Samsung Advanced Institute of Technology

Rethinking Entropy in Test-Time Adaptation: The Missing Piece from Energy Duality

Mincheol Park[1], *Heeji Won*[2], *Won Woo Ro*[3], *and Suhyun Kim*[4]

[1]Samsung Advanced Institute of Technology, [1]Samsung Electronics,
[2]Korea University, [3]Yonsei University, [4]Kyung Hee University

SAMSUNG

NEURAL INFORMATION PROCESSING SYSTEMS

KOREA UNIVERSITY 1905

YONSEI UNIVERSITY 1885 연세대학교

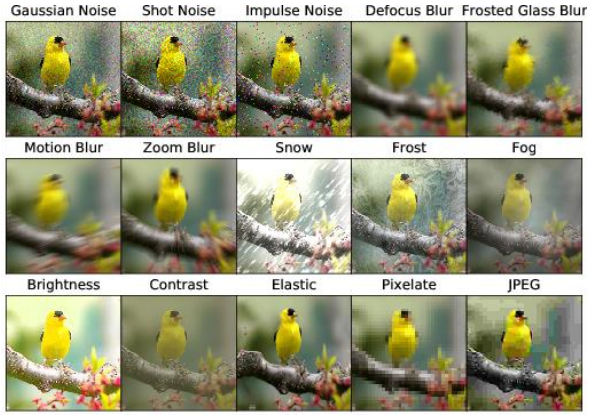
KYUNG HEE UNIVERSITY

Covariate Shifts and Fully Test-Time Adaptation

- **Assumption that test data comes from the same distribution as training data is often broken in practice**

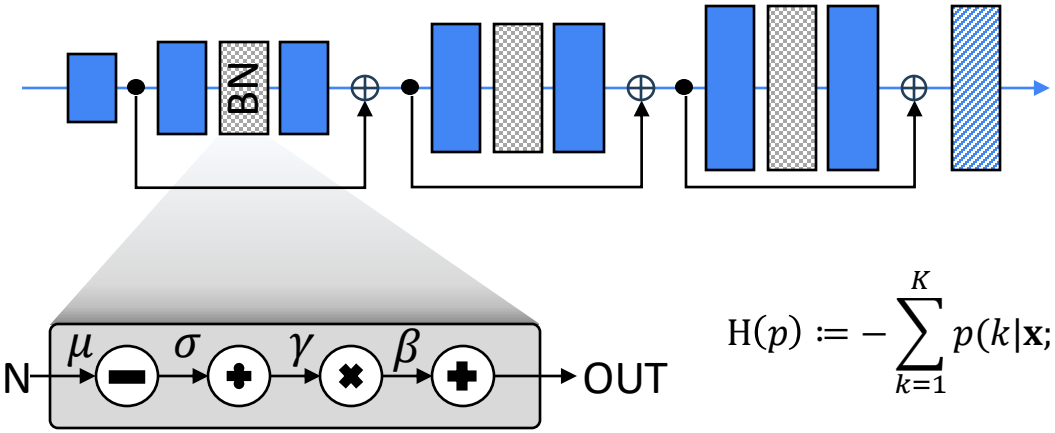
$D_S = \{(\mathbf{x}_i^s, \mathrm{y}_i^s)\}_{i=1}^N$ $(\mathbf{x} \in \mathbb{R}^D, \mathrm{y} \in \mathbb{R})$ **"Goldfinch"**

$D_T = \{(\mathbf{x}_i^t, \mathrm{y}_i^t)\}_{i=1}^M$



$\mathbf{x}_{N+1} \neq \mathbf{x}$

Existing Test-time adaptation (TTA) solution[1]



