Unified Entropy Optimization for Open-Set Test-Time Adaptation

Zhengqing Gao[1,2] Xu-Yao Zhang[1,2*] Cheng-Lin Liu[1,2]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
gaozhengqing2021@ia.ac.cn {xyz, liucl}@nlpr.ia.ac.cn

Abstract

Test-time adaptation (TTA) aims at adapting a model pre-trained on the labeled source domain to the unlabeled target domain. Existing methods usually focus on improving TTA performance under covariate shifts, while neglecting semantic shifts. In this paper, we delve into a realistic open-set TTA setting where the target domain may contain samples from unknown classes. Many state-of-the-art closed-set TTA methods perform poorly when applied to open-set scenarios, which can be attributed to the inaccurate estimation of data distribution and model confidence. To address these issues, we propose a simple but effective framework called unified entropy optimization (UniEnt), which is capable of simultaneously adapting to covariate-shifted in-distribution (csID) data and detecting covariate-shifted out-of-distribution (csOOD) data. Specifically, UniEnt first mines pseudo-csID and pseudo-csOOD samples from test data, followed by entropy minimization on the pseudo-csID data and entropy maximization on the pseudo-csOOD data. Furthermore, we introduce UniEnt+ to alleviate the noise caused by hard data partition leveraging sample-level confidence. Extensive experiments on CIFAR benchmarks and Tiny-ImageNet-C show the superiority of our framework. The code is available at https://github.com/gaozhengqing/UniEnt.

1. Introduction

Deep neural networks (DNNs) have achieved great success in recent years when the training and test data are drawn i.i.d. from the same distribution. However, in many real-world applications, this strict assumption is difficult to hold. Models deployed in practice can encounter different types of distribution shifts. On the one hand, the model needs to be able to address semantic shifts, *i.e.*, identify samples from unknown classes, which has given rise to problems such as out-of-distribution (OOD) detection [15, 16, 19, 32, 56] and open-set recognition [4, 5, 21, 43]. On the other hand, the model needs to be robust to covariate shifts and have good generalization performance to different styles and domains. Many efforts have been devoted to reduce the performance gap of DNNs under covariate shifts, such as domain generalization [45, 47, 58, 59] and domain adaptation [11, 50]. Among various studies addressing covariate shifts, test-time adaptation (TTA) has recently received increasing attention because its practicality: neither source domain data nor target domain labels are required [27, 28, 33, 35, 44, 46].

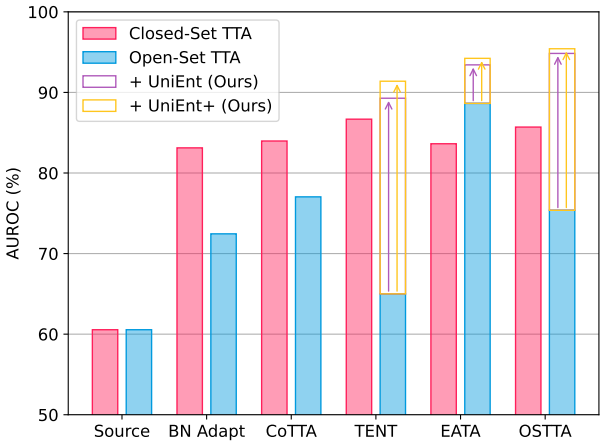


Figure 1. Existing TTA methods exhibit performance degradation with unknown classes included, while our methods can improve them significantly. We compare BN Adapt [33], CoTTA [46], TENT [44], EATA [35], and OSTTA [27].

Nevertheless, most of the existing TTA methods [33, 35, 44, 46] focus only on solving the covariate shift and ignoring the semantic shift. We believe that this is impractical since we cannot guarantee the test samples contain only the classes seen in the training phase. Many recent works [27, 28] have realized this and made some initial attempts. Figure 2 illustrates the differences between the traditional closed-set TTA and the novel open-set TTA set-

*Corresponding author.

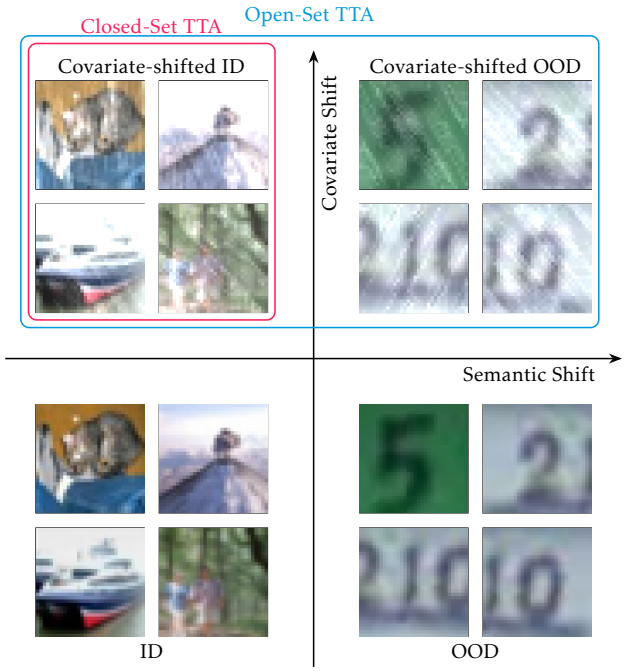


Figure 2. Comparison between closed-set TTA and open-set TTA.

tings. First, we need to clarify that in the literature on OOD detection, out refers specifically to "outside the semantic space", whereas in the literature on OOD generalization, out refers specifically to "outside the covariate space". Here we follow the terminology used in [56]. According to the different types of distribution shifts, we divide the real-world data into four types:

- In-distribution (ID) data is the most common data we typically use to train a model, with a limited number of classes.
- Out-of-distribution (OOD) data contains some open classes that have not been seen before in ID data, with the same style and domain as ID data.
- Covariate-shifted ID (csID) data and ID data have the same classes and differ in styles and domains.
- Covariate-shifted OOD (csOOD) data is different from ID data in both classes and domains.

The open-set TTA setting takes into account both csID data and csOOD data.

