DM-POSA: Enhancing Open-World Test-Time Adaptation with Dual-Mode Matching and Prompt-Based Open Set Adaptation

Shiji Zhao[1,2], **Shao-Yuan Li**[1,2,3 *], **Chuanxing Geng**[1,2], **Sheng-Jun Huang**[1,2], **Songcan Chen**[1,2]
[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
[3]State Key Lab. for Novel Software Technology, Nanjing University, P.R. China
{zhaosj, lisy, gengchuanxing, huangsj, s.chen}@nuaa.edu.cn

Abstract

The need to generalize the pre-trained deep learning models to unknown test-time data distributions has spurred research into test-time adaptation (TTA). Existing studies have mainly focused on closed-set TTA with only covariate shifts, while largely overlooking open-set TTA that involves *semantic shifts*, i.e., *unknown open-set classes*. However, addressing adaptation to unknown classes is crucial for open-world safety-critical applications such as autonomous driving. In this paper, we emphasize that accurate identification of the open-set samples is rather challenging in TTA. The entanglement of semantic shift and covariate shift mutually confuse the network's discriminative capability. This co-interference further exacerbates considering the single-pass data nature and low latency requirement. With this understanding, we propose **D**ual-mode **M**atching and **P**rompt-based **O**pen **S**et **A**daptation (DM-POSA) for open-set TTA to enhance discriminative feature learning and unknown classes distinguishment with minimal time cost. DM-POSA identifies open-set samples via dual-mode matching strategies, including model-parameter-based and feature-space-based matching. It also optimizes the model with a random pairing discrepancy loss, enhancing the distributional difference between open-set and closed-set samples, thus improving the model's ability to recognize unknown categories. Extensive experiments show the superiority of DM-POSA over state-of-the-art baselines on both closed-set class adaptation and open-set class detection.

1 Introduction

Deep neural networks have achieved great success in a wide range of machine learning tasks. Nevertheless, they often exhibit brittleness and vulnerability when confronted with data distribution shifts.Therefore, enhancing the robustness of deep models against distribution shifts has become a critical and actively researched area.

*Corresponding author: Shao-Yuan Li.

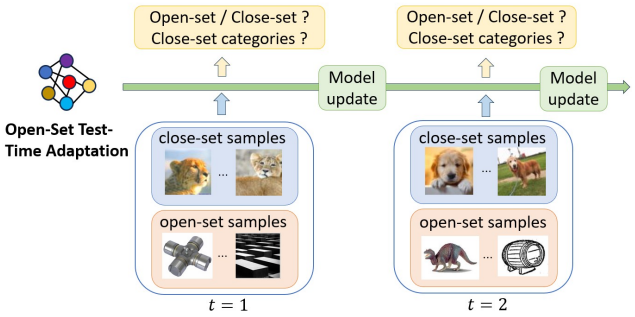


Figure 1: Open-Set Test-Time Adaptation

Test-time adaptation (TTA) emerges as a significant research frontier through adapting pre-trained models to unforeseen deployment distribution shifts. Aligning well with real-world scenarios, TTA has motivated many research efforts, including test time normalization [Schneider *et al.*, 2020; Nado *et al.*, 2020], entropy minimization [Wang *et al.*, 2021], self-supervised learning [Sun *et al.*, 2020; Liu *et al.*, 2021], contrastive learning [Chen *et al.*, 2022], data augmentation [Zhang *et al.*, 2022], uncertainty-aware optimization [Niu *et al.*, 2022], online continual adaptation [Boudiaf *et al.*, 2022; Wang *et al.*, 2022; Zhang *et al.*, 2023], the small batch size adaptation [Niu *et al.*, 2023; Zhao *et al.*, 2024], as well as TTA with pre-trained vision-language models [Hakim *et al.*, 2024; Feng *et al.*, 2023].

Nevertheless, there is a noticeable gap in the literature concerning the issue of semantic shift in TTA. As depicted in Fig.1, in open-world applications that require models to be deployed in diverse environments, the test domain not only experiences covariate shift (domain/style shift) but also suffers from semantic shift (open-set classes/new categories). Existing TTA methods encounter significant challenges in such situations. Fig.2 illustrates the performance drop of representative TTA approaches when the test data contains open-set samples. These methods lack robust loss functions or mechanisms to identify open-set classes. When the new categories are incorporated as closed-set samples for model adaptation, they tend to mislead the covariate shift estimation and alignment.

