Learning from Unknown for Open-Set Test-Time Adaptation

Taki Hasan Rafi[1] Amit Agarwal[2] Hitesh L. Patel[2] Dong-Kyu Chae[1*]
[1]Hanyang University, Seoul, South Korea
[2]Oracle AI, USA
{takihr, dongkyu}@hanyang.ac.kr

Abstract

Deep learning models often struggle to maintain performance when the training and testing data come from different distributions. Test-time adaptation (TTA) addresses this by adapting a pre-trained model to an unlabeled target domain under distribution shifts. A more challenging setting is open-set TTA (OSTTA), where the target domain may contain unknown samples outside the source classes. Existing OSTTA methods primarily detect and discard such unknowns, relying only on known samples for adaptation. In this work, we argue that unknown samples can also provide valuable cues for improving adaptation. We propose ***LU-OSTTA*** *(**l**earning from **u**nknown for OSTTA), a simple yet effective framework that leverages both in-distribution and semantically useful out-of-distribution samples. Our approach introduces: (i) a class-conditioned dynamic energy threshold to separate OOD samples more reliably, (ii) an optimal transport–based pseudo-label refinement to mitigate noise under distribution shifts, and (iii) an adaptive prototype weighting strategy that emphasizes semantically aligned target samples while down-weighting harmful ones. Experimental results demonstrate that our LU-OSTTA consistently outperforms state-of-the-art TTA and OSTTA methods, highlighting the benefits of utilizing rather than discarding unknown samples. Our code is available at: https://github.com/takihasan/LU-OSTTA.*

1. Introduction

Deep learning has emerged as a key solution for computer vision tasks when the training and testing data are derived from the same distribution. However, in real-world settings, this assumption often does not hold due to discrepancies between the training and test samples [9, 24, 42]. This is primarily because train (source) and test (target) data are collected from different distributions. Moreover, target data can be corrupted during online batch adaptation.

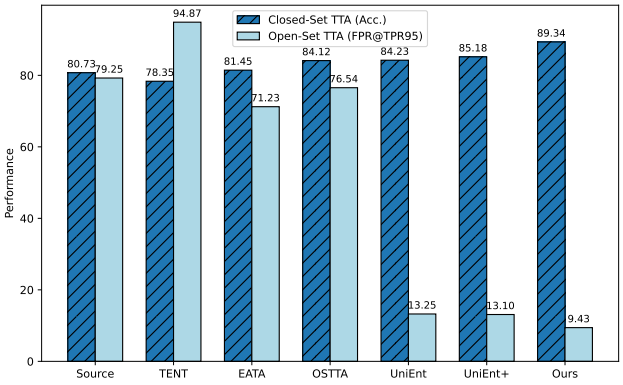

Figure 1. Existing TTA methods exhibit performance degradation when unknown classes are included, while OSTTA methods handle open-set distributions more effectively but still fall short. Our method significantly improves performance in the open-set setting. We compare Source [85], TENT [72], CoTTA [74], EATA [56], OSTTA [38], UniEnt [24], and UniEnt+ [24] on CIFAR-10-C.

To address these limitations, test-time adaptation (TTA) [7, 54, 60, 61, 74, 83, 88, 90] methods provide a holistic solution by enhancing the learning capabilities of a source model without requiring access to the source data, thereby improving robustness and consistency. They also help mitigate inherent privacy concerns during deployment.

Unlike traditional machine learning setup, where the model stops learning after the training dataset has been fully used, in the Test-Time Training [6, 23, 49, 68] (TTT) paradigm, model learns from test data for improving robustness to handle distribution shifts. In TTT settings, it introduces a self-supervised auxiliary rotation prediction task. However, it offers a source-free learning facility but requires altering the source training protocol, which eventually limits the practicability of the fully test-time adaptation task. Distribution shift can happen often, but further model tuning is not usually feasible due to lack of labels in target domain and computational resources. Meanwhile, TTA offers online adaptation of unlabeled target data without explicit fine-tuning [10]. Existing TTA methods are based on

*Corresponding author.

entropy minimization, where models adopt the samples that contribute to lower entropy. But the performance degrades if the model encounters open classes during adaptation.

On the other hand, out-of-distribution (OOD) detection [21,22,47,58] aims to detect samples not fully related to the training samples. But, models tend to classify OOD data as one of the in-distribution (ID) data that hinders the predictive performance [25, 53]. Existing methods are focused on dealing with covariate shift instead of semantic shift. Following recent literature, semantic shift refers to "out of the semantic space". In this work, we define *unknown samples* as those exhibiting semantic shift, while known samples include both ID and covariate-shifted variants.

Motivation. In practical deployment, the target domain often contains strong unknown OOD samples. Existing OSTTA methods [24, 38] mainly attempt to separate or discard such samples. However, we argue that not all unknowns are harmful: certain unknown samples can provide valuable semantic cues, reducing prediction entropy in ways similar to known samples. Prior works overlook these *semantically useful unknowns*, discarding them indiscriminately. As shown in Fig. 1, source models fail to adapt effectively under both closed- and open-set conditions. Standard TTA methods [56, 72] degrade in open-set scenarios due to incorrect normalization statistics and ineffective OOD filtering. OSTTA methods [24, 38] improve robustness but still underperform, as they cannot distinguish helpful unknowns from harmful ones. This motivates us to explicitly leverage semantically useful unknown samples for more effective open-set adaptation.

Our Approach. Motivated by these observations, we propose LU-OSTTA, a framework that handles unknown-class samples often encountered in real-world deployment. Instead of solely detecting and rejecting harmful samples during adaptation, our method leverages semantically useful unknown samples through a weighted mechanism for open-set adaptation. To enhance robustness in open-set settings, we introduce a dynamic energy threshold (DET), a class-conditioned variant of the energy threshold that dynamically adjusts on a per-class basis to better distinguish between weak and strong OOD samples. Since noisy labels are common under distribution shift and further increased when OOD samples are encountered during test time, we propose employing optimal transport (OT) to refine pseudo labels by aligning teacher and student predictions, thereby alleviating noise due to the open-set setting. To further separate and exploit semantically helpful unknowns, we propose an adaptive prototype weighting strategy, which maintains source and target prototype memory banks, and assigns semantic weights based on similarity to source prototypes. We incorporate contrastive loss with entropy-based regularization to ensure that only semantically useful and informative unknown samples contribute to adaptation, while down-weighting strong OOD samples. Lastly, to demonstrate the effectiveness of our proposed approach, we conducted comprehensive experiments under the open-set protocol described in [24].

Contributions. In summary, our main contributions are:

- We identify a critical limitation of prior works: discarding all unknown samples ignores their potential to enhance adaptation. We instead propose leveraging semantically helpful unknowns for improved open-set performance.
- We introduce a dynamic energy threshold (DET) that performs class-conditioned thresholding, enabling more reliable separation of known and unknown samples. We further refine pseudo-labels via optimal transport, improving supervision under distribution shift.
- We design adaptive prototype weighting that emphasizes semantically aligned unknown samples while down-weighting harmful ones during adaptation.
- Our method is evaluated on multiple open-set TTA protocols, where it exhibits significantly higher performance compared to other OSTTA and TTA methods.