Existing TTA methods make extensive use of entropy objective, which proves to be very effective. We first experimentally verify that existing TTA methods [27, 33, 35, 44, 46, 54] degrade the classification accuracy of known classes when open-set classes are included, which is consistent with the conclusions drawn from some recent studies [27, 28]. In addition, as shown in Fig. 1, the detection performance of unknown classes is also impaired, which has not received enough attention in previous studies. We attribute the performance degradation to the following two points. First, the presence of open-set samples leads to the incorrect estimation of normalization statistics by the model, leading to errors in updating affine parameters. Second, entropy minimization on samples from unknown classes forces the model to output confident predictions, undermining the model's confidence and leading to a decrease in the model's ability to distinguish between known classes and unknown classes.

With the aforementioned causes in mind, we propose three techniques to enhance the robustness of existing TTA methods under open-set setting. We first propose a distribution-aware filter to preliminarily distinguish between csID samples and csOOD samples. Specifically, we observe that the cosine similarity between the features extracted by the source model and the source domain prototypes can reflect the semantic shift, and we use this property to distinguish samples. We then propose a unified entropy optimization framework (UniEnt) to address the aforementioned challenges. UniEnt minimizes the entropy of csID samples while maximizing the entropy of csOOD samples simultaneously. Furthermore, we propose UniEnt+ using a sample-level weighting strategy to avoid the error caused by noisy data partition.

We summarize the contributions of this paper as follows.

- We first delve into the performance of existing methods under closed-set TTA and open-set TTA settings. We then summarize two reasons for the performance degradation of existing methods with open-set classes included.
- We propose a unified entropy optimization framework, which consists of a distribution-aware filter to distinguish csID and csOOD samples, entropy minimization on csID samples to obtain good classification performance on known classes and entropy maximization on csOOD samples to obtain good detection performance on unknown classes.
- Our proposed framework can be flexibly applied to many existing TTA methods and substantially improves their performance under open-set setting. Comprehensive experiments demonstrate the effectiveness of our approach.

2. Related Work

Test-time adaptation. Among all the approaches to solving covariate shifts, test-time adaptation has received much attention because of its challenging setting of accessing only the source model and unlabelled target data. Some of the initial work [24, 31, 33, 39, 44, 51] focused on improving TTA performance by estimating batch normalization statistics using test data and designing unsupervised objective functions, *e.g.*, TENT [44] proposed to optimize the affine parameters of batch normalization by minimizing the entropy of model outputs. These works mainly focus on static

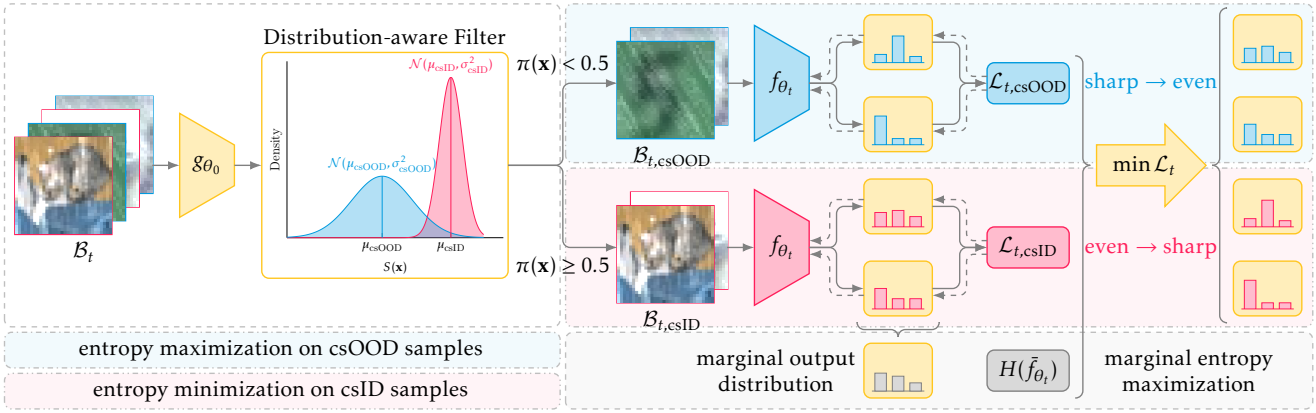


Figure 3. Illustration of the unified entropy optimization (UniEnt) framework. At timestamp t, mini-batch $\mathcal{B}_t$ may contain samples from csID and csOOD. First, we filter csOOD samples by csOOD score $S(\mathbf{x})$. Then, we perform entropy minimization for csID samples and entropy maximization for csOOD samples, we also adopt marginal entropy maximization to pervent model collapse. After optimization, we can yield better classification and detection performance tradeoff.

TTA and do not take into account the changes in the domain. After adapting to a target domain, the adapted model is reset to the one pretrained on the source domain to adapt to the next domain. Later, some work [35, 46] proposed the continual TTA setting where the model needs to adapt to a series of continuously changing target domains without knowing the domain labels. This poses new challenges for TTA: catastrophic forgetting and error accumulation. CoTTA [46] addresses the above issues through teacher-student model structure with data augmentation and stochastic recovery, while EATA [35] addresses the above issues through sample selection and anti-forgetting regularizer.

Robust test-time adaptation. Recently, several works have paid more attention to the robustness of TTA methods. LAME [2], NOTE [12] and RoTTA [53] focus on the performance of TTA methods under non-i.i.d. correlated sampling of test data. SITA [24] and MEMO [57] explore techniques for performing TTA on a single image. ODS [60] addresses case with label shift. OSTTA [27] pays attention to the performance degradation caused by long-term TTA. OWTTT [28] and OSTTA [27] consider the scenarios where the test data includes unknown classes. SAR [36] comprehensively analyzed the impact of mixed domain shifts, small batch sizes, and online imbalanced label distribution shifts on TTA performance. It is worth noting that there are some differences between the settings proposed by OWTTT [28] and OSTTA [27], the samples of unknown classes in OWTTT [28] are drawn from OOD, while the samples of unknown classes in OSTTA [27] are drawn from csOOD. We adopt the setting proposed in OSTTA [27] because of its practicality and challenging nature. First, the unknown class samples we encounter during TTA are likely to experience the same covariate shift. Second, it is more difficult to distinguish between csID samples and csOOD samples than between csID samples and OOD samples.

