i = \begin{cases} 1, & \text{if } F_i < \gamma \\ 0, & \text{otherwise.} \end{cases}, d = 1, \cdots, D.$$

Here, γ is threshold value which is δ-quantile of F

Experimental Results

CIFAR-10C Results

Time	$t \longrightarrow$															
Method	Gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic	pixelate	jpeg	Mean
Source	72.33	65.71	72.92	46.94	54.32	34.75	42.02	25.07	41.30	26.01	9.30	46.69	26.59	58.45	30.30	43.51
BN Adapt	28.08	26.12	36.27	12.82	35.28	14.17	12.13	17.28	17.39	15.26	8.39	12.63	23.76	19.66	27.30	20.44
Pseudo-label	26.70	22.10	32.00	13.80	32.20	15.30	12.70	17.30	17.30	16.50	10.10	13.40	22.40	18.90	25.90	19.80
TENT-online$^+$	24.80	23.52	33.04	11.93	31.83	13.71	10.77	15.90	16.19	13.67	7.86	12.05	21.98	17.29	24.18	18.58
TENT-continual	24.80	**20.60**	28.60	14.40	31.10	16.50	14.10	19.10	18.60	18.60	12.20	20.30	25.70	20.80	24.90	20.70
CoTTA	23.92	21.40	25.95	11.82	27.28	12.56	10.48	15.31	14.24	13.16	7.69	11.00	18.58	13.83	17.17	16.29 (0.02)
PETAL (S-Res)	23.44	21.20	**25.50**	11.80	**27.22**	12.54	10.45	15.14	14.31	12.89	7.61	10.72	18.42	13.83	17.37	16.16 (0.02)
PETAL (FIM)	**23.42**	21.13	25.68	**11.71**	27.24	**12.19**	**10.34**	**14.76**	**13.91**	**12.65**	**7.39**	**10.49**	**18.09**	**13.36**	**16.81**	**15.95** (0.04)

Classification error rate (%) for CIFAR10-to-CIFAR10C with the highest corruption of severity level 5

CIFAR-100C Results

Time	$t \longrightarrow$															
Method	Gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic	pixelate	jpeg	Mean
Source	73.00	68.01	39.37	29.32	54.11	30.81	28.76	39.49	45.81	50.30	29.53	55.10	37.23	74.69	41.25	46.45
BN Adapt	42.14	40.66	42.73	27.64	41.82	29.72	27.87	34.88	35.03	41.50	26.52	30.31	35.66	32.94	41.16	35.37
Pseudo-label	38.10	36.10	40.70	33.20	45.90	38.30	36.40	44.00	45.60	52.80	45.20	53.50	60.10	58.10	64.50	46.20
TENT-continual	**37.20**	**35.80**	41.70	37.90	51.20	48.30	48.50	58.40	63.70	71.10	70.40	82.30	88.00	88.50	90.40	60.90
CoTTA	40.09	37.67	39.77	26.91	37.82	28.04	26.26	32.93	31.72	40.48	24.72	26.98	32.33	28.08	33.46	32.48 (0.02)
PETAL (S-Res)	38.37	36.43	38.69	**25.87**	37.06	27.34	25.55	32.10	31.02	38.89	24.38	**26.38**	31.79	27.38	32.98	31.62 (0.04)
PETAL (FIM)	38.26	36.39	**38.59**	25.88	**36.75**	**27.25**	**25.40**	**32.02**	**30.83**	**38.73**	**24.37**	26.42	**31.51**	**26.93**	**32.54**	**31.46** (0.04)