Previously only a few works have explored the open-set challenge in TTA. OWTTT [Li *et al.*, 2023] firstly pro-

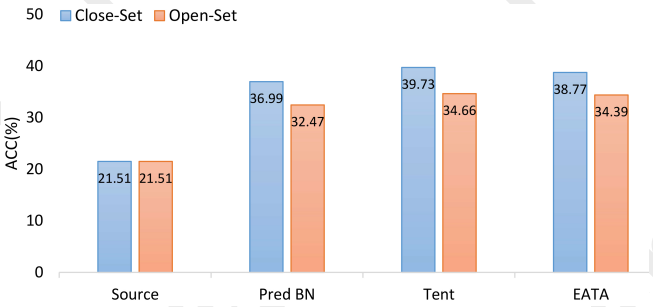


Figure 2: Performance degradation of traditional TTA methods in the presence of open-set samples.

posed the open-world test-time training/adaptation (OWTTT) concept, referring to the semantic shift as strong out-of-distribution (OOD) and the covariate shift as weak OOD. Built under the self-training framework, OWTTT lessened the influence of incorrectly pseudo-labelled strong OOD samples by pruning out samples far from the source domain prototypes. OSTTA [Lee *et al.*, 2023] and UniEnt [Gao *et al.*, 2024] analyzed that open-set samples undermine the model's confidence and thus corrupt the extensively used entropy minimization objective. OSTTA [Lee *et al.*, 2023] treated the samples whose confidence values are lower and misaligned with 'wisdom of crowds' as open-set and filtered them out. UniEnt [Gao *et al.*, 2024] distinguished open-set samples based on the cosine similarity between target sample features and the source domain prototypes, then conducted entropy minimization and maximization respectively on the closed-set and open-set samples.

These works open the door and shed insightful light on the area of open-set TTA. However, the sophisticateness of open-set TTA leaves large room for further exploration. The key challenge of open-set TTA lies in that, the dual factors of semantic shift detection and covariate shift alignment entangle and interfere with one another. Imprecise detection of open-set classes would confuse the network's discriminative capability on the closed-set covariate shift samples. This confirmation bias exacerbates encountering the single-pass nature of TTA, which intrinsically challenges effective feature learning within constrained adaptation time.

Based on straightforward measures defined over the model predictions or feature space similarity, the above open-set sample detection efforts are limited for difficult open-set samples, which resemble some closed-set classes. Additionally, simply discarding the identified open-set samples misses the potential use of the hidden information within them.

With these concerns, we propose DM-POSA, an open-set TTA method based on dual-mode matching and pairing discrepancy loss. The core idea of DM-POSA is to improve the model's ability to distinguish open-set samples by accurate open-set sample recognition and robust open-set representation learning. Upon the arrival of each test batch, DM-POSA quickly identifies open-set samples through a dual-mode matching strategy including model-parameter-based and feature-space-based matching. It further enhances the model's adaptation to difficult open-set samples by optimizing the pairing discrepancy loss, thereby effectively avoiding the negative impact of open-set data on model performance.

In summary, our contributions are as follows:

- We analyze and emphasize that due to the entanglement of covariate shift and semantic shit, accurate identification of open-set samples is rather challenging, especially under the single-pass data protocol and low latency requirement of TTA.
- We propose a timely DM-POSA approach using dual-mode matching to quickly identify reliable closed-set and potential open-set samples in streaming data. Additionally, it introduces a visual prompt-based pairing discrepancy loss to enhance the distributional difference between open-set and closed-set categories, improving the recognition of challenging open-set samples.
- We conduct extensive experimental validations on multiple datasets, and show the superiority of DM-POSA in distinguishing open-set samples while ensuring closed-set classification accuracy.

2 Related Works

Test-Time Adaptation Research on test-time adaptation has emerged rapidly in recent years. Early works mainly focused on adjustments to the batch normalization (BN) layers by recalculating normalization statistics for each test batch of data [Nado *et al.*, 2020; Schneider *et al.*, 2020]. Subsequently, entropy minimization-based methods gained prominence. TENT [Wang *et al.*, 2021] optimizes batch normalization parameters via entropy minimization, while Memo [Zhang *et al.*, 2022] reduces output entropy across augmented inputs to improve robustness. TEA [Yuan *et al.*, 2024] aligns the model's distribution to test data by transforming classifiers into energy-based models. Pseudo-labeling methods include AdaContrast [Chen *et al.*, 2022], which combines contrastive learning with pseudo-label refinement, and NC-TTT [Osowiechi *et al.*, 2024], which improves adaptation by distinguishing noisy features via contrastive learning. RMT [Döbler *et al.*, 2023] uses a mean teacher model with symmetric cross-entropy for consistency loss. Non-parametric approaches, like LAME [Boudiaf *et al.*, 2022], adjust outputs without changing model parameters, leveraging Laplacian matrix adjustments. AdaNPC [Zhang *et al.*, 2023] uses memory-based voting to iteratively align source and target distributions.