OOD detection. For models deployed in real-world scenarios, the ability of OOD detection is crucial. Recent studies in OOD detection can be roughly divided into two categories. The first type of approaches [15, 19, 20, 30, 32] is devoted to design sophisticated score functions and input-output transformations. MSP [15] uses the maximum softmax probability to detect OOD samples. ODIN [30] and generalized ODIN [20] further introduces temperature scaling, input preprocessing and confidence decompose to improve OOD detection performance. The second type of approaches instead regularizes the model by exploring the additional outlier data [16, 23, 48, 52, 55]. For example, OE [16] encourages the model to output low-confidence predictions for anomalous data. WOODS [23] on the other hand utilizes unlabelled wild data to improve the detection performance. SCONE [1] considers both OOD detection and OOD generalization for the first time. It is worth noting that all the methods mentioned above are designed for the training phase. Recently, AUTO [49] propose to optimize the network using unlabeled test data at test time to imporve OOD detection performance.

3. Methodology

3.1. Problem Setup

Let $\mathcal{D}_s = \{\mathbf{x}_i, y_i\}_{i=1}^{N_s}$ be the source domain dataset with label space $\mathcal{Y}_s = \{1, \cdots, C_s\}$, and $\mathcal{D}_t = \{\mathbf{x}_j, y_j\}_{j=1}^{N_t}$ be the target domain dataset with label space $\mathcal{Y}_t = \{1, \cdots, C_t\}$, where C_s and C_t denote the number of classes in the source and target domain datasets, respectively. C_s is equal to C_t for closed-set TTA while $C_s < C_t$ always holds for open-set TTA. Given a model f_{θ_0} pre-trained on $\mathcal{D}_s$, TTA aims to adapt the model to $\mathcal{D}_t$ without target labels accessible.

To be specific, we denote the mini-batch of test samples at timestamp t as $\mathcal{B}_t$ and the adapted model as f_{θ_t}. The main objective of open-set TTA is to correctly predict the classes in $\mathcal{Y}_s$ while reject the classes in $\mathcal{Y}_t \setminus \mathcal{Y}_s$ using the adapted model f_{θ_t}, especially in the presence of large data distribution shifts.

3.2. Preliminaries

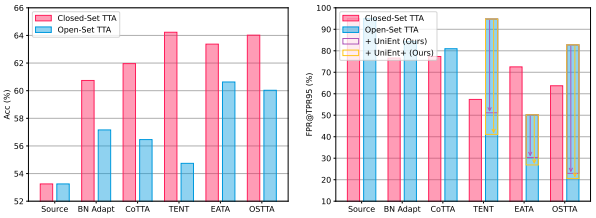
For closed-set TTA, a common practice [44] is to adapt the model by minimizing the unsupervised entropy objective:

$$\min_{\theta_t} \mathcal{L}_t = \frac{1}{\|\mathcal{B}_t\|} \sum_{\mathbf{x} \in \mathcal{B}_t} H(f_{\theta_t}(\mathbf{x})) - \lambda H(\bar{f}_{\theta_t}), \quad (1)$$

where $H(f_{\theta_t}(\mathbf{x})) = -\sum_{c=1}^{C} f_{\theta_t}^c(\mathbf{x}) \log f_{\theta_t}^c(\mathbf{x})$ denotes the entropy of the softmax output $f_{\theta_t}(\mathbf{x})$, $\bar{f}_{\theta_t} = \frac{1}{\|\mathcal{B}_t\|} \sum_{\mathbf{x} \in \mathcal{B}_t} f_{\theta_t}(\mathbf{x})$ represents the average softmax output over the mini-batch $\mathcal{B}_t$, and λ is a hyperparameter used to balance the two terms in the loss function. In previous studies [3, 6, 29, 31], marginal entropy $H(\bar{f}_{\theta_t})$ has been widely adopted to prevent model collapse, *i.e.*, predicting all input samples to the same class.

3.3. Motivation

There is no label of the test data to provide supervised information during TTA, an entropy minimization or a self-training strategy is widely adopted in existing methods. While previous studies [27, 33, 35, 44, 46, 54] focused on improving the performance of closed-set TTA, we empirically find that they exhibit performance degradation with open-set samples included. As shown in Fig. 4, We first compare the performance of existing TTA methods under different settings. Specifically, we conduct closed-set experiments on CIFAR-100-C [14], *i.e.*, updating the model and measuring the performance of the adapted model with only the test samples from known classes, and the open-set counterparts are extracted from Tab. 1. Experimental results show that applying existing methods to open-set TTA leads to the degradation of both the classification performance on known classes and the detection performance on unknown classes. We argue the degradation is caused by the following two reasons. First, the introduce of samples from unknown classes leads to the incorrect estimation of normalization statistics by the model, which results in unreliable updating of the model parameters. Second, entropy minimization-based methods achieved competitive closed-set results by making the model confident on the predictions. However, minimizing entropy on samples from unknown classes destroys the model confidence, which is an undesirable result. We believe that a good model confidence is very important, especially in open-set TTA, because it can tell us how much can we trust the adapted model's predictions.



(a) Acc↑ (b) FPR@TPR95↓

Figure 4. Performance comparison of existing TTA methods under closed-set and open-set settings.

3.4. Distribution-aware Filter

We first model the open-set data distribution as shown in Eq. (2):

$$\mathcal{P}_{\text{OPEN}} := \pi \mathcal{P}_{\text{csID}} + (1 - \pi) \mathcal{P}_{\text{csOOD}}, \quad (2)$$

where $\pi \in [0, 1]$. Equation (2) contains two distributions that the model may encounter during TTA:

- Covariate-shifted ID $\mathcal{P}_{\text{csID}}$ shares the label space with the training data, whereas the input space suffers from style and domain shifts.
- Covariate-shifted OOD $\mathcal{P}_{\text{csOOD}}$ differs from those of the training data in both the label space and the input space.

We define the csOOD score for each test sample as:

$$S(\mathbf{x}) = \nu \left(\max_c \frac{g_{\theta_0}(\mathbf{x}) \cdot p_c}{\|g_{\theta_0}(\mathbf{x})\| \|p_c\|} \right), \quad (3)$$

where $\nu(\cdot)$ denotes min-max normalization with the range of $[0, 1]$, g_{θ_0} denotes the feature extractor of source domain pre-trained model, p_c denotes the source domain prototype of class c.