Goal: $p_s(\mathbf{x}) \cong p_t(\mathbf{x})$

1. Wang et al., "Tent: Fully Test-Time Adaptation by Entropy Minimization," *ICLR, 2021*

First Motivation in Entropy Minimization

- **$p_s(\mathbf{x}) \cong p_t(\mathbf{x})$ is necessary for mitigation of covariate shifts**
  - It is necessary to introduce a quantity that quantifies how the data likely belongs to the marginal parameterized $\boldsymbol{\theta}$
  - Energy maps data to a deterministic scalar by summing over the probable classes
    - ✓ a larger negative value represents more likely (or highly observable) data under the distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$
- **EM pushes the energy to the logit of the most confident class, while confident classes become confident**

$$p(y|\mathbf{x};\boldsymbol{\theta}) = \frac{\exp(f_{\boldsymbol{\theta}}(\mathbf{x})[\mathrm{y}])}{\sum_c \exp(f_{\boldsymbol{\theta}}(\mathbf{x})[\mathrm{c}])} \quad \textbf{(Softmax)}$$

$$E_{\boldsymbol{\theta}}(\mathbf{x}) \triangleq -\log \sum_{k=1}^{K} \exp(f_{\boldsymbol{\theta}}(\mathbf{x})[\mathrm{k}]) \quad \textbf{(Energy)}$$

$$\mathrm{H}(p) \coloneqq -\sum_{k=1}^{K} p(k|\mathbf{x};\boldsymbol{\theta}) \log p(k|\mathbf{x};\boldsymbol{\theta}) \quad \textbf{(Entropy)}$$

Lemma 1. Conjugate Relation

Suppose $\mathbf{z}$ represents the model's logit, and $\mathbf{g}$ denotes the gradient of the concave function $E_{\boldsymbol{\theta}}$ w.r.t. the logit $\mathbf{z}$. The concave conjugate of $E_{\boldsymbol{\theta}}(\mathbf{z})$ is defined as $E^*_{\boldsymbol{\theta}}(\mathbf{g}) = \min_{\mathbf{z}}\{\mathbf{g}^T\mathbf{z} - E_{\boldsymbol{\theta}}(\mathbf{z})\}$.

Then, the gradient $\mathbf{g}$ corresponds negatively to the Softmax, i.e., $\mathbf{g} = \nabla_{\mathbf{z}} E_{\boldsymbol{\theta}}(\mathbf{z}) = -\mathbf{p}(\mathbf{x})$, and the conjugate function $E^*_{\boldsymbol{\theta}}(\mathbf{g})$ becomes the negative entropy of $\mathbf{p}(\mathbf{x})$:

$$E^*_{\boldsymbol{\theta}}(\mathbf{g}) = H(\mathbf{p}) = -\mathbf{p}(\mathbf{x})^T \log \mathbf{p}(\mathbf{x})$$

Lemma 2. Fenchel-Moreau Theorem

Primal function $E_{\boldsymbol{\theta}}(\mathbf{z})$ and its conjugate function $E^*_{\boldsymbol{\theta}}(\mathbf{g})$ exhibit bi-duality. The primal function can be completely recovered from its conjugate function $E^*_{\boldsymbol{\theta}}(\mathbf{g})$. Thus, energy and entropy satisfy the following relationship:

$$E_{\boldsymbol{\theta}}(\mathbf{z}) = \min_{\mathbf{p}}\{-\mathbf{p}^T\mathbf{z} - H(\mathbf{p})\}$$

When $H(\mathbf{p}) \to 0$, $E_{\boldsymbol{\theta}}(\mathbf{z}) \to -z_{k^*}$

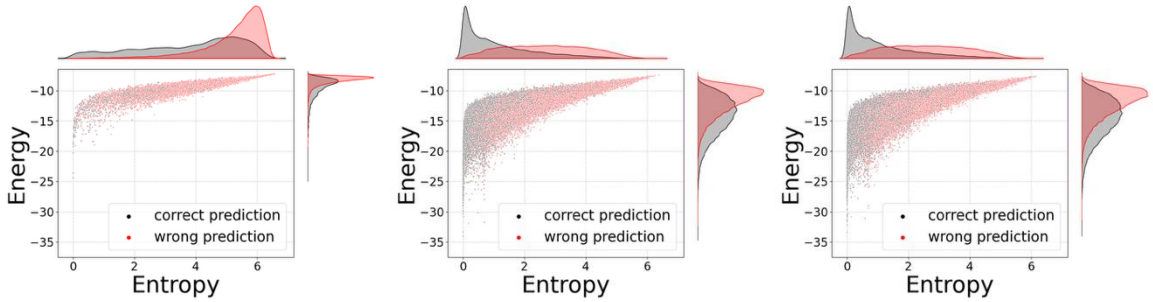
$\mathbf{p}$ is more confident[1], ideally converge to one-hot