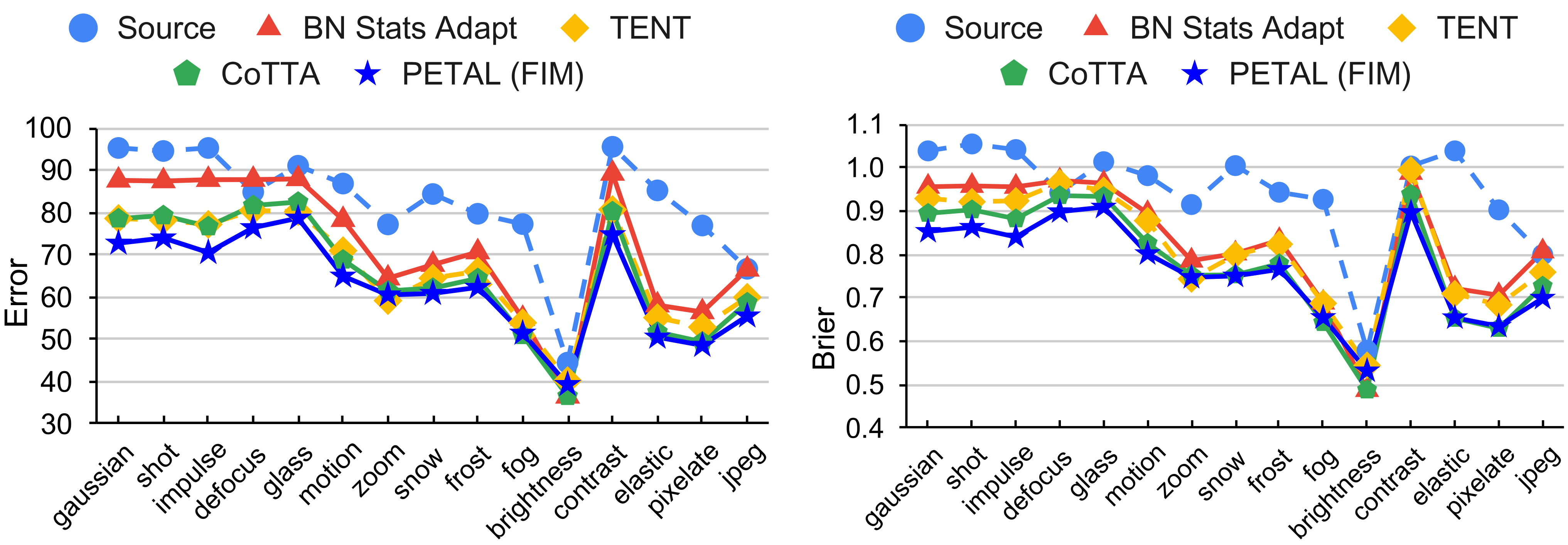
Classification error rate (%) for CIFAR100-to-CIFAR100C with the highest corruption of severity level 5

ImageNet-C Results

Metric \ Method	Source	BN Adapt	TENT	CoTTA	PETAL (FIM)
Error (%)	82.35	72.07	66.52	63.18	**62.71**
NLL	5.0701	3.9956	3.6076	3.3425	**3.3252**
Brier	0.9459	0.8345	0.8205	0.7681	**0.7663**

Classification error rate (%) for ImageNet-to-ImageNetC averaged over all corruption types and over 10 diverse corruption orders with the highest corruption of severity level 5

ImageNet-C Results



ImageNet-to-ImageNetC results averaged over 10 different corruption orders with level 5 corruption severity

ImageNet-3DCC Results

Metric \ Method	Source	BN Adapt	TENT	CoTTA	PETAL (FIM)
Error (%)	69.21	67.32	95.93	59.91	**59.61**
NLL	5.0701	3.9956	3.6076	3.3425	**3.3252**
Brier	3.9664	3.7163	19.0408	3.2636	**3.2560**

Classification error rate (%) for ImageNet-to-ImageNet3DCC averaged over all corruption types and over 10 diverse corruption orders with the highest corruption of severity level 5

Summary

- Focused on lifelong test-time adaptation (LTTA) set-up
- Addressed the problem of LTTA from a probabilistic perspective
- Proposed a novel approach PETAL:
  - Naturally gives student-teacher framework + regularizer
  - Better Uncertainty Estimates
  - Can be extended for Bayesian SSL when labeled and unlabeled data distributions are not same
- Developed a data-driven Fisher information matrix based parameter restoration
- Achieved state-of-the-art results on various lifelong TTA benchmark datasets