Open-Set Test-Time Adaptation Several works have explored open-set test-time adaptation, which extends TTA by addressing scenarios where the test data includes samples unseen during the training phase. They are designed either within a self-training framework or on the basis of entropy minimization. OWTTT [Li *et al.*, 2023] designs an adaptive outlier data pruning strategy and represents out-of-distribution samples through dynamic prototype expansion. By calculating the similarity between the test data and the entire prototype pool, it can distinguish whether the sample is open-set. OSTTA [Lee *et al.*, 2023] proposes a population-based sample selection strategy that reduces the negative impact of open-set samples by filtering out samples with low confidence in the model. UniEnt [Gao *et al.*, 2024] introduces an integrated framework, which combines pseudo-label

generation with close-set entropy minimization and open-set entropy maximization.

3 Proposed Approach: DM-POSA

3.1 Preliminary

Let $D_s = \{(x_i, y_i)\}_{i=1}^{N_s}$ be the source domain dataset with label space $Y_s = \{1, \cdots, C_s\}$, and $D_t = \{x_j\}_{j=1}^{N_t}$ be the target domain dataset with label space $Y_t = \{1, \cdots, C_t\}$, where C_s and C_t denote the number of classes in the source and target domain datasets, respectively. In open-set TTA, the label space of the target domain satisfies $Y_t = Y_s \cup Y_o$, where Y_s is the set of known (closed-set) classes and Y_o is the set of unknown (open-set) classes, with $Y_s \cap Y_o = \emptyset$. The test data arrives sequentially in mini-batches, denoted as B_t at timestamp t. Given a model f_{θ_0} pre-trained on D_s, the objective of open-set TTA is to correctly predict the classes in Y_s while rejecting the classes in Y_o using the adapted model f_{θ_t}, especially under significant distribution shifts between D_s and D_t. Specifically, we assume a covariate shift scenario where $P_s(X_s) \neq P_t(X_s)$ but $P_s(Y_s|X_s) = P_t(Y_s|X_s)$, ensuring the label conditional distribution remains consistent across domains. At each time step t, the model needs to determine whether a sample x_j belongs to Y_s or Y_o, classify it if it belongs to Y_s, and update itself to obtain f_{θ_t}.

Due to the co-occurrence of covariate shift and semantic shift, effectively adapting pre-trained models to open-set TTA with low latency is rather challenging. In this paper, we propose Dual-mode Matching and Prompt-based Open Set Adaptation (DM-POSA) to tackle this problem. DM-POSA first quickly distinguishes open-set samples through an efficient dual-mode pattern matching based on parameter update and feature change. Then makes better use of them through a visual prompt enhanced divergence loss to amplify the distribution difference between open-set and closed-set samples, which induces robust representation learning and better hard-to-distinguish open-set sample detection. The overall framework of DM-POSA is shown in Figure 3.

3.2 Pattern Matching Based on Parameter Update

We first propose a pattern-matching strategy based on model parameter updates for identifying open-set samples. This design is based on entropy minimization. Based on the convergence of gradient descent [Robbins and Monro, 1951], it can be guaranteed that under an appropriate learning rate, after the model computes the gradient for a single sample and updates, the entropy of the label prediction for the sample will decrease when the common entropy minimization loss $L(\theta, x) = H(\theta, x)$ is used:

$$H(f_{\theta_{t+1}}(x)) \leq H(f_{\theta_t}(x)). \tag{1}$$

However, since the model update is done over an entire batch of data, for a single sample x, the entropy of its label prediction may not necessarily decrease. This depends on how well the desired gradient update direction for the sample matches the average gradient update direction of the entire batch. Thus, we define the sample gradient matching as follows:

$$d(g_x, \bar{g}) = \frac{1}{L}\sum_{l=1}^{L} \frac{< g_x, \bar{g} >}{||g_x|| \cdot ||\bar{g}||}, \quad \bar{g} = \frac{1}{N}\sum_{i=1}^{N} g_{x_i}, \tag{2}$$

where L is the number of layers the model updates. For open-set samples, their gradient behaves like noise, and the update direction is often erroneous and erratic. Therefore, $\bar{g}$ is dominated by the updates from close-set samples. Using the distribution of $d(g_x, \bar{g})$ to distinguish open-set samples becomes a good strategy.