1. Choi et al., "Adaptive Energy Alignment for Accelerating Test-Time Adaptation," *ICLR, 2025*

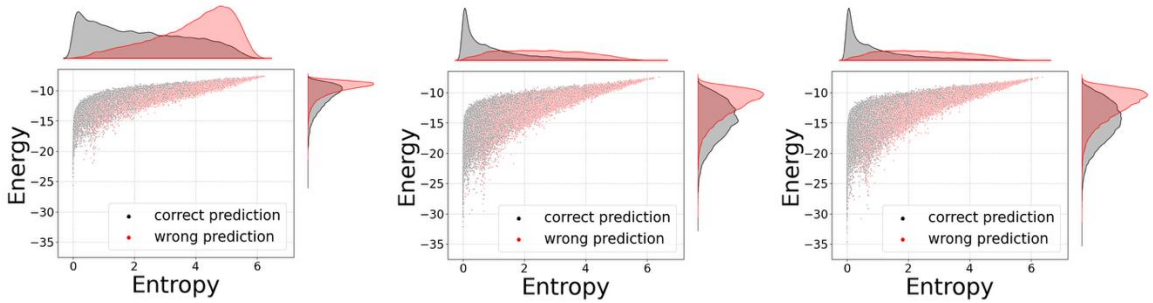
Second Motivation in Entropy Minimization

- **To make entropy approach zero, an additional goal should be to guide the logits**

No adapt | EM (SAR) | ReTTA



Contrast (severity 5) in ImageNet-C[1]



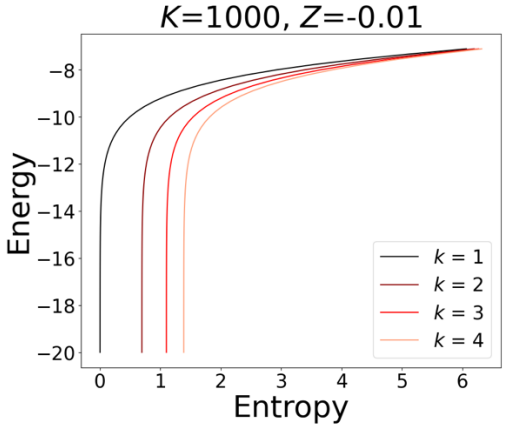
JPEG (severity 5) in ImageNet-C[1]

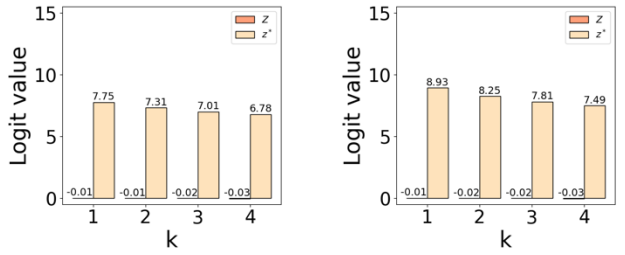
Theorem 1. Energy-entropy equation

Suppose the logit of the model $f_{\boldsymbol{\theta}}$ is defined over K classes, where k classes are assigned a primary logit z^* with strong influence, and the remaining $K-k$ classes share a singular logit Z with minimal influence. Then, the closed-form equation for the energy-entropy relationship based on the conditioned logits is given by:

$$H(E_{\boldsymbol{\theta}}) = -(1 - C(k)e^{E_{\boldsymbol{\theta}}}) \log\left(\frac{1 - C(k)e^{E_{\boldsymbol{\theta}}}}{k}\right) - C(k)e^{E_{\boldsymbol{\theta}}}(Z + E_{\boldsymbol{\theta}})$$

where $C(k) = (K-k)e^{Z}$





(a) Contrast (b) JPEG

1. Hendrycks and Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," *ICLR, 2019*

Objective from Energy-Based Modeling (EBM)

TTA through EBMs[1]

$$E_{\boldsymbol{\theta}}(\mathbf{x}) \triangleq -\log \sum_{k=1}^{K} \exp(f_{\boldsymbol{\theta}}(\mathbf{x})) \quad \text{(Free energy)}$$

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{\exp(-E_{\boldsymbol{\theta}}(\mathbf{x}))}{Z_{\boldsymbol{\theta}}} \quad \text{Marginal density}$$