2. Related Work

Test-Time Adaptation. Test-time adaptation (TTA) [56,57, 72, 83] is a promising direction in domain adaptation research. TTA aims to generalize to unseen target domains using a pre-trained source model. Unlike traditional domain adaptation, TTA does not assume that the source and target data come from the same distribution. A notable entropy-minimization-based TTA method, TENT [72], minimizes prediction entropy to update the affine parameters γ and β in batch normalization, thereby improving model adaptation. Based on this criterion, the entropy minimization loss function becomes the sole objective for updating batch normalization layers. EATA [56] introduces a sample selection strategy that excludes high-entropy samples from adaptation. SAR [57] adopts gradient-based sample selection, encouraging model weights to converge toward flat minima with small gradients. DeYO [39] further argues that entropy alone is insufficient for robust adaptation and proposes object-destructive transformations as an additional filtering step. In contrast to sample selection, some TTA methods [18, 51, 55, 69] focus on improving the optimization objective so that all samples can contribute to adaptation. Other approaches, such as refining pseudo labels [1, 20, 50] or batch normalization [59], have also demonstrated strong TTA performance. Recently, open-set TTA (OSTTA) has gained attention due to its practical significance. OSTTA methods [24, 38] account for the possibility that target data may contain unknown samples. These approaches detect both covariate-shifted in-distribution (csID)

and out-of-distribution (csOOD) data, and adapt only the csID samples to achieve robust OSTTA. Another work, ODS [92], considers label shift for open-world TTA. Moreover, TTA has been widely applied across diverse tasks, including object detection [8, 36, 63, 71], action recognition [78], super-resolution [19, 89], visual question answering [46, 76], and video understanding [2, 4, 44, 86], demonstrating its versatility as a practical solution.

OOD Detection. Out-of-distribution (OOD) detection is essential for building safe machine learning systems, as it enables the identification of samples that fall outside the training distribution. This capability is crucial for anomaly detection, open-set recognition, novelty detection, detecting and adapting to covariate domain shifts [3, 11, 32, 80]. OOD detection methods can be broadly categorized into several directions [80]. Post-hoc methods [67, 73] improve OOD detection without modifying pre-trained models. Other approaches, such as ReAct [66], adjust model activations to yield more informative energy scores. A widely used baseline is softmax-based detection, where prediction entropy is computed from the softmax output. Temperature scaling further calibrates model uncertainty by rescaling logits [66], while Generalized Entropy (GEN) [48] enhances softmax-based scoring. Another line of work leverages additional outlier data for OOD detection [35, 79, 87]. OE [30] encourages low model confidence for OOD samples, whereas WOODS [35] exploits unlabeled "wild" data to improve detection performance. Gradient-based approaches, such as ODIN [43] and MDS [40], apply input perturbations during inference as a pre-processing step to improve OOD discrimination. GradNorm [33] and Approximate Mass [26] instead use gradient norms to define OOD scores. More recently, test-time OOD detection has emerged as a way to improve robustness [81, 82]. AUTO [81] introduces a test-time modification strategy using stochastic gradient decent to reduce model confidence on potential OOD samples. OODD [82] proposes a dynamic OOD dictionary that accumulates OOD features during test time to enhance detection.

Open-Set Domain Adaptation. Open-Set Domain Adaptation (OSDA) addresses the challenge of unknown classes that exist in the target domain but are absent in the source domain. The goal of OSDA is to reduce the distribution gap between source and target domains while simultaneously detecting unknown classes. Existing approaches [45, 64, 65, 75] typically group unknown-class samples together and transfer knowledge from known classes for alignment. For example, OSBP [65] employs pseudo labels to guide classifier learning while rejecting unknown samples during source–target alignment. STA [45] introduces a weighting mechanism that distinguishes between known and unknown samples, gradually adjusting their importance for feature alignment. OVANET [64] adopts a one-vs-all classifier to model inter- and intra-class distances, minimizing sample entropy to better detect unknown samples. UADAL [34] proposes a segregation strategy to identify samples that are far from the source distribution. Finally, OSLPP [75] learns a discriminative common subspace between source and target features and progressively selects unknown-class samples to improve model training.

Discussion. Robustness under distribution shifts can be achieved by TTA, OOD detection, and OSDA methods; however, these methods can still impose limitations when encountering unknown samples. Existing OSTTA methods rely on rejecting unknown samples that can potentially help adaptation via relevant semantic information. On the other hand, OOD detection can detect unknown samples but this paradigm does not leverage them during adaptation. Moreover, OSDA methods do not offer privacy for source data that hinders the main goal of test-time settings. In our method, we address and combine all these issues by introducing a framework that leverages semantically useful unknown samples under an open-set TTA setting, alleviating the challenges in conventional TTA, OOD detection and OSDA methods.

3. Method

3.1. Preliminary

Task Definition. During TTA, a source model adapts to a target domain that contains n_t samples in the test set $\mathcal{D}_t$, which were not observed in the source domain. Let the source domain be $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with labeled samples, and the target domain be $\mathcal{D}_t = \{(x_i^t)\}_{i=1}^{n_t}$ containing only unlabeled samples, as y_t^i is not available during testing. Additionally, we designate $\mathcal{C}_s$ to be the set of source classes and $\mathcal{C}_t$ to be the target classes. In closed-set TTA settings, $\mathcal{C}_s$ is equal to $\mathcal{C}_t$. While in open-set TTA, $|\mathcal{C}_s| < |\mathcal{C}_t|$ always holds, where $\mathcal{C}_s \subset \mathcal{C}_t$, and $\mathcal{C}_t \setminus \mathcal{C}_s$ is called unknown classes. So, given a model f_θ pre-trained on $\mathcal{D}_s$, we aim to adapt it to $\mathcal{D}_t$ without accessing the target labels. Specifically, given a mini-batch $\mathcal{B}$ from the test set, the goal is to adapt samples from known classes $\mathcal{C}_s$ and also adapt classes from $\mathcal{C}_t$ that are weak OODs but considered as unknown.

Framework Overview. To address the open-set test-time adaptation (OSTTA) problem, we propose a framework, **LU-OSTTA**, which leverages both in-distribution and out-of-distribution samples under distribution shifts that are often encountered in the target domain. An overview of the framework is illustrated in Fig. 2. LU-OSTTA is composed of three key modules: **(1)** dynamic-energy threshold, **(2)** OT-based pseudo-label refinement, and **(3)** adaptive-prototype weighting. First, the dynamic-energy threshold enables fine-grained separation between in-distribution and OOD samples by adjusting thresholds on a per-class basis, preventing over-reliance on a global cutoff. Next, the OT-based pseudo-label refinement calibrates noisy pseudo-labels caused by distribution shifts through distribution-

level alignment between teacher and student predictions. Finally, the adaptive-prototype weighting emphasizes semantically aligned target samples while down-weighting highly uncertain ones, ensuring that only informative OOD samples contribute to adaptation. Our framework effectively utilizes semantically useful unknown samples rather than discarding them, leading to more robust and efficient online adaptation under open-set conditions. Further descriptions of the proposed modules are discussed below.