As shown in Fig. 5, we empirically found that $S(\mathbf{x})$ can distinguish between csID samples and csOOD samples. To be more specific, the distribution of $S(\mathbf{x})$ appears to be bimodal, and its two peaks indicate csID and csOOD modes, respectively. In order to select the optimal threshold, we model the distribution of $S(\mathbf{x})$ as a Gaussian mixture model (GMM) with two components, where the component with larger mean corresponds to the csID samples, and vice versa:

$$\begin{aligned} \mathcal{P}(\mathbf{x}) =& \pi(\mathbf{x}) \mathcal{N}(\mathbf{x} \mid \mu_{\text{csID}}, \sigma_{\text{csID}}^2) \\ &+ (1 - \pi(\mathbf{x})) \mathcal{N}(\mathbf{x} \mid \mu_{\text{csOOD}}, \sigma_{\text{csOOD}}^2), \end{aligned} \quad (4)$$

where $\pi(\mathbf{x})$ denotes the probability that $S(\mathbf{x})$ belongs to the csID component, μ_{csID}, σ_{csID}^2 and μ_{csOOD}, σ_{csOOD}^2 represent the mean and variance of the csID and csOOD components, respectively. Further, $\pi(\mathbf{x})$ can be easily obtained using the EM algorithm.

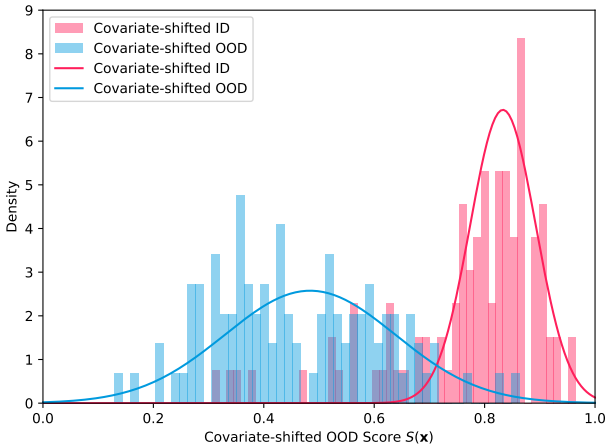


Figure 5. The csOOD score $S(\mathbf{x})$ presents a bimodal distribution.

Then, we can split $\mathcal{B}_t$ into $\mathcal{B}_{t,\text{csID}}$ and $\mathcal{B}_{t,\text{csOOD}}$ through Eq. (5):

$$\begin{aligned}\mathcal{B}_{t,\text{csID}} &= \{\mathbf{x} \mid \mathbf{x} \in \mathcal{B}_t \wedge \pi(\mathbf{x}) \geq 0.5\} \\ \mathcal{B}_{t,\text{csOOD}} &= \{\mathbf{x} \mid \mathbf{x} \in \mathcal{B}_t \wedge \pi(\mathbf{x}) < 0.5\}\end{aligned}, \tag{5}$$

where $\mathcal{B}_{t,\text{csID}}$ and $\mathcal{B}_{t,\text{csOOD}}$ are the mini-batches of pseudo csID and pseudo csOOD samples at timestamp t, respectively.

3.5. Unified Entropy Optimization

UniEnt. Based on the previous sections, we consider minimizing the entropy of the model's predictions of the samples from known classes, which can solve the inaccurate estimation of the data distribution and yield more reliable adaptation. However, the samples from unknown classes have not been explored effectively. Inspired by previous work [16, 23, 49], we propose to make the model produce approximately uniform predictions via entropy maximization instead, which can solve the inaccurate estimation of the model confidence and help distinguish known classes samples from unknown classes samples. The overall test-time optimization objective can be written as:

$$\mathcal{L}_{t,\text{csID}} = \frac{1}{\|\mathcal{B}_{t,\text{csID}}\|} \sum_{\mathbf{x} \in \mathcal{B}_{t,\text{csID}}} H(f_{\theta_t}(\mathbf{x})), \tag{6}$$

$$\mathcal{L}_{t,\text{csOOD}} = \frac{1}{\|\mathcal{B}_{t,\text{csOOD}}\|} \sum_{\mathbf{x} \in \mathcal{B}_{t,\text{csOOD}}} H(f_{\theta_t}(\mathbf{x})), \tag{7}$$

$$\min_{\theta_t} \mathcal{L}_t = \mathcal{L}_{t,\text{csID}} - \lambda_1 \mathcal{L}_{t,\text{csOOD}} - \lambda_2 H(\bar{f}_{\theta_t}), \tag{8}$$

where λ_1 and λ_2 are trade-off hyperparameters.

UniEnt+. In the distribution-aware filter, we distinguish csID samples from csOOD samples roughly, which inevitably introduces some noise. To address this problem, we propose a weighting scheme to achieve entropy minimization for known classes and entropy maximization for unknown classes at the same time. The objective can be reformulated as follows:

$$\begin{aligned}\min_{\theta_t} \mathcal{L}_t = &\frac{1}{\|\mathcal{B}_t\|} \sum_{\mathbf{x} \in \mathcal{B}_t} \pi(\mathbf{x}) H(f_{\theta_t}(\mathbf{x})) \\ &- \lambda_1 \frac{1}{\|\mathcal{B}_t\|} \sum_{\mathbf{x} \in \mathcal{B}_t} (1 - \pi(\mathbf{x})) H(f_{\theta_t}(\mathbf{x})) \\ &- \lambda_2 H(\bar{f}_{\theta_t})\end{aligned}. \tag{9}$$

4. Experiments

4.1. Setup

Datasets. Following previous studies, we evaluate our proposed methods on the widely used corruption benchmark datasets: CIFAR-10-C, CIFAR-100-C, and Tiny-ImageNet-C [14]. Each dataset contains 15 types of corruptions with 5 severity levels, all our experiments are conducted under the most severe corruption level 5. Pre-trained models are trained on the clean training set and tested and adapted on the corrupted test set. Following OSTTA [27], we apply the same corruption type to the original SVHN [34] and ImageNet-O [18] test sets to generate the SVHN-C and ImageNet-O-C datasets. We use SVHN-C and ImageNet-O-C as the covariate shifted OOD datasets for CIFAR-10/100-C and Tiny-ImageNet-C, respectively.