Goal

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x})} \log p_{\boldsymbol{\theta}}(\mathbf{x}) = -\mathbb{E}_{p(\mathbf{x})} \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}$$

$$= \underbrace{-\mathbb{E}_{p(\mathbf{x})} \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\mathbf{x})}_{\text{Positive}} + \underbrace{\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x})} \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\mathbf{x})}_{\text{Negative}} \quad \text{(Contrastive)}$$

SGLD[2] $\quad \mathbf{x}^{k+1} \leftarrow \mathbf{x}^{k} - \frac{\epsilon^2}{2} \nabla_{\mathbf{x}} E_{\boldsymbol{\theta}}(\mathbf{x}^{k}) + \epsilon \mathbf{z}$

(Assume MCMC step=1) Generative process

Lemma 3. Fisher Divergence

The one-step SGLD update initialized from $\mathbf{x} \sim p(\mathbf{x})$ approximates the gradient of the Fisher divergence between the true distribution $p(\mathbf{x})$ and the model distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ as follows:

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x})} \log p_{\boldsymbol{\theta}}(\mathbf{x}) \cong \frac{\epsilon^2}{2} \nabla_{\boldsymbol{\theta}} D_F(p(\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{x})) + \mathrm{o}(\epsilon^2)$$

Derivation of objective based on EBM

$$D_F\big(p(\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{x})\big) = \mathbb{E}_{p(\mathbf{x})} \left[\frac{1}{2} \| \nabla_{\mathbf{x}} \log p(\mathbf{x}) - \nabla_{\mathbf{x}} \log p_{\boldsymbol{\theta}}(\mathbf{x}) \|^2 \right]$$

(Score Matching)[3]

$$\cong \mathbb{E}_{p(\mathbf{x})} [\mathrm{Tr}(\nabla_{\mathbf{x}}^2 \log p_{\boldsymbol{\theta}}(\mathbf{x})) + \frac{1}{2} \| \nabla_{\mathbf{x}} \log p_{\boldsymbol{\theta}}(\mathbf{x}) \|^2]$$

$$= \mathbb{E}_{p(\mathbf{x})} [\sum_{i=1}^{d} \frac{\partial^2 E_{\boldsymbol{\theta}}(\mathbf{x})}{(\partial x_i)^2} + \frac{1}{2} \sum_{i=1}^{d} \left(\frac{\partial E_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_i} \right)^2]$$

1. Yuan et al., "TEA: Test-time Energy Adaptation," *CVPR*, 2024
2. Du et al., "Improved Contrastive Divergence Training of Energy-Based Models," *ICML*, 2021
3. Hyvarinen and Dayan, "Estimation of Non-normalized Statistical Models by Score Matching," *Journal of Machine Learning Research 6(4)*, 2005

Sliced Score Matching

- Energy-based modeling allow the model to reshape its likelihood landscape in response to data
- Generative sampling is unstable → Sampling-free loss function
- Computing score matching requires evaluating the trace of Hessian → sensitive the sharp local curvature
- SSM matches inner products of score functions → scalable at high-dimensional data

$$D_F\big(p(\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{x})\big) = \mathbb{E}_{p(\mathbf{x})}[\mathrm{Tr}(\underbrace{\nabla_{\mathbf{x}}^2 \log p_{\boldsymbol{\theta}}(\mathbf{x})}_{\text{Hessian}}) + \frac{1}{2}\|\nabla_{\mathbf{x}} \log p_{\boldsymbol{\theta}}(\mathbf{x})\|^2]$$

Sliced Score Matching (SSM)[1]

$$D_{SF}\big(p(\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{x})\big) = \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{v})}\left[\frac{1}{2}\left\|\mathbf{v}^T\boldsymbol{\nabla}_{\mathbf{x}}\log p(\mathbf{x}) - \mathbf{v}^T\boldsymbol{\nabla}_{\mathbf{x}}\log p_{\hat{\boldsymbol{\theta}}}(\mathbf{x})\right\|^2\right]$$

$$\ell_{SSM}(\boldsymbol{\theta}) = \frac{1}{|\mathfrak{B}_t|}\sum_{\mathbf{x}\in\mathfrak{B}_t}\left[\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2 E_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_i \partial x_j}v_i v_j + \frac{1}{2}\sum_{i=1}^{d}\left(\frac{\partial E_{\boldsymbol{\theta}}(\mathbf{x})}{\partial x_i}v_i\right)^2\right]$$