3.2. Dynamic-Energy Threshold for OOD Detection

OOD sample detection is a binary classification problem that provides a boundary between in- and out-of-distribution samples. However, previous methods [29, 53, 77] rely on softmax scores to distinguish between these samples. But models have the tendency to assign arbitrarily high confidence for OOD samples. Energy-based scores overcome this limitation [12, 47] by mapping each sample to a scalar that is lower for ID-data but higher for OOD data. Here, a density function of data $p(x)$ that has a low likelihood score can be considered as an OOD sample. Energy of a given input (x, y) is $E(x, y) = -f(x)$ where $f(x)$ is a discriminative neural classifier. Free energy function is defined as $E(\mathbf{x}; f) = -T \cdot \log \sum_i^K e^{f_i(\mathbf{x})/T}$, where T is the softmax temperature. Based on the energy free function, OOD detection is defined as:

$$g(\mathbf{x}; \tau, f) = \begin{cases} 1, & \text{if } -E(\mathbf{x}; f) > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where τ is the energy threshold and $g(x)$ is the OOD detector. The threshold is derived carefully from the in-distribution data for better OOD detection and separability [47]. However, a fixed global threshold can eliminate too many unlabeled samples which can potentially be useful OOD samples. Moreover, naively eliminating OOD samples can eventually negatively affect the model's performance, especially in class-imbalance settings. To address this limitation, we propose a *dynamic-energy threshold* that can dynamically adjust the threshold on a per-class basis. To achieve this, we apply a class-conditioned energy statistics (μ_c, σ_c), and define a class-conditioned threshold,

$$\tau_c = \mu_c + \beta\sigma_c \tag{2}$$

where β is a scaling hyperparameter. Given a test sample x with a pseudo label $\hat{y}$. We assign a soft weight based on its normalized energy:

$$w_c(\mathbf{x}) = \exp\left(-\gamma \cdot \frac{E(\mathbf{x}; f) - \mu_c}{\sigma_c + \varepsilon}\right) \tag{3}$$

where γ is a hyperparameter that controls the sharpness of down-weighting, and ε enforces numerical stability. This weighting strategy ensures low energy samples such as ID-samples obtain higher weights while OOD-samples are down-weighted rather than fully rejected by other works [24, 41]. Finally, our adaptation incorporates these dynamic weights into the unsupervised loss:

$$\mathcal{L}_{\text{dyn}}(\theta) = \frac{1}{\sum_{\mathbf{x}\in\mathcal{B}_t} w_{\hat{y}}(\mathbf{x})} \sum_{\mathbf{x}\in\mathcal{B}_t} w_{\hat{y}}(\mathbf{x}) \cdot \ell(\mathbf{x}; \theta) \tag{4}$$

Here, $\ell(\mathbf{x}; \theta)$ refers to the unsupervised loss during TTA. Our dynamic thresholding based on the energy score allows the model to retain useful OOD samples that are semantically aligned with source classes, while rejecting extreme outliers. As a result, it can adaptively balance between ID samples and extracting useful information from OOD data instead of just rejecting them during TTA.

3.3. Optimal Transport for Refined Pseudo Labels

In TTA, unreliable pseudo labels often occur due to dynamic distribution shifts in the test set. However, models often show high confidence on OOD samples, which that can be incorrect. Higher confidence leads to noise during adaptation that hinders the model's performance. Taking inspiration from OT due to its success in domain adaptation tasks [5, 14, 15, 62], we propose **O**ptimal **T**ransport based **P**seudo-**L**abel refinement **(OTPL)** strategy. We adopt a student-teacher framework [20, 70, 74], where the student model performs adaptation and teacher model provides stable prediction. Instead of sharing the weights with the student model, the teacher model updates its weights using *exponential moving average* weights of the student model.

$$\theta'_T = \alpha\theta'_{T-1} + (1 - \alpha)\theta_T \tag{5}$$

Here, α is a smoothing factor. Let $p_T(\mathbf{x})$ and $p_S(\mathbf{x})$ be the predictive distributions of a teacher and student model respectively. In a general setting, consistency regularization is enforced to ensure similarity between $p_T(\mathbf{x})$ and $p_S(\mathbf{x})$ on unlabeled target samples. We leverage OT to align the teacher and student distributions instead of minimizing their discrepancy for each sample. Given a cost function $C(p_T, p_S)$ that minimizes their disagreement, we compute OT coupling such as:

$$\min_{T\in\Pi(p_T,p_S)} \langle T, C\rangle + \varepsilon H(T) \tag{6}$$

where $T \in \mathbb{R}_+^{n\times m}$ is the transport plan, $\Pi(p_T, p_S)$ denotes the set of couplings with marginals p_T and p_S, and $H(T)$ is the entropy regularizer. This formulation is efficiently solved with the Sinkhorn algorithm [17]. Distribution-level alignment between teacher and student predictions is achieved by OT plan T. Thus, we define refined pseudo-labels as:

$$q(\mathbf{x}_i) = \sum_j T_{ij} \tag{7}$$

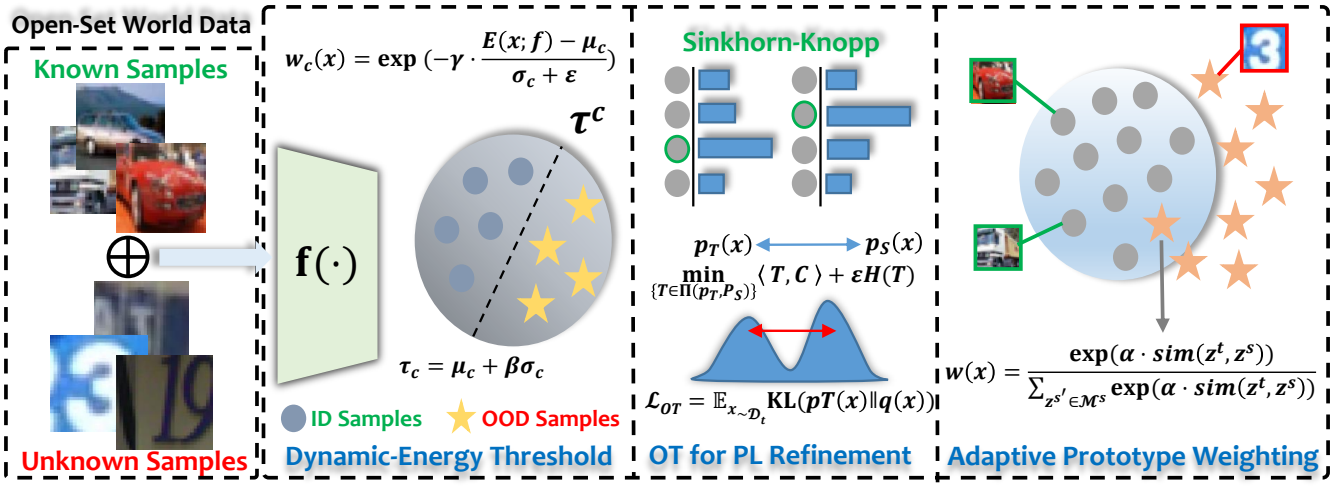

Figure 2. An overview of our *learning from unknown* OSTTA (LU-OSTTA) method. Real-world data can contain unknown samples during testing. Our framework deals with these unknown samples, and learn useful information from them to improve open-set adaptation performance. We identify known and unknown samples, then refine noisy pseudo labels. Lastly, we learn semantically useful unknown samples via a weighted contrastive objective.

where $q(\mathbf{x}_i)$ represents a smoothed distribution over classes that aligns both teacher and student predictions. These refined labels alleviate the risk of overconfidence in model predictions and ensure correct pseudo-labels. Furthermore, the student model is then optimized with KL divergence loss between its prediction and the OT refined pseudo-labels:

$$\mathcal{L}_{OT} = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_t} \mathrm{KL}\big(p_S(\mathbf{x}) \parallel q(\mathbf{x})\big) \tag{8}$$

This ensures the student model remains consistent with the teacher while having better and calibrated pseudo-labels. This OT-based pseudo-label refinement method jointly ensures calibration and consistency between pseudo-labels while preventing noisy labels that often occur during distribution shifts. It also ensures class-level structure in the target domain due to down-weighting unreliable predictions. It is worth mentioning that the OT-based refinement strategy complements our dynamic-energy thresholding because pseudo-labels are retained after prior class-conditioned thresholding, further calibrated to guide the adaptation process effectively.