Evaluation protocols. Following recent research [27, 35, 44, 46], we evaluate TTA methods under continuously changing domains without resetting the parameters after each domain. At test time, the corrupted images are provided to the model in an online fashion. After encountering a mini-batch of test data, the model makes predictions and updates parameters immediately. The predictions of test data arriving at timestamp t will not be affected by any test data arriving after timestamp t. We construct the mini-batch using the same number of csID samples and csOOD samples. Regarding the model's adaptation performance on csID data, we use the accuracy metric. To evaluate whether the adapted model can detect csOOD data robustly, we measure the area under the receiver operating characteristic curve (AUROC) and the false positive rate of csOOD samples when the true positive rate of csID samples is at 95% (FPR@TPR95). As we pursue a good trade-off between the classification accuracy on csID data and the detection accuracy on csOOD data, we also report the open-set classification rate (OSCR) [9] to measure the balanced performance.

Baseline methods. We mainly compare our method with two types of pervious methods in TTA: 1) entropy-free methods: **Source** directly evaluates the test data using the source model without adaptation. **BN Adapt** [33]

Method	CIFAR-10-C				CIFAR-100-C				Average			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [54]	81.73	77.89	79.45	68.44	53.25	60.55	94.98	39.87	67.49	69.22	87.22	54.16
BN Adapt [33]	84.20	80.40	76.84	72.13	57.16	72.45	84.29	47.10	70.68	76.43	80.57	59.62
CoTTA [46]	85.77	85.89	72.40	77.26	56.46	77.04	80.96	48.95	71.12	81.47	76.68	63.11
TENT [44]	79.38	65.39	95.94	56.73	54.74	65.00	94.79	42.24	67.06	65.20	95.37	49.49
+ UniEnt	**84.31** (+4.93)	92.28 (+26.89)	36.74 (-59.20)	80.32 (+23.59)	**59.07** (+4.33)	89.28 (+24.28)	51.14 (-43.65)	56.26 (+14.02)	**71.69** (+4.63)	90.78 (+25.59)	43.94 (-51.43)	68.29 (+18.81)
+ UniEnt+	84.03 (+4.65)	**93.18** (+27.79)	**32.74** (-63.20)	**80.62** (+23.89)	58.58 (+3.84)	**91.39** (+26.39)	**41.09** (-53.70)	**56.36** (+14.12)	71.31 (+4.25)	**92.29** (+27.09)	**36.92** (-58.45)	**68.49** (+19.01)
EATA [35]	80.92	84.32	71.66	72.63	**60.63**	88.64	50.18	57.24	70.78	86.48	60.92	64.94
+ UniEnt	84.31 (+3.39)	**97.15** (+12.83)	**13.25** (-58.41)	82.99 (+10.36)	59.75 (-0.88)	93.42 (+4.78)	30.36 (-19.82)	57.99 (+0.75)	72.03 (+1.26)	95.29 (+8.81)	21.81 (-39.12)	70.49 (+5.55)
+ UniEnt+	**85.18** (+4.26)	96.97 (+12.65)	14.28 (-57.38)	**83.67** (+11.04)	59.71(-0.92)	**94.23** (+5.59)	**26.87** (-23.31)	**58.19** (+0.95)	**72.45** (+1.67)	**95.60** (+9.12)	**20.58** (-40.35)	**70.93** (+6.00)
OSTTA [27]	**84.44**	72.74	77.02	65.17	**60.03**	75.37	82.75	51.35	**72.24**	74.06	79.89	58.26
+ UniEnt	82.46 (-1.98)	96.20 (+23.46)	16.37 (-60.65)	80.51 (+15.34)	58.69 (-1.34)	94.84 (+19.47)	22.95 (-59.80)	57.28 (+5.93)	70.58 (-1.66)	95.52 (+21.47)	19.66 (-60.23)	68.90 (+10.64)
+ UniEnt+	84.30 (-0.14)	**97.38** (+24.64)	**11.56** (-65.46)	**82.91** (+17.74)	58.93 (-1.10)	**95.42** (+20.05)	**20.59** (-62.16)	**57.69** (+6.34)	71.62 (-0.62)	**96.40** (+22.35)	**16.08** (-63.81)	**70.30** (+12.04)

Table 1. Results of different methods on CIFAR benchmarks. ↑ indicates that larger values are better, and vice versa. All values are percentages. The **bold** values indicate the best results, and the underlined values indicate the second best results.

Method	Tiny-ImageNet-C			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [54]	22.29	53.79	93.41	16.29
BN Adapt [33]	37.00	61.06	90.90	28.50
TENT [44]	28.96	49.78	95.96	19.02
+ UniEnt	37.23 (+8.27)	**63.92** (+14.14)	89.72 (-6.24)	**30.18** (+11.16)
+ UniEnt+	**37.31** (+8.35)	63.83 (+14.05)	**89.12** (-6.84)	30.12 (+11.10)
EATA [35]	37.09	57.55	93.22	27.91
+ UniEnt	37.54 (+0.45)	**64.34** (+6.79)	**89.23** (-3.99)	30.59 (+2.68)
+ UniEnt+	**38.65** (+1.56)	62.30 (+4.75)	90.88 (-2.34)	**30.95** (+3.04)
OSTTA [27]	**37.29**	55.66	94.34	**27.74**
+ UniEnt	33.72 (-3.57)	**62.69** (+7.03)	89.67 (-4.67)	26.63 (-1.11)
+ UniEnt+	34.47 (-2.82)	61.28 (+5.62)	**89.56** (-4.78)	26.65 (-1.09)

Table 2. Results of different methods on Tiny-ImageNet-C.

updates batch normalization statistics with the test data during TTA. **CoTTA** [46] adopts the teacher-student architecture to provide weight-averaged and augmentation-averaged pseudo-labels to reduce error accumulation, combined with stochastic restoration to avoid catastrophic forgetting. 2) entropy-based methods: **TENT** [44] estimates normalization statistics and optimizes channel-wise affine transformations through entropy minimization. **EATA** [35] selects reliable and non-redundant samples for model adaptation, the former achieve prediction entropy lower than a pre-defined threshold and the latter have diverse model outputs. In addition, the fisher regularization is introduced to prevent catastrophic forgetting. **OSTTA** [27] uses the wisdom of crowds to filter out the samples with lower confidence values in the adapted model than in the original model. Our methods can be easily applied to existing entropy-based methods without additional modification. Regarding applying our methods to EATA and OSTTA, we apply the filtering methods and keep everything else the same.