However, obtaining the gradient for a single sample is not trivial. It requires passing each sample through the model and computing its gradient, which incurs high time overhead compared to parallelizing over an entire batch of data. To address this, we propose a more efficient computation method based on the model parameter update perspective, leveraging the model's prediction confidence to construct a new open-set score $ods(x)$:

$$\begin{aligned} \theta_t^{'} &= \alpha * \theta_{t-1}^{'} + (1-\alpha) * \theta_t, \\ ods(x) &= \sum_{i=1}^{C} \left(f_{\theta_t}(x)_i - f_{\theta_t^{'}}(x)_i \right) \cdot \mathbb{I}\left(i = c\right), \end{aligned} \tag{3}$$

where $c = \arg\max_{c' \in C} f_{\theta_t}(x)_{c'}$, θ_t is the model parameter updated by full gradient steps at each time t, $\theta_t^{'}$ is the parameter updated via a moving average, and α is the momentum for updating. $ods(x)$ measures the model's confidence change in the maximum class between two consecutive model parameter updates.

Nevertheless, since the source-domain model itself is not perfect, the indicator function in Eq.3 may point to the wrong class thus leading to an unstable ods score. To address this, we propose a new feature-space-based pattern-matching strategy in the next section.

3.3 Pattern Matching Based on Feature Change

Assume that the model consists of a feature extractor f and a classifier h. The individual dimensions of the features $f(x)$ extracted by the model can be viewed as different matching patterns, while the final class prediction $h(f(x)) = [W_1 f(x) + b_1, W_2 f(x) + b_2, \cdots, W_C f(x) + b_C]$ represents the pattern belonging to each class for the extracted features.

However, there exist some hard-to-detect open-set samples, which are mistakenly assigned similar patterns by the model, resulting in their features being difficult to distinguish and leading to misclassification with high confidence. As shown in Figure 4, through Grad-CAM heatmaps[Selvaraju *et al.*, 2017], we observe the following phenomenon: the class activation map of a correctly classified close-set sample mainly focuses on the real features of the target object, whereas the class activation map of an open-set sample focuses on spurious features, which construct patterns similar to those of close-set samples and deceive the classifier h.

Based on this observation, we propose a feature space masking strategy, which covers parts of the original image's pixels, forcing the close-set sample's feature pattern to lose its distinctiveness. For open-set samples, since their feature patterns are inherently spurious, the masking has little impact on them. By analyzing the magnitude of the feature pattern change, we can effectively differentiate between open-set and close-set samples.

$$m_{h,w} = \mathbb{I}(\text{Bernoulli}(1-\rho)), \tag{4}$$

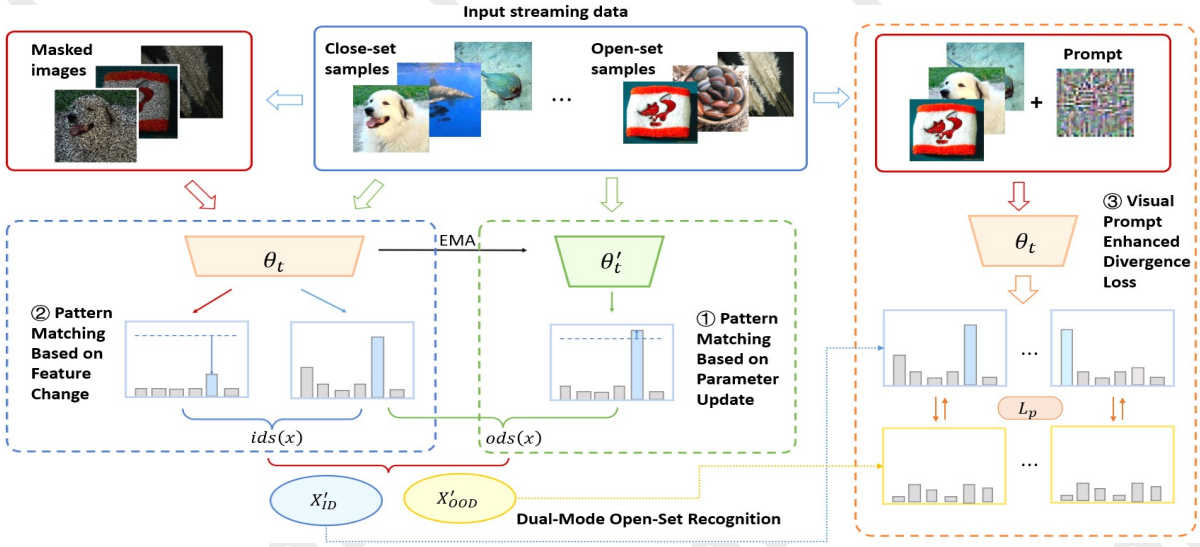
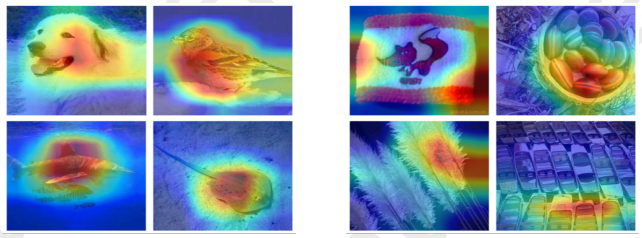