3.4. Adaptive-Prototype Weighting

Although dynamic-energy thresholding and OT-based refinement encourage the use of better pseudo labels and OOD samples for adaptation, there are still target samples that are highly uncertain and far from the source classes. Naively rejecting them can diminish the potential of utilizing unknown samples. On the other hand, utilizing them directly for adaptation can cause the risk of negative transfer due to the mismatch between known and unknown classes [45]. To address this, we introduce an *adaptive-prototype weighting* mechanism that selectively emphasizes target samples based on their semantic proximity to source prototypes. We maintain source and target memory banks $\mathcal{M}^s$ and $\mathcal{M}^t$ to store prototypes respectively.

$$\begin{aligned} \mathcal{M}^s &= [m_1^s, m_2^s, \ldots, m_{N_s}^s], \\ \mathcal{M}^t &= [m_1^t, m_2^t, \ldots, m_{N_t}^t] \end{aligned} \tag{9}$$

Here, feature vector of x_i is stored in m_i, and updated after each batch with a momentum v. $\mathcal{M}^s$ does not store source data or features extracted during test time, hence acts as a set of frozen source prototypes.

$$m_i \leftarrow v m_i + (1 - v)\mathbf{f}_i \tag{10}$$

We perform k-means clustering to get clusters $C_s = [c_1^s, c_2^s, \ldots, c_{N_s}^s]$ and $C_t = [c_1^t, c_2^t, \ldots, c_{N_t}^t]$ on $\mathcal{M}^s$ and $\mathcal{M}^t$ [84]. We also retrieve normalized source and target prototypes $\{\mu_j^s\}_{j=1}^{N_s}$ and $\{\mu_j^t\}_{j=1}^{N_t}$. Using the source model, we extract features $\mathbf{f}_i$ and compute similarity distribution between $\mathbf{f}_i$ and $\{\mu_j^t\}_{j=1}^n$ (which is similar for $\{\mu_j^s\}_{j=1}^n$ too):

$$\mathcal{P}_{i,j}^s = \frac{\exp(\mu_j^t \cdot \mathbf{f}_i / \tau)}{\sum j = 1^k \exp(\mu_j^t \cdot \mathbf{f}_i / \tau)} \tag{11}$$

where τ is temperature parameter. This distribution captures how well a sample aligns with different target clusters. As target domain samples are unlabeled and far from the source distribution due to shift, it is hard to determine which target samples contain semantically useful information among OOD. To address this, we compute a semantic weight $w(x)$ for each target OOD sample $\mathbf{x}$ based on its similarity with source prototypes:

$$w(\mathbf{x}) = \frac{\exp(\alpha \cdot \mathrm{sim}(z^t, z^s))}{\sum_{z^{s'} \in \mathcal{M}^s} \exp(\alpha \cdot \mathrm{sim}(z^t, z^{s'}))}, \tag{12}$$

Table 1. Results of different methods on CIFAR-10/100-C benchmarks. (↑) indicates that larger values are better, and vice versa. All values are percentages. We present Source, TTA methods, OSTTA methods, and Our method respectively. We underline the second best score, and best scores are in **bold**. And improvements (±) compared to the second best score are presented.

Method	CIFAR-10-C				CIFAR-100-C			
	Acc.↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc.↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [85]	80.73	76.44	79.25	67.44	53.68	61.55	93.20	39.71
TENT [72]	78.35	64.79	94.87	55.92	55.71	66.23	93.89	41.34
CoTTA [74]	83.34	84.07	72.40	77.45	55.33	77.23	80.78	48.32
EATA [56]	81.45	83.76	71.23	72.56	60.78	87.24	94.88	42.28
OSTTA [38]	84.12	72.45	76.54	65.32	60.18	75.34	82.12	50.83
UniEnt [24]	84.23	95.45	13.25	83.45	59.45	92.21	23.24	57.31
UniEnt+ [24]	85.18	96.12	13.10	84.27	60.32	92.43	23.69	58.23
S-OSTTA [37]	86.23	96.74	11.24	85.03	61.34	95.37	23.75	58.35
Ours	**89.34**$_{(+3.11)}$	**98.56**$_{(+1.82)}$	**9.43**$_{(-1.81)}$	**88.74**$_{(+3.71)}$	**64.76**$_{(+3.42)}$	**97.67**$_{(+2.30)}$	**20.34**$_{(-3.41)}$	**62.45**$_{(+4.10)}$

where z^t and z^s are the target and source embeddings. And α is a scaling factor. This weighting enforces target samples that are close to the source prototype as candidates during adaptation. To align target samples with the source prototype, we enforce contrastive loss with $w(\mathbf{x})$:

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{B}_t|}\sum_{\mathbf{x}\in\mathcal{B}_t} w(\mathbf{x})\cdot\log\frac{\exp\left(\text{sim}(z^t, z^s)/\tau\right)}{\sum_{z^{s'}\in\mathcal{M}^s}\exp\left(\text{sim}(z^t, z^{s'})/\tau\right)}. \tag{13}$$

This ensures useful OOD samples contribute during adaptation while highly uncertain samples are down-weighted. Finally, to prevent mode collapse and ensure balanced predictions across classes, we add an entropy regularizer:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{|\mathcal{B}_t|}\sum_{\mathbf{x}\in\mathcal{B}_t}\sum_{c=1}^{K} p(c|\mathbf{x})\log p(c|\mathbf{x}), \tag{14}$$

Here, $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{ent}}$ ensure that semantically useful OOD samples guide the adaptation process while avoiding mode collapse and maintaining a diverse decision boundary.

3.5. Training Objective

During test-time adaptation, the student model f_θ is optimized on unlabeled target samples $\mathcal{D}_t$ using a combination of dynamic-energy loss $\mathcal{L}_{\text{dyn}}$, OT-based pseudo-label refinement loss $\mathcal{L}_{OT}$, $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{ent}}$. For a target batch $\mathcal{B}_t$, the total training objective can be written as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dyn}} + \lambda_{OT}\mathcal{L}_{OT} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} \tag{15}$$

where λ balances the contribution of all losses. Minimizing $\mathcal{L}_{\text{total}}$ encourages the student model to adapt robustly to the target domain while leveraging potentially useful unknown samples instead of discarding them.

4. Experiments

4.1. Experimental Setup

Datasets. We follow previous TTA studies; we evaluate our method with commonly used corruption benchmark datasets: CIFAR-10-C, CIFAR-100-C, and Tiny-ImageNet-C [28]. Each of the datasets has 15 different corruptions with 5 severity levels from 1 to 5, and we use the highest level. Models pre-trained with clean source training data (CIFAR-10, CIFAR-100, and Tiny-ImageNet) are adapted to corrupted datasets. Following [24, 38], we apply the identical corruption to SVHN [52] and ImageNet-O [31] datasets, and create SVHN-C and ImageNet-O-C datasets to incentivize the open-set setting with unknown classes. All open-set samples were resized as the closed-set samples and to the same corruption level.