Implementation details. For experiments on CIFAR benchmarks, following pervious studies [6, 27, 31], we use the WideResNet [54] with 40 layers and widen factor of 2. The model pre-trained with AugMix [17] is available from RobustBench [7]. For Tiny-ImageNet-C, we pre-train ResNet50 [13] on the Tiny-ImageNet [26] training set, as OSTTA [27] did. The model is initialized with the pre-trained weights on ImageNet [8] and optimized for 50 epochs using SGD [38] with a batch size of 256. The initial learning rate is set to 0.01 and adjust using a cosine annealing schedule. During TTA, we use Adam [25] optimizer with the batch size of 200 for all experiments. The learning rate is set to 0.001 and 0.01 for entropy-based methods (TENT [44], EATA [35], OSTTA) and CoTTA [46], respectively. We use the energy score [32] to measure the ability of the adapted model to detect unknown classes. Furthermore, following T3A [22], we use the weights of the linear classifier as the source domain prototypes, and thus our approach is source-free. Entropy-based methods update only the affine parameters, while CoTTA updates all parameters.

4.2. Results

CIFAR benchmarks. We first conduct experiments on the most common CIFAR benchmarks, and the results are presented in Tab. 1. From Tab. 1, we can see that UniEnt and UniEnt+ significantly improve the performance of three different existing TTA methods. For example, on CIFAR-10-C, UniEnt improves the Acc, AUROC, FPR@TPR95 and OSCR of TENT [44] by 4.93%, 26.89%, 59.20% and 23.59% respectively, while UniEnt+ improves the Acc, AUROC, FPR@TPR95 and OSCR of TENT by 4.65%, 27.79%, 63.20% and 23.89% respectively.

In more detail, we can observe that TENT [44] and OSTTA [27] perform even worse than Source and BN methods that do not update model parameters in some cases (OSCR decreases by 3.27%∼15.40%), which indicates that some existing TTA methods cannot effectively update model parameters with open-set classes included. This can be attributed to the fact that these methods ignore the distribution variations introduced by open-set samples, resulting in the unreliable estimation of normalization statistics and model confidence.

Tiny-ImageNet-C. We then conduct experiments on a more challenging dataset Tiny-ImageNet-C, and the results are summarized in Tab. 2. As shown in Tab. 2, consistent with

Method	$\mathcal{L}_{t,\text{csID}}$	$\mathcal{L}_{t,\text{csOOD}}$	CIFAR-10-C				CIFAR-100-C			
			Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
TENT [44]	✗	✗	79.38	65.39	95.94	56.73	54.74	65.00	94.79	42.24
	✓	✗	**85.04** (+5.66)	81.80 (+16.41)	68.89 (-27.05)	73.57 (+16.84)	**59.30** (+4.56)	86.09 (+21.09)	63.65 (-31.14)	55.55 (+13.31)
	✓	✓	84.31 (+4.93)	**92.28** (+26.89)	**36.74** (-59.20)	**80.32** (+23.59)	59.07 (+4.33)	**89.28** (+24.28)	**51.14** (-43.65)	**56.26** (+14.02)
EATA [35]	✗	✗	80.92	84.32	71.66	72.63	**60.63**	88.64	50.18	57.24
	✓	✗	**85.53** (+4.61)	82.94 (-1.38)	67.95 (-3.71)	74.85 (+2.22)	60.46 (-0.17)	88.53 (-0.11)	54.30 (+4.12)	57.26 (+0.02)
	✓	✓	84.31 (+3.39)	**97.15** (+12.83)	**13.25** (-58.41)	**82.99** (+10.36)	59.75 (-0.88)	**93.42** (+4.78)	**30.36** (-19.82)	**57.99** (+0.75)
OSTTA [27]	✗	✗	84.44	72.74	77.02	65.17	**60.03**	75.37	82.75	51.35
	✓	✗	**84.86** (+0.42)	84.96 (+12.22)	62.66 (-14.36)	75.84 (+10.67)	58.95 (-1.08)	90.62 (+15.25)	44.79 (-37.96)	56.50 (+5.15)
	✓	✓	82.46 (-1.98)	**96.20** (+23.46)	**16.37** (-60.65)	**80.51** (+15.34)	58.69 (-1.34)	**94.84** (+19.47)	**22.95** (-59.80)	**57.28** (+5.93)

Table 3. Ablation study on CIFAR benchmarks. We investigate the effectiveness of $\mathcal{L}_{t,\text{csID}}$ and $\mathcal{L}_{t,\text{csOOD}}$ in Eq. (8) for UniEnt.

Method		0.1	0.2	0.5	1.0	Δ
TENT [44]	+ UniEnt	(59.09, 89.11, 51.68, 56.20)	(59.07, 89.28, 51.14, 56.26)	(58.92, 89.59, 50.16, 56.22)	(58.76, 89.95, 48.92, 56.21)	(0.33, 0.84, 2.76, 0.06)
	+ UniEnt+	(58.64, 91.18, 41.79, 56.34)	(58.58, 91.39, 41.09, 56.36)	(58.41, 91.68, 40.22, 56.33)	(58.12, 91.89, 39.68, 56.13)	(0.52, 0.71, 2.11, 0.23)
EATA [35]	+ UniEnt	(59.50, 93.34, 30.72, 57.72)	(59.75, 93.42, 30.36, 57.99)	(59.37, 92.56, 34.98, 57.40)	(59.58, 93.82, 28.29, 57.97)	(0.38, 1.26, 6.69, 0.59)
	+ UniEnt+	(59.73, 93.47, 30.25, 58.00)	(59.81, 93.88, 27.84, 58.17)	(59.71, 94.23, 26.87, 58.19)	(59.62, 93.47, 30.37, 57.91)	(0.19, 0.76, 3.50, 0.28)
OSTTA [27]	+ UniEnt	(58.85, 93.89, 26.59, 57.14)	(58.82, 94.32, 24.94, 57.24)	(58.69, 94.84, 22.95, 57.28)	(57.88, 94.80, 23.51, 56.51)	(0.97, 0.95, 3.64, 0.77)
	+ UniEnt+	(59.25, 94.19, 24.62, 57.54)	(59.15, 94.84, 22.29, 57.69)	(58.93, 95.42, 20.59, 57.69)	(58.20, 95.65, 20.12, 57.06)	(1.05, 1.46, 4.50, 0.63)

Table 4. Performance of UniEnt and UniEnt+ with varying λ_1 on CIFAR-100-C. The values in the table are presented as (Acc, AUROC, FPR@TPR95, OSCR). Δ is the difference between the maximum and minimum values when λ_1 take different values. Smaller Δ values represent better robustness.

previous analysis, UniEnt and UniEnt+ still achieve better performance. Numerically, UniEnt improves the Acc, AUROC, FPR@TPR95 and OSCR of TENT [44] by 8.27%, 14.14%, 6.24% and 11.16% respectively, while UniEnt+ improves the Acc, AUROC, FPR@TPR95 and OSCR of TENT by 8.35%, 14.05%, 6.84% and 11.10% respectively.