1. Song et al., "Sliced Score Matching: A Scalable Approach to Density and Score Estimation," *UAI, 2020*

Targeted Class Convergence and ReTTA Loss

- **Guidance of the logits toward a zero-entropy region**
  - By leveraging the model's discriminative power, ReTTA treats the most probable class as the target class
- **Total loss for TTA, ReTTA**
  - $\ell_{ReTTA}(\boldsymbol{\theta}) = \ell_{EM}(\boldsymbol{\theta}) + \lambda_1(\alpha)\ell_{SSM}(\boldsymbol{\theta}) + +\lambda_2\ell_{TCC}(\boldsymbol{\theta})$
  - $\lambda_1(\alpha)$ is a self-adjusting coefficient

Targeted Class Convergence (TCC)

$$\ell_{TCC}(\boldsymbol{\theta}) = \frac{1}{|\mathfrak{B}_t|}\sum_{\mathbf{x}\in\mathfrak{B}_t}\left[-\log\left(\frac{\exp(f_\theta(\mathbf{x})[\tilde{y}])}{\sum_k \exp(f_\theta(\mathbf{x})[k])}\right)\right]$$

$$\min_{\alpha\in[0,1]} \|\alpha\nabla_{\boldsymbol{\theta}}\underbrace{\ell_{EM}(\boldsymbol{\theta})}_{\text{Entropy}} + (1-\alpha)\nabla_{\boldsymbol{\theta}}\ell_{SSM}(\boldsymbol{\theta})\|_2^2$$ **(MOO problem)** [1]

Self-adjusting Coefficient

$$\alpha = \frac{(\nabla_{\boldsymbol{\theta}}\ell_{SSM}(\boldsymbol{\theta}) - \nabla_\theta\ell_{EM}(\boldsymbol{\theta}))^T \cdot \nabla_\theta\ell_{SSM}(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}}\ell_{EM}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\ell_{SSM}(\boldsymbol{\theta})\|^2}$$

$$\lambda_1(\alpha) = \max\left(\min\left(\frac{1-\alpha}{\alpha}, 1\right), 0\right)$$

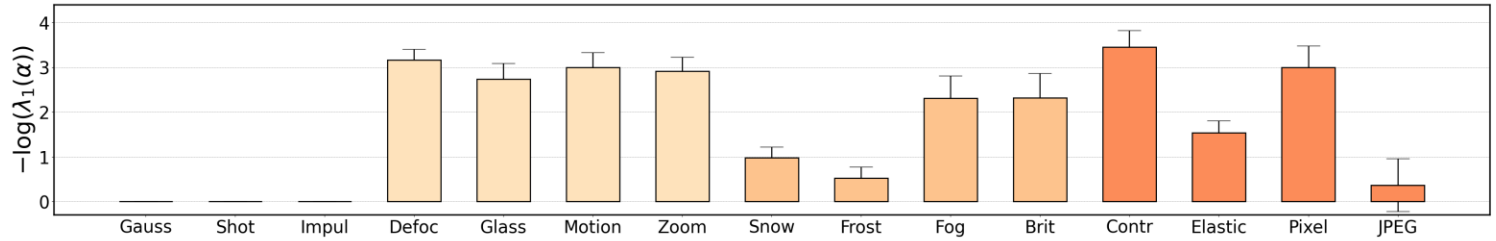


Figure 2: Breakdown of the self-adjusting coefficient λ_1 during total TTA iterations on ImageNet-C (severity 5), based on Table 1. The negative-log scale has zero corresponding to $\lambda_1 = 1$, and higher values indicate near-zero λ_1. The four colors represent Noise, Blur, Weather, and Digital groups.