Evaluation Settings. We follow the common TTA protocol where the model is evaluated under continuously changing domains without resetting its parameters after adapting to each domain. During inference, corrupted target samples are streamed in an online manner. For each mini-batch, the model first predicts the labels and updates its parameters using the same mini-batch. To construct the mini-batches, we sample an equal number of known (in-distribution) and unknown (out-of-distribution) samples to simulate the open-set scenario. We adopt evaluation setup from [24]. We report the classification accuracy on known target samples to measure the adaptation performance. Second, to evaluate the separation of known and unknown samples, we report the area under the ROC curve (AUROC), false positive rate at 95% TPR (FPR@95) and the Open-Set Classification Rate (OSCR), which jointly measure classification and rejection performance across decision thresholds.

Implementation Details. We follow the previous standard TTA [56, 72] and OSTTA [24, 37] setups; we use WideResNet-40 for CIFAR benchmarks, pre-trained on a clean dataset using AugMix augmentation provided by RobustBench [16]. For Tiny-ImageNet-C, the ResNet-50 model is pre-trained on Tiny-ImageNet. During TTA, the Adam optimizer with a learning rate of 0.001 and a batch size of 200 is used for all experiments. Particularly, learning rate is set to 0.01 for TENT [72], EATA [56], CoTTA [74] and OSTTA [38].

Table 2. Results of different methods on Tiny-ImageNet benchmark. (↑) indicates that larger values are better, and vice versa. All values are percentages. Notations are followed from Tab. 1.

Method	Tiny-ImageNet-C			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [85]	22.14	54.86	93.32	16.12
TENT [72]	28.56	48.10	95.32	19.41
CoTTA [74]	30.12	57.24	81.78	23.45
EATA [56]	32.92	58.31	83.66	25.64
OSTTA [38]	37.46	55.78	93.78	27.67
UniEnt [24]	37.23	56.02	91.25	27.98
UniEnt+ [24]	38.00	57.12	90.34	28.43
S-OSTTA [37]	39.12	58.90	89.36	29.89
Ours	**42.78**$_{(+3.66)}$	**62.34**$_{(+3.44)}$	**86.57**$_{(-2.79)}$	**32.47**$_{(+2.58)}$

4.2. Experimental Results

CIFAR Benchmarks. We conduct experiments on both CIFAR-10-C and CIFAR-100-C benchmarks, where the SVHN-C dataset is included an open-set unknown dataset for evaluation. We compare our method with source-only, representative TTA methods, and OSTTA methods. As presented in Tab. 1, our method significantly improves performance across all metrics (Acc., AUROC, FPR@TPR95, and OSCR) in both closed-set and open-set TTA. This improvement is due to the inclusion of semantically useful unknown samples during adaptation. However, directly using the pre-trained source [85] model results in a significant drop in accuracy and a 69.82% higher FPR@TPR95 compared to our method, as it is unable to distinguish between known and unknown samples in the CIFAR-10-C dataset. For example, our method improves TENT [72] by 10.99%, 33.77%, 85.44%, and 32.82% across the evaluation metrics, while it improves the best-performing open-set TTA method S-OSTTA [37] by 3.11%, 0.82%, 1.81%, and 3.71%, respectively. On the CIFAR-100-C dataset, our method also consistently exhibits better performance. In contrast, TENT [72] and OSTTA [38] show performance drops in the OSCR metric, as these models naively update parameters with unknown samples included, failing to achieve considerable performance. Finally, our consistent improvements suggest the effectiveness of selectively incorporating semantically useful unknown samples, rather than naively adopting them as in TENT [72] or OSTTA [38].

Tiny-ImageNet Benchmark. We also conduct experiments on the Tiny-ImageNet benchmark, as presented in Tab. 2. Our method consistently outperforms all other approaches (Source, TTA, and OSTTA) across all evaluation metrics. In particular, it achieves performance gains of 3.66%, 3.44%, 2.79%, and 2.58% over the second-best method, S-OSTTA [37]. These improvements demonstrate that our proposed modules: DET, OT-PL Refinement, and APW play a crucial role in improving model performance by effectively leveraging helpful unknown samples for adaptation.

4.3. Further Analysis

Ablation Study. We verify the effect and contribution of each proposed component (DET, OTPL, APW) in Tab. 3. We use CIFAR-10-C as the closed-set and SVHN-C as the open-set datasets. Without known and unknown sample separation, the model intends to minimize the entropy of open-set samples, which leads to a severe performance drop in all metrics. Adding our dynamic-energy thresholding (DET) alone drastically improves performance due to its flexible class-conditioned known and unknown sample separation strategy, which indicates our proposed DET can well distinguish samples. Moreover, our OT-based pseudo-label refinement also proves effective as it improves all metrics by a large margin similar to DET. Combining both DET and OTPL achieves similar results to other OSTTA methods such as UniEnt and UniEnt+. Finally, our adaptive prototype weighting further improves the overall results and outperforms all methods. This indicates that our APW incentivizes learning useful unknown samples, which helps achieve better performance and robustness.

Table 3. Ablation study of DET, OTPL and APW.

DET	OTPL	APW	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
✗	✗	✗	64.23	61.33	96.31	48.42
✓	✗	✗	82.27	79.93	39.45	68.73
✗	✓	✗	74.34	76.14	42.59	66.56
✓	✓	✗	83.45	95.61	14.56	83.11
✓	✓	✓	**89.34**	**98.56**	**9.43**	**88.74**

Different Confidence Thresholds. Inspired by the experimental setting in [38], we evaluate our dynamic-energy threshold (DET) against other thresholding methods such as MSP [29], Max Logit [27], and Energy [47]. For the evaluation, we train TENT [72] and compare our DET with these methods. We include the SVHN-C dataset alongside CIFAR-10-C to ensure an open-set setting. From Tab. 4, we can see that across all evaluation metrics—AUROC, FPR@TPR95, and OSCR; our method shows superior performance. Another advantage is that using class-conditioned energy statistics helps to immediately distinguish between known and unknown samples, whereas existing OOD methods require a large batch to determine the threshold with the AUROC score. Hence, our method is more flexible and better suited for OSTTA tasks.

Table 4. Comparison of different methods on CIFAR-10-C. (↑) indicates higher is better, (↓) indicates lower is better.

Method	CIFAR-10		
	AUROC↑	FPR@TPR95↓	OSCR↑
MSP [29]	48.08	93.34	42.90
Max Logit [27]	50.51	90.67	45.66
Energy [47]	51.25	90.89	47.76
DET (Ours)	**79.93**	**42.59**	**68.73**

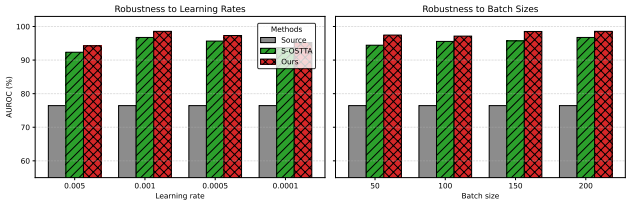

Figure 3. AUROC on CIFAR-10-C with different learning rate and batch sizes. Source [85] and S-OSTTA [37] are compared here.

Different Number of Unknown Classes. We measure the performance under different numbers of unknown classes, as it is a crucial criterion to understand how the model is deviating when the number of unknown classes increases. In Tab. 5, we can see that TENT [72] and other methods have consistent fluctuations when the number of classes increases, hence complexity. However, our proposed method has shown more robust and stabilized performance when the number of unknown classes increases.

Table 5. OSCR of different methods including ours on CIFAR-10-C under different number of unknown classes.