4.3. Analysis

Ablation study. To verify the effectiveness of different components in $\mathcal{L}_t$ (Eq. (8)), we conduct extensive ablation studies on CIFAR benchmarks. The results are summarized in Tab. 3. Compared with the baselines without $\mathcal{L}_{t,\text{csID}}$ and $\mathcal{L}_{t,\text{csOOD}}$ (the same as TENT [44], EATA [35] and OSTTA [27]), introducing $\mathcal{L}_{t,\text{csID}}$ improves the classification accuracy of known classes, which indicates that our proposed distribution-aware filter can well distinguish the samples of known classes from the samples of unknown classes. It is worth noting that the introduction of $\mathcal{L}_{t,\text{csID}}$ also leads to better detection performance of unknown classes, which is consistent with the findings obtained in a recent study [43]. With the addition of $\mathcal{L}_{t,\text{csOOD}}$, the model's detection performance of unknown classes has been further improved. Considering the trade-off between the two, UniEnt achieves the optimal OSCR values in most cases.

Hyperparameter sensitivity. We perform sensitivity analyses on the hyperparameters λ_1 and λ_2, as summarized in Tab. 4 and Tab. 5. We first investigate the effect of λ_1 on CIFAR-100-C, with λ_1 taking values from $\{0.1, 0.2, 0.5, 1.0\}$ and λ_2 holds constant. The experimental results show that our methods are robust to the value of λ_1, the gaps between the best and worst values of Acc, AUROC, FPR@TPR95 and OSCR are 1.05%, 1.46%, 6.69% and 0.77%, respectively. We then examine how λ_2 affects csID classification and csOOD detection, with λ_2 taking values from $\{0.1, 0.2, 0.5, 1.0\}$ and λ_1 holds constant. It is easy to conclude from the results that a larger λ_2 leads to better csOOD detection performance, yet at the same time, it may lose some of the csID classification performance, and vice versa. Numerically, different values of λ_2 will result in the maximum performance differences of 15.49%, 7.51%, 35.06% and 15.41% for Acc, AUROC, FPR@TPR95 and OSCR, respectively.

Performance under different number of unknown classes. The number of unknown classes is an important measure representing the complexity of the open-set. We examine the impact of different numbers of unknown classes. Specifically, we perform experiments on the CIFAR-10-C dataset and control the number of unknown classes to vary from 2 to 10, keeping the number of samples constant. From Tab. 6, we can see that TENT [44] fluctuates with different number of classes while the proposed UniEnt and UniEnt+ are more robust to different number of unknown classes.

Performance under different ratios of csOOD to csID samples. We also perform experiments with different ratios of the number of csOOD samples to the number of csID samples, and the results are displayed in Tab. 7. We vary

Method		0.1	0.2	0.5	1.0	Δ
TENT [44]	+ UniEnt	(59.44, 87.02, 60.32, 55.93)	(59.07, 89.28, 51.14, 56.26)	(58.09, 92.87, 33.24, 56.23)	(56.62, 94.53, 25.26, 55.24)	(2.82, 7.51, 35.06, 1.02)
	+ UniEnt+	(59.19, 87.95, 57.31, 56.04)	(58.58, 91.39, 41.09, 56.36)	(56.71, 94.57, 25.02, 55.34)	(53.13, 94.93, 24.19, 52.01)	(6.06, 6.98, 33.12, 4.35)
EATA [35]	+ UniEnt	(60.54, 88.14, 55.48, 57.15)	(60.06, 89.45, 50.99, 57.16)	(59.75, 93.42, 30.36, 57.99)	(58.26, 95.07, 22.18, 57.02)	(2.28, 6.93, 33.30, 0.97)
	+ UniEnt+	(60.35, 89.49, 50.20, 57.44)	(60.51, 91.03, 42.50, 58.02)	(59.71, 94.23, 26.87, 58.19)	(59.03, 95.28, 21.20, 57.81)	(1.48, 5.79, 29.00, 0.75)
OSTTA [27]	+ UniEnt	(58.69, 94.84, 22.95, 57.28)	(56.63, 95.43, 21.02, 55.46)	(49.85, 93.77, 32.12, 48.59)	(43.89, 91.19, 47.50, 42.41)	(14.80, 4.24, 26.48, 14.87)
	+ UniEnt+	(59.15, 94.84, 22.29, 57.69)	(57.55, 95.82, 18.91, 56.43)	(50.31, 94.09, 30.05, 49.11)	(43.66, 91.78, 43.35, 42.28)	(15.49, 4.04, 24.44, 15.41)

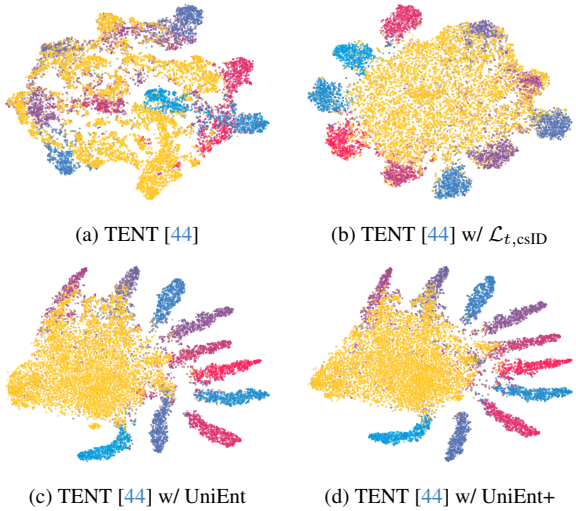
Table 5. Performance of UniEnt and UniEnt+ with varying λ_2 on CIFAR-100-C. Δ is the difference between the maximum and minimum values when λ_2 take different values.