1. Sener and Koltun, "Multi-Task Learning as Multi-Objective Optimization," *NeurIPS, 2018*

Experimental Results

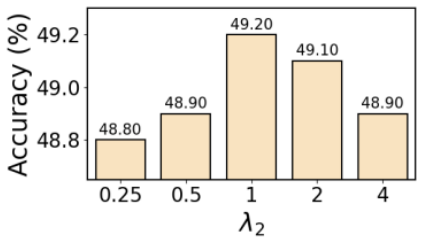
- **ReTTA outperforms current methods under covariate shifts for real-world dataset**
  - ReTTA shows also better performance under label distribution shifts in test scenarios

Mild	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
ResNet-50 (BN)	2.2	2.9	1.8	17.9	9.8	14.8	22.5	16.9	23.3	24.4	58.9	5.4	16.9	20.7	31.7	18.0
MEMO	7.5	8.8	8.9	19.8	13	20.7	27.7	25.3	28.7	32.2	61.0	11.0	23.8	33.0	37.6	23.9
Tent	29.2	31.2	30.1	28.1	27.7	41.4	49.4	47.2	41.5	57.7	67.4	29.2	54.8	58.5	52.4	43.1
EATA	34.9	37.1	35.8	33.4	33.0	47.1	52.7	51.6	45.7	60.0	68.1	44.4	57.9	60.6	55.1	47.8
SAR	30.6	30.6	31.3	28.5	28.5	41.9	49.4	47.1	42.2	57.5	67.3	37.8	54.6	58.4	52.1	43.9
DeYO	35.6	37.9	37.1	33.8	34.1	48.5	52.8	52.7	**46.4**	60.6	68.0	46.1	58.4	61.5	55.7	48.6
TEA*	16.8	17.5	17.5	15.8	16.0	27.3	39.9	35.3	33.9	49.0	65.7	17.9	45.1	50.2	41.3	32.6
AEA	26.2	26.8	27.3	24.2	20.8	40.3	48.1	47.3	41.4	56.0	65.7	9.5	53.4	56.7	49.5	39.5
ReTTA (ours)	**37.3**$_{\pm 0.0}$	**39.7**$_{\pm 0.2}$	**38.9**$_{\pm 0.2}$	**34.5**$_{\pm 0.3}$	**34.1**$_{\pm 0.0}$	**49.3**$_{\pm 0.2}$	**53.1**$_{\pm 0.2}$	**52.7**$_{\pm 0.1}$	46.1$_{\pm 0.2}$	**60.7**$_{\pm 0.1}$	**68.2**$_{\pm 0.1}$	**47.6**$_{\pm 0.3}$	**58.6**$_{\pm 0.0}$	**61.5**$_{\pm 0.0}$	**56.0**$_{\pm 0.0}$	**49.2**$_{\pm 0.0}$

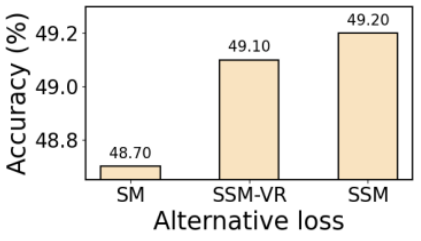
Table 1: Comparisons with baseline TTA methods on ImageNet-C at severity level 5 under mild scenario in terms of accuracy (%). * TEA was not publicly reported and was tested directly.