Method	2	4	6	8	10
Source	72.45	69.67	69.33	69.17	68.05
TENT	49.44	49.23	50.56	48.25	50.56
UniEnt+	78.45	78.46	78.51	78.34	76.98
S-OSTTA	80.34	79.26	79.76	80.00	78.28
Ours	**84.12**	**83.44**	**83.33**	**83.67**	**83.00**

Model Robustness to Hyper-Parameters. In real-world applications, models are often deployed on small devices. However, it is crucial to have a small batch size and set an optimal learning rate to stabilize performance. In Fig. 3, we can see our method can flexibly outperform S-OSTTA [37] even under a batch size of 50. On the other hand, the results do not fluctuate even across low to high learning rates. These findings demonstrate the scalability and robustness of our method when hyper-parameters are not optimal.

Different Ratios of Known and Unknown. We follow the setup from [24, 37, 38] to experiment with different ratios of known and unknown samples. But in the main experiments, it is set to 1:1. We consider an imbalanced data ratio between known and unknown samples, where the open-set dataset has a lower (0.1/0.5) ratio, or higher data ratio (1.5/2.0). In Tab. 6, we observe that the TTA method such as TENT [72] has a significant drop when we add a higher number of open-set data. In contrast, our method exhibits consistency and less sensitivity to different data ratios; hence it is more suitable in real-world applications where the data ratio is often imbalanced.

Per-batch Inference Time. TTA methods are often deployed in small devices, and they are often resource-intensive. In Tab. 7, we compare the inference latency per batch. As we can see our method achieves better OSCR than other high-performing methods while preserving similar inference time.

Table 6. OSCR of different ratios of known and unknown on CIFAR-10-C.

Method	0.1	0.5	1.5	2.0
Source	40.00	40.30	38.97	38.18
TENT	47.76	44.27	44.89	41.79
UniEnt+	56.43	57.85	56.12	55.04
S-OSTTA	60.23	60.43	59.90	58.67
Ours	**63.56**	**63.30**	**62.92**	**62.76**

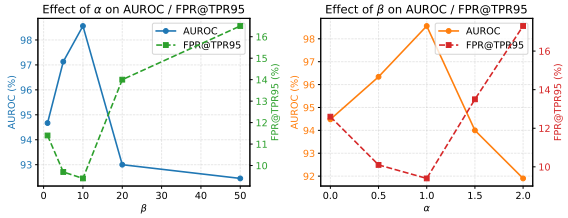

Figure 4. AUROC (%) and FPR@TPR95 on CIFAR-10-C with different α and β values.

Ablation on α and β. We perform analysis for β and α values from eq. 2 and 12. In Fig. 4, we present the performance across different β and α values. We observe that β = 1.0 maintains the best performance, whereas α = 10 complements the performance on CIFAR-10-C. A larger β value can assume more target samples as in-distribution known samples, and vice versa. On the other hand, smaller α value is better when the target prototypes are noisier and classes are not well separated.

Table 7. Comparison of different methods for per-batch inference latency [37] on CIFAR-10-C.

Method	CIFAR-10-C						
	TENT	EATA	CoTTA	OSTTA	UniEnt	S-OSTTA	**Ours**
Inference Time (ms)↓	26.34	57.76	887.34	59.89	75.34	90.34	93.34
OSCR↑	55.92	72.56	77.45	65.32	83.45	85.03	**88.74**

5. Conclusion

To tackle the challenging paradigm of open-set test-time adaptation (OSTTA) effectively, we propose LU-OSTTA, which leverages both known and unknown samples during adaptation, unlike previous methods that overlook unknown samples. We introduce a dynamic-energy threshold to separate known and unknown samples in a class-conditioned manner. To mitigate the effect of noisy pseudo-labels, we employ an optimal transport-based pseudo-label refinement. We further propose an adaptive prototype weighting strategy to ensure effective learning from semantically useful unknown samples. Experiments across different settings confirm that our method consistently outperforms existing TTA and OSTTA approaches in the open-set settings.

Acknowledgement

This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(**RS-2025-25422680**, Metacognitive AGI Framework and its Applications, and **RS-2020-II201373**, Artificial Intelligence Graduate School Program (Hanyang University)).

References

[1] Amit Agarwal and Kulbhushan Pachauri. Pseudo labelling for key-value extraction from documents, 2023. US Patent 11,823,478. 2

[2] Amit Agarwal, Srikant Panda, Angeline Charles, Hitesh Laxmichand Patel, Bhargava Kumar, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, and Dong-Kyu Chae. MVTamperBench: Evaluating robustness of vision-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18804–18828, Vienna, Austria, July 2025. Association for Computational Linguistics. 3

[3] Amit Agarwal, Srikant Panda, and Kulbhushan Pachauri. FS-DAG: Few shot domain adapting graph networks for visually rich document understanding. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal, editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 100–114, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics. 3

[4] Fatemeh Azimi, Sebastian Palacio, Federico Raue, Jörn Hees, Luca Bertinetto, and Andreas Dengel. Self-supervised test-time adaptation on video data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3439–3448, 2022. 3

[5] Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020. 4

[6] Alexander Bartler, Andre Bühler, Felix Wiewel, Mario Döbler, and Bin Yang. Mt3: Meta test-time training for self-supervised test-time adaption. In *International Conference on Artificial Intelligence and Statistics*, pages 3080–3090. PMLR, 2022. 1

[7] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022. 1

[8] Shilei Cao, Juepeng Zheng, Yan Liu, Baoquan Zhao, Ziqi Yuan, Weijia Li, Runmin Dong, and Haohuan Fu. Exploring test-time adaptation for object detection in continually changing environments. *arXiv preprint arXiv:2406.16439*, 2024. 3

[9] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 1

[10] MingCai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In *Forty-second International Conference on Machine Learning*. 1

[11] Wenxi Chen, Raymond A Yeh, Shaoshuai Mou, and Yan Gu. Leveraging perturbation robustness to enhance out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4724–4733, 2025. 3

[12] Hyunjun Choi, Hawook Jeong, and Jin Young Choi. Balanced energy regularization loss for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15691–15700, 2023. 4

[13] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 13

[14] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017. 4

[15] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016. 4

[16] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020. 6

[17] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 4

[18] Qi Deng, Shuaicheng Niu, Ronghao Zhang, Yaofo Chen, Runhao Zeng, Jian Chen, and Xiping Hu. Learning to generate gradients for test-time adaptation via test-time training layers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16235–16243, 2025. 2

[19] Zeshuai Deng, Zhuokun Chen, Shuaicheng Niu, Thomas Li, Bohan Zhuang, and Mingkui Tan. Efficient test-time adaptation for super-resolution with second-order degradation and reconstruction. *Advances in Neural Information Processing Systems*, 36:74671–74701, 2023. 3

[20] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023. 2, 4

[21] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *Advances in Neural Information Processing Systems*, 35:37199–37213, 2022. 2

[22] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances*

in neural information processing systems, 34:7068–7081, 2021. 2

[23] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022. 1

[24] Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23975–23984, 2024. 1, 2, 4, 6, 7, 8, 13, 14, 15

[25] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[26] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019. 3

[27] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 7

[28] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 6

[29] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 4, 7

[30] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 3

[31] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 6

[32] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960, 2020. 3

[33] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 3

[34] JoonHo Jang, Byeonghu Na, Dong Hyeok Shin, Mingi Ji, Kyungwoo Song, and Il-Chul Moon. Unknown-aware domain adversarial learning for open-set domain adaptation. *Advances in Neural Information Processing Systems*, 35:16755–16767, 2022. 3