Method	2	4	6	8	10	Δ
Source [54]	70.84	69.28	69.32	69.18	68.44	2.40
BN Adapt [33]	72.56	72.48	72.52	72.44	72.14	0.42
TENT [44]	49.51	48.29	51.74	49.53	50.97	3.45
+ UniEnt	78.71	78.39	78.28	78.13	77.82	0.89
+ UniEnt+	78.65	78.23	78.23	78.07	77.68	0.97

Table 6. OSCR of UniEnt and UniEnt+ on CIFAR-10-C under different number of unknown classes.

Method	0.2	0.4	0.6	0.8	1.0	Δ
Source [54]	40.00	40.03	39.98	39.92	39.87	0.16
BN Adapt [33]	49.91	49.55	48.92	47.97	47.10	2.81
TENT [44]	47.68	44.12	44.06	42.90	42.16	5.52
+ UniEnt	56.84	57.48	57.13	56.77	56.26	1.22
+ UniEnt+	57.15	57.59	57.24	56.88	56.33	1.26

Table 7. OSCR of UniEnt and UniEnt+ on CIFAR-100-C under different ratios of csOOD to csID samples.



(a) TENT [44] (b) TENT [44] w/ $\mathcal{L}_{t,\text{csID}}$

(c) TENT [44] w/ UniEnt (d) TENT [44] w/ UniEnt+

Figure 6. T-SNE visualization on CIFAR-10-C test set with SVHN-C as csOOD. red → blue denotes csID samples and yellow denotes csOOD samples.

the data ratio from 0.2 to 1.0. It can be observed that our proposed methods are insensitive to the variation of the data ratio while TENT [44] is more sensitive, and thus can be applied to different data ratio cases.

T-SNE visualization. To illustrate the effects of different methods on csID classification and csOOD detection, we visualize the feature representations of CIFAR-10-C test samples with SVHN-C test samples as csOOD samples via T-SNE [42] in Fig. 6. It can be observed that the features from known classes and unknown classes adapted by TENT [44] are mixed together, while UniEnt and UniEnt+ can better separate them. Furthermore, we observe that filtering out csOOD samples (w/ $\mathcal{L}_{t,\text{csID}}$) can not only improve the classification performance on known classes, but also the detection performance on unknown classes.

5. Conclusion

This paper presents a unified entropy optimization framework for open-set test-time adaptation that can be flexibly applied to various existing TTA methods. We first delve into the performance of existing methods under open-set TTA setting, and attribute the performance degradation to the unreliable estimation of normalization statistics and model confidence. To address the above issues, we then propose a distribution-aware filter to preliminary distinguish csID samples from csOOD samples, followed by entropy minimization on csID samples and entropy maximization on csOOD samples. In addition, we propose to leverage sample-level confidence to reduce the noise from hard data partition. Extensive experiments reveal that our methods outperform state-of-the-art TTA methods in open-set scenarios. We hope that more studies can focus on the robustness of TTA methods under open-set, which can facilitate the application of these methods in real scenarios.

Acknowledgements. This work has been supported by the National Science and Technology Major Project (2022ZD0116500), National Natural Science Foundation of China (U20A20223, 62222609, 62076236), CAS Project for Young Scientists in Basic Research (YSBR-083), and Key Research Program of Frontier Sciences of CAS (ZDBS-LY-7004).

References

[1] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *ICML*, 2023. 3

[2] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *CVPR*, 2022. 3

[3] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, 2022. 4

[4] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, 2020. 1

[5] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2021. 1

[6] Sungha Choi, Seunghan Yang, Seokeon Choi, and Sungrack Yun. Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In *ECCV*, 2022. 4, 6

[7] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2021. 6, 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[9] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *NeurIPS*, 2018. 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 1

[12] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *NeurIPS*, 2022. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 1

[14] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 4, 5, 1

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 3, 2

[16] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 1, 3, 5

[17] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2020. 6, 1

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 5, 1

[19] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022. 1, 3, 2

[20] 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 *CVPR*, 2020. 3

[21] Hongzhi Huang, Yu Wang, Qinghua Hu, and Ming-Ming Cheng. Class-specific semantic reconstruction for open set recognition. *IEEE TPAMI*, 2022. 1

[22] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *NeurIPS*, 2021. 6

[23] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *ICML*, 2022. 3, 5

[24] Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. Sita: Single image test-time adaptation. *arXiv preprint arXiv:2112.02355*, 2021. 2, 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015. 6

[27] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *ICCV*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[28] 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 *ICCV*, 2023. 1, 2, 3

[29] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 4

[30] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 3

[31] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *ICLR*, 2023. 2, 4, 6

[32] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 1, 3, 6, 2

[33] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 1, 2, 4, 5, 6, 8, 3

[34] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[35] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[36] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *ICLR*, 2023. 3

[37] Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. Rdumb: A simple approach that questions our progress in continual test-time adaptation. *arXiv:2306.05401*, 2023. 1

[38] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 6, 1

[39] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *NeurIPS*, 2020. 2

[40] Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yugang Jiang. Deeper insights into vits robustness towards common corruptions. *arXiv:2204.12143*, 2022. 1

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[43] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2022. 1, 7

[44] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 1, 2, 4, 5, 6, 7, 8, 3

[45] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *TKDE*, 2022. 1

[46] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6

[47] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, 2021. 1

[48] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *ICCV*, 2021. 3

[49] 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, 5

[50] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 1

[51] Fuming You, Jingjing Li, and Zhou Zhao. Test-time batch statistics calibration for covariate shift. *arXiv preprint arXiv:2110.04065*, 2021. 2

[52] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *ICCV*, 2019. 3

[53] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *CVPR*, 2023. 3

[54] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 2, 4, 6, 8, 1, 3

[55] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *WACV*, 2023. 3

[56] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 1, 2

[57] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *NeurIPS*, 2022. 3

[58] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 1

[59] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE TPAMI*, 2022. 1

[60] 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 *ICML*, 2023. 3