Figure 3: The overall framework of proposed DM-POSA approach. DM-POSA first quickly distinguishes open-set samples through an efficient dual-mode pattern matching based on parameter update and feature change. Then makes better use of them through a visual prompt enhanced divergence loss, which helps to amplify the distribution difference between open-set and closed-set samples, thus inducing robust representation learning and better hard-to-distinguish open-set sample detection.



(a) close-set sample (b) open-set sample

Figure 4: Difference of close-set and open-set class activation maps

where ρ is the probability of masking. We define a new close-set score $ids(x)$ as follows:

$$ids(x) = \sum_{i=1}^{C} \left(f_{\theta_t}(x)_i - f_{\theta_t}(x \odot m)_i\right) \cdot \mathbb{I}\left(i = c\right), \quad (5)$$

where $c = \arg\max_{c' \in C} f_{\theta_t}(x)_{c'}$. Samples with higher ids scores belong to the close-set class, while those with lower ids scores are more likely to belong to the open-set class.

3.4 Dual-Mode Open-Set Recognition

To combine the two previously proposed open-set sample recognition strategies based on mode matching, we use their intersection as the criterion. For any batch of data X_t arriving at time t, the close-set sample set $\mathcal{X}'_{ID}$ and open-set sample set $\mathcal{X}'_{OOD}$ identified are as follows:

$$\begin{aligned} \mathcal{X}'_{ID} &= \{x \mid ids(x) > \text{median}(ids(X_t)) \cap ods(x) \geq 0\} \\ \mathcal{X}'_{OOD} &= \{x \mid ids(x) \leq \text{median}(ids(X_t)) \cap ods(x) < 0\} \end{aligned} \quad (6)$$

To implement TTA, we apply the entropy minimization loss only to the samples identified as close-set:

$$L_{en} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} h(f(x_i))_c \log h(f(x_i))_c, \quad (7)$$

where $x_i \in \mathcal{X}'_{ID}$, $N = |\mathcal{X}'_{ID}|$, and C is the number of close-set classes.

3.5 Visual Prompt Enhanced Divergence Loss

The open-set recognition strategy based on dual-mode matching can identify open-set samples, but it simply discards them and fails to effectively utilize the information contained therein. We further propose a visual prompt enhanced divergence loss to amplify the differences between open-set and closed-set samples, thus inducing robust representation learning and better detecting hard open-set samples.

Suppose the true open-set sample set is $\mathcal{X}_{OOD}$, and the true close-set sample set is $\mathcal{X}_{ID}$. Assume that, at each moment t, the open-set sample set identified in the current batch of data is $\mathcal{X}'_{OOD}$, and the close-set sample set is $\mathcal{X}'_{ID}$. It can be approximated that $\mathcal{X}'_{OOD} \subseteq \mathcal{X}_{OOD}$, $\mathcal{X}'_{ID} \subseteq \mathcal{X}_{ID}$. The class probability distribution of each sample can be expressed as a matrix $P_{\text{ID}} \in \mathbb{R}^{b \times C}$, where the i-th row is the class probability distribution of sample $\mathbf{x}_i$:

$$P = [p_1; p_2; \ldots; p_b], \quad (8)$$

Each $p_i = \left[p^{(i,1)}, p^{(i,2)}, \ldots, p^{(i,C)}\right]$ is the class probability distribution of sample x_i. We measure the distributional divergence between open-set and close-set samples as follows:

$$D(p(x_i), p(x_j)) = \sum_{c=1}^{C} p^{(i,c)} \log \frac{p^{(i,c)}}{p^{(j,c)}}. \quad (9)$$