[35] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022. 3

[36] Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2022. 3

[37] Byung-Joon Lee, Jin-Seop Lee, and Jee-Hyong Lee. Stabilizing open-set test-time adaptation via primary-auxiliary filtering and knowledge-integrated prediction. *arXiv preprint arXiv:2508.18751*, 2025. 6, 7, 8, 13, 14, 15

[38] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16380–16389, 2023. 1, 2, 6, 7, 8, 13, 14, 15

[39] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*, 2024. 2

[40] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 3

[41] Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11836–11846, 2023. 4

[42] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025. 1

[43] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 3

[44] Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22952–22961, 2023. 3

[45] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2927–2936, 2019. 3, 5

[46] Jin Liu, Jialong Xie, Fengyu Zhou, and Shengfeng He. Question type-aware debiasing for test-time visual question answering model adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10805–10816, 2024. 3

[47] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 2, 4, 7

[48] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23946–23955, 2023. 3

[49] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or

thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021. 1

[50] Jing Ma. Improved self-training for test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23701–23710, 2024. 2

[51] Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2555–2565, 2024. 2

[52] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, 2011. 6

[53] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 2, 4

[54] A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip HS Torr. Tipi: Test time adaptation with transformation invariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24162–24171, 2023. 1

[55] Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. *arXiv preprint arXiv:2404.01650*, 2024. 2

[56] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 1, 2, 6, 7, 13, 14

[57] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 2

[58] Srikant Panda, Amit Agarwal, Gouttham Nambirajan, and Kulbhushan Pachauri. Out of distribution element detection for information extraction, 2025. US Patent App. 18/347,983. 2

[59] Hyejin Park, Jeongyeon Hwang, Sunung Mun, Sangdon Park, and Jungseul Ok. Medbn: Robust test-time adaptation against malicious test samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5997–6007, 2024. 2

[60] Taki Hasan Rafi, Amit Agarwal, Hitesh L. Patel, and Dong-Kyu Chae. Towards robust continual test-time adaptation via neighbor filtration. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5161–5165. ACM, 2025. 1, 13

[61] Taki Hasan Rafi, Serbeter Karlo, Amit Agarwal, Hitesh Patel, Bhargava Kumar, and Dong-Kyu Chae. Instance-aware test-time adaptation for domain generalization. In *Proceedings of the 30th International Conference on Database Systems for Advanced Applications (DASFAA)*, Singapore City, Singapore, 2025. to appear. 1

[62] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 737–753. Springer, 2017. 4

[63] Xiaoqian Ruan and Wei Tang. Fully test-time adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1038–1047, 2024. 3

[64] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 9000–9009, 2021. 3

[65] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 153–168, 2018. 3

[66] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021. 3

[67] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pages 20827–20840. PMLR, 2022. 3

[68] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 1

[69] Mingkui Tan, Guohao Chen, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Peilin Zhao, and Shuaicheng Niu. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[70] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 4

[71] Edwin Thomas, Amit Agarwal, Sandeep Jana, and Kulbhushan Pachauri. Model augmentation framework for domain assisted continual learning in deep learning, 2025. US Patent App. 18/406,905. 3

[72] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2, 6, 7, 8, 13, 14, 15

[73] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022. 3

[74] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 1, 4, 6, 7, 13, 14, 15

[75] Qian Wang, Fanlin Meng, and Toby P Breckon. Progressively select and reject pseudolabeled samples for open-set domain adaptation. *IEEE Transactions on Artificial Intelligence*, 5(9):4403–4414, 2024. 3

[76] Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu. Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia*, 26:2137–2147, 2023. 3

[77] Guoxuan Xia and Christos-Savvas Bouganis. Augmenting softmax information for selective classification with out-of-distribution data. In *Proceedings of the Asian Conference on Computer Vision*, pages 1995–2012, 2022. 4

[78] Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Modality-collaborative test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26732–26741, 2024. 3

[79] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 3

[80] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. 3

[81] Puning Yang, Jian Liang, Jie Cao, and Ran He. Auto: Adaptive outlier optimization for online test-time ood detection. *arXiv preprint arXiv:2303.12267*, 2023. 3

[82] Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. Oodd: Test-time out-of-distribution detection with dynamic dictionary. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30630–30639, 2025. 3

[83] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023. 1, 2

[84] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13834–13844, 2021. 5

[85] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 1, 6, 7, 8, 14

[86] Runhao Zeng, Qi Deng, Huixuan Xu, Shuaicheng Niu, and Jian Chen. Exploring motion cues for video test-time adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1840–1850, 2023. 3

[87] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5531–5540, 2023. 3

[88] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18971–18981, 2023. 1

[89] Lei Zhang, Jiangtao Nie, Wei Wei, and Yanning Zhang. Unsupervised test-time adaptation learning for effective hyperspectral image super-resolution with unknown degeneration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5008–5025, 2024. 3

[90] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. *arXiv preprint arXiv:2306.03536*, 2023. 1

[91] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 13

[92] Zhi Zhou, Lan-Zhe Guo, Lin-Han Jia, Dingchu Zhang, and Yu-Feng Li. Ods: Test-time adaptation in the presence of open-world data shift. In *International Conference on Machine Learning*, pages 42574–42588. PMLR, 2023. 3

A. Evaluation Baselines

We mainly focused on comparing our proposed method with three types of other methods. (1) Entropy-free method: **The** source model trained with clean datasets is directly tested under an open-set setting. (2) Entropy-based/continual TTA methods: **TENT** [72] estimates the normalization statistics and optimizes the model parameters based on entropy minimization. **CoTTA** [74] adopts the mean-teacher method to improve pseudo labels and provide a weighted average of these labels to mitigate error accumulation. It also introduces a stochastic restoration module to enforce continual adaptation by avoiding catastrophic forgetting. **EATA** [56] reduces the effect of noisy samples with high entropy by employing an active sample selection criterion. To alleviate the issue of forgetting, they introduce a Fisher regularizer to constrain model parameters. (3) Open-set TTA methods: **OSTTA** [38] uses a filtering technique based on the confidence values of the adopted model compared with the original source model, where low confidence samples appear to be noisy. **UniEnt** [24] uses entropy minimization and maximization with a distribution-aware filtering method for both covariate shifted in-and out-of-distribution samples. Furthermore, **UniEnt+** [24] alleviates the noisy samples by leveraging sample-level confidence. Lastly, **Stabilized OSTTA** [37] uses an auxiliary filtering method to validate data from the primary filtering mechanism and also employs knowledge-integrated prediction to calibrate the output of the adopted model.