Label Shifts	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
ResNet-50 (GN)	17.9	19.9	17.9	19.7	11.3	21.3	24.9	40.4	47.4	33.6	69.3	36.3	18.7	28.4	52.2	30.6
MEMO	18.4	20.6	18.4	17.1	12.7	21.8	26.9	40.7	46.9	34.8	69.6	36.4	19.2	32.2	53.4	31.3
Tent	3.6	4.2	4.4	16.5	5.9	26.9	28.4	17.9	26.2	2.3	72.2	46.1	7.3	52.3	56.2	24.7
EATA	25.7	28.6	24.8	18.5	19.6	24.1	28.4	35.3	33.0	41.2	65.2	33.3	28.0	42.4	43.1	32.7
SAR	33.7	36.9	35.3	19.3	20.3	33.8	29.8	21.9	44.7	34.9	71.9	46.7	6.6	52.3	56.2	36.3
DeYO	42.5	44.9	43.8	22.2	16.3	41.0	13.2	52.2	**51.5**	39.7	73.4	**52.6**	46.9	59.3	59.3	43.9
TEA*	0.4	0.4	0.4	0.2	0.1	0.4	1.2	1.1	1.3	0.4	13.5	0.5	0.3	0.3	5.0	1.7
ReTTA (ours)	**42.7**$_{\pm 0.3}$	**45.1**$_{\pm 0.1}$	**44.2**$_{\pm 0.2}$	**29.4**$_{\pm 2.5}$	**22.9**$_{\pm 5.8}$	**41.1**$_{\pm 0.1}$	**34.4**$_{\pm 14.4}$	**52.8**$_{\pm 0.5}$	51.1$_{\pm 0.1}$	**58.5**$_{\pm 0.2}$	**73.5**$_{\pm 0.1}$	49.8$_{\pm 0.2}$	**48.4**$_{\pm 0.7}$	**59.8**$_{\pm 0.3}$	**59.3**$_{\pm 0.0}$	**47.5**$_{\pm 0.4}$
VitBase (LN)	9.4	6.7	8.3	29.1	23.4	34.0	27.1	15.8	26.4	47.4	54.7	44.0	30.5	44.5	47.6	29.9
MEMO	21.6	17.4	20.6	37.1	29.6	40.6	34.4	25.0	34.8	55.2	65.0	54.9	37.4	55.5	57.7	39.1
Tent	33.9	1.8	27.2	54.8	52.9	58.6	54.3	12.4	11.7	69.7	76.3	66.3	59.6	69.7	66.6	47.7
EATA	36.2	34.7	35.5	43.4	44.3	49.3	48.5	53.2	53.5	62.3	72.7	18.8	58.0	64.7	62.8	49.2
SAR	42.3	34.9	44.1	50.0	50.5	55.6	53.1	59.7	47.2	66.2	75.2	50.3	60.1	67.3	65.0	54.8
DeYO	53.5	36.0	54.6	57.6	58.7	63.7	46.2	67.6	66.0	**73.2**	77.9	66.7	69.0	73.5	70.3	62.3
TEA*	6.9	13.2	14.6	0.9	1.4	7.1	3.1	0.6	1.4	66.9	73.7	62.1	1.4	68.2	63.8	25.7
ReTTA (ours)	**54.0**$_{\pm 0.1}$	**55.0**$_{\pm 0.1}$	**55.2**$_{\pm 0.1}$	**57.8**$_{\pm 0.2}$	**58.7**$_{\pm 0.2}$	**64.7**$_{\pm 0.1}$	**58.5**$_{\pm 7.5}$	**69.0**$_{\pm 0.4}$	**67.1**$_{\pm 0.1}$	71.2$_{\pm 0.2}$	**77.9**$_{\pm 0.0}$	**67.6**$_{\pm 1.0}$	**69.8**$_{\pm 0.4}$	**74.1**$_{\pm 0.2}$	**71.6**$_{\pm 0.3}$	**64.8**$_{\pm 0.5}$

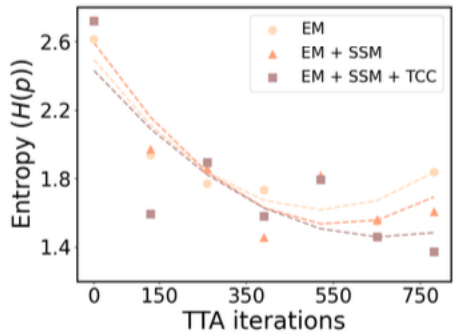
Table 2: Comparisons with baseline TTA methods on ImageNet-C (severity 5) under online label shifts (imbalance ratio=∞) in accuracy (%). * TEA was not publicly reported and was tested directly.



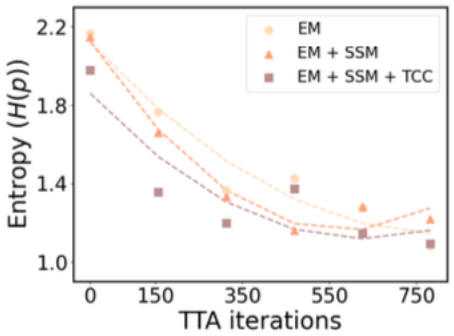
(a) Effects of varying λ_2



(b) Effects of alternatives for SSM



(a) Zoom



(b) Fog

Samsung Advanced Institute of Technology

Thank you

Exhibit Hall C.D.E (San Diego)

Mincheol