B. Pseudo Code

Algorithm 1: LU-OSTTA

Input: Pre-trained source model f_θ, target batch $\mathcal{B}_t$, source prototypes $\mathcal{M}^s$
Output: Adapted student model f_θ
for *each target batch* $\mathcal{B}_t$ **do**
  `// Step 1: Dynamic-Energy OOD Detection`
  for *each sample* $\mathbf{x} \in \mathcal{B}_t$ **do**
    Compute energy: $E(\mathbf{x}; f)$
    Derive class-conditioned threshold: $\tau_c = \mu_c + \beta\sigma_c$
    Compute dynamic weight: $w_c(\mathbf{x}) = \exp\left(-\gamma \cdot \frac{E(\mathbf{x};f)-\mu_c}{\sigma_c+\varepsilon}\right)$
  Compute weighted loss: $\mathcal{L}_{\text{dyn}}(\theta)$ from Eq. (7)
  `// Step 2: OT-based Pseudo-Label Refinement`
  Obtain teacher distribution $p_T(\mathbf{x})$ and student $p_S(\mathbf{x})$
  Compute OT plan T with cost $C(p_T, p_S)$ via Sinkhorn
  Refine pseudo-label: $q(\mathbf{x}_i) = \sum_j T_{ij}$
  Compute loss: $\mathcal{L}_{OT} = \mathbb{E}\ \text{KL}(p_S(\mathbf{x}) \| q(\mathbf{x}))$
  `// Step 3: Adaptive-Prototype Weighting`
  Update target memory bank $\mathcal{M}^t$ with momentum
  Cluster $\mathcal{M}^s, \mathcal{M}^t$ into prototypes $\{\mu_j^s\}, \{\mu_j^t\}$
  For each target embedding z^t:
    Compute semantic weight: $w(\mathbf{x}) = \frac{\exp(\alpha \cdot \text{sim}(z^t, z^s))}{\sum_{z^{s'} \in \mathcal{M}^s} \exp(\alpha \cdot \text{sim}(z^t, z^{s'}))}$
    Compute contrastive loss: $\mathcal{L}_{\text{con}}$ from Eq. (15)
    Compute entropy regularizer: $\mathcal{L}_{\text{ent}}$ from Eq. (16)
  `// Step 4: Update Model`
  Compute total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dyn}} + \lambda_{OT}\mathcal{L}_{OT} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}}$$

  Update model parameters $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}$

C. More Results

Additional Results on CIFAR Benchmarks. We perform additional experiments with Places365-C [91] and Texture-C [13] datasets. We follow the same setup and evaluation metric from [37], we further add harmonic mean (H-S) of accuracy and AUROC. In Tab. 8, we demonstrate the performance with CIFAR-10-C by adding different open-set environments. our method consistently outperforms S-OSTTA method in all metrics. Existing open-set TTA methods exhibit considerable performance, but lack achieving higher performance compared to our method. But S-OSTTA performed closely with our method, but the margin is significant. On the other hand, Tab. 9, we follow the similar trend as our method outperforms other methods with CIFAR-100-C benchmark as well. Similarly, S-OSTTA achieved the second best score in all metrics.

Additional Results on Tiny-Imagenet Benchmark. In Tab. 10, we perform experiment with Places365-C [91] and Texture-C [13] datasets as open-set environment and Tiny-ImageNet-C as the close-set environment. Tny-Imagenet-C has poses more challenging tasks as it has 200 classes. In both open-set datasets, our method significantly performs other methods, and demonstrates its capability to handle open-set environments.

Performance under Continual Settings. We follow a similar setup [60,74]. In standard TTA setting, corruption types change abruptly in the highest severity level (e.g. 1-5), where 1 is the lowest and 5 is the highest. However, in

Table 8. Results of different methods on CIFAR-10-C benchmark. ↑ indicates that larger values are better. All values are percentages. We present Source , TTA methods , OSTTA methods , and Our method respectively. We underline the second best score, and best scores are in **bold**. Improvements (±) compared to the second best score are also presented.

Method	Places365-C			Textures-C		
	Acc↑	AUROC↑	H-S↑	Acc↑	AUROC↑	H-S↑
Source [85]	82.46	83.32	82.89	82.46	82.51	82.48
TENT [72]	55.31	54.23	54.76	70.23	68.45	69.33
CoTTA [74]	84.67	82.34	83.49	84.10	79.78	81.88
EATA [56]	84.78	80.35	82.51	81.87	78.24	80.01
OSTTA [38]	84.56	76.45	80.30	81.56	69.43	75.01
UniEnt [24]	84.78	88.64	86.67	82.75	84.43	83.58
UniEnt+ [24]	84.56	89.57	86.99	80.45	89.65	84.80
S-OSTTA [37]	88.23	94.12	91.08	87.56	97.51	92.27
Ours	**90.41**$_{(+1.85)}$	**94.78**$_{(+0.66)}$	**92.54**$_{(+1.46)}$	**90.43**$_{(+2.87)}$	**98.74**$_{(+1.23)}$	**94.40**$_{(+2.13)}$

Table 9. Results of different methods on CIFAR-100-C benchmark. ↑ indicates that larger values are better. All values are percentages.

Method	Places365-C			Textures-C		
	Acc↑	AUROC↑	H-S↑	Acc↑	AUROC↑	H-S↑
Source [85]	53.45	65.34	58.92	53.45	62.65	57.55
TENT [72]	26.45	60.10	36.86	29.80	61.56	40.49
CoTTA [74]	55.77	72.81	63.51	51.45	67.68	58.47
EATA [56]	53.90	71.35	61.47	50.45	58.29	53.28
OSTTA [38]	60.32	72.65	66.21	58.76	65.30	61.84
UniEnt [24]	59.39	77.19	67.33	57.78	73.43	64.64
UniEnt+ [24]	58.76	78.67	67.31	56.45	73.89	64.42
S-OSTTA [37]	62.78	85.41	72.07	62.10	93.23	75.17
Ours	**64.32**$_{(+1.54)}$	**88.67**$_{(+3.26)}$	**74.93**$_{(+2.86)}$	**65.23**$_{(+3.13)}$	**95.78**$_{(+2.55)}$	**78.03**$_{(+2.86)}$

Table 10. Results of different methods on Tiny-ImageNet benchmark. ↑ indicates that larger values are better. All values are percentages.

Method	Places365-C			Textures-C		
	Acc↑	AUROC↑	H-S↑	Acc↑	AUROC↑	H-S↑
Source [85]	28.24	67.69	40.05	28.24	71.67	40.49
TENT [72]	40.78	65.76	50.38	34.87	46.67	39.87
CoTTA [74]	41.56	72.45	53.19	56.46	72.45	63.49
EATA [56]	44.32	77.34	56.23	42.87	65.23	51.67
OSTTA [38]	47.67	75.24	58.37	45.72	60.23	51.34
UniEnt [24]	46.87	78.25	58.57	44.45	64.72	51.96
UniEnt+ [24]	45.23	78.13	57.44	44.32	63.52	51.67
S-OSTTA [37]	48.24	84.08	61.23	47.89	82.80	60.45
Ours	**50.76**$_{(+2.52)}$	**86.87**$_{(+2.79)}$	**64.04**$_{(+2.81)}$	**49.90**$_{(+2.01)}$	**85.40**$_{(+2.60)}$	**63.84**$_{(+3.39)}$

continual setting, we experiment this severity level under a sequence by gradually changing severity for the 15 different corruption types. And the we change the corruption types gradually from lowest to highest, so that the distribution shift within each corruption is also gradual. Following previous method [74], we randomly shuffle 10 different corruption types then report average error rate over ten different sequences, shown in Tab. 11. We can see, our method outperforms both TTA and OSTTA settings by a significant margin in CIFAR-10-C dataset.

Table 11. Experiments on CIFAR-10-to-CIFAR-10-C by gradually changing. The severity level changes from lowest to highest and the corruption type changes when the severity level is lowest. Results are presented in mean over 10 randomly shuffled corruption types. Lower is better.

Avg. Error (%)↓	**Source**	**TENT [72]**	**CoTTA [74]**	**OSTTA [38]**	**UniEnt [24]**	**UniEnt+ [24]**	**S-OSTTA [37]**	**Ours**
CIFAR-10-C	26.5	33.6	12.2	24.6	11.3	11.2	9.4	**8.1**