For all samples, to achieve efficient computation, we design a random pairwise distribution divergence loss. First, let

$$B = \min(|\mathcal{X}'_{ID}|, |\mathcal{X}'_{OOD}|). \quad (10)$$

Assume that $\{u_1, u_2, \ldots, u_B\}$ is the set of predicted close-set samples, and $\{v_1, v_2, \ldots, v_B\}$ is the set of predicted open-set samples. Define a random mapping π:

$$\pi : \{1, 2, \ldots, B\} \to \{1, 2, \ldots, B\} \tag{11}$$

The corresponding random pairing relationship is: $(u_i, v_{\pi(i)})_{i=1}^{B}$. The loss function based on these paired samples can be expressed as:

$$L_p = -\frac{1}{B} \sum_{i=1}^{N} \sum_{j=1}^{M} D(p(x_i), p(x_j)) \tag{12}$$

The advantage of the proposed random-pairing distribution divergence loss is that it is relatively robust to the division of open-set and close-set samples. Even if some close-set samples are misclassified, the probability of pairing them with samples of the same class is small ($p < \frac{1}{C}$), and the resulting noise gradients are also small.

To handle the indistinguishable open-set samples, we incorporate visual prompts and introduce some extra learnable parameters. Firstly, we define the indistinguishable open-set samples as follows: For an open-set sample $x \in \mathcal{X}_{OOD}$, there exists a close-set sample $x' \in \mathcal{X}_{ID}$, such that

$$\begin{aligned} &\mathbf{1}^T |f(x) - f(x')| \leq \delta_1 \\ &\mathbf{1}^T |h(f(x)) - h(f(x'))| \leq \delta_2, \end{aligned} \tag{13}$$

where δ_1 and δ_2 are very small positive numbers. It requires both the representation and label prediction of the open-set sample to be sufficiently similar to a close-set sample.

The proposed method aims to improve the discriminability of indistinguishable open-set samples by using visual prompts. Let ϵ denote the visual prompt, which is a learnable parameter matrix of the same size as the image x. For any image x, applying the visual prompt is equivalent to adding the prompt to x: $x = x + \epsilon$. For indistinguishable open-set samples $x \in \mathcal{X}_{OOD}$, the final optimization goal of the prompt design is that, for any close-set sample $x' \in \mathcal{X}_{ID}$, the following conditions should hold:

$$\begin{aligned} &\mathbf{1}^T |f(x + \epsilon) - f(x' + \epsilon)| > \delta_1 \\ &\mathbf{1}^T |h(f(x + \epsilon)) - h(f(x' + \epsilon))| > \delta_2. \end{aligned} \tag{14}$$

Therefore, when calculating the random pairing discrepancy loss, we will substitute x with $x + \epsilon$ for the calculation. the overall training objective is

$$L = L_{en}(x) + \lambda_1 L_p(x + \epsilon), \tag{15}$$

where λ_1 is a hyperparameter that balances the weight of L_p. It is worth noting that we select the common entropy minimization as the basic loss. However, our method also has the versatility to be extended to other approaches.

4 Experiments

4.1 Experimental Setup

Datasets We select three domain-shifted datasets, Cifar10-C, Cifar100-C, and TinyImageNet-C [Hendrycks and Dietterich, 2019], as the close-set data for testing. These datasets cover 15 different types of perturbations, with each perturbation type having 5 levels of severity, where level 5 corresponds to the most severe disturbance to the original data distribution. The pre-trained models are trained on the close-set categories. In line with UniEnt [Gao *et al.*, 2024], we use SVHN dataset as the open-set data for Cifar10/100-C and the ImageNet-O as the open-set data for TinyImageNet-C.

Evaluation Protocol We follow the single-pass protocol from TTAC [Su *et al.*, 2024]. In this protocol, images with domain shifts arrive in a streaming fashion in batches. At each time step t, the model encounters a mini-batch of test data, which must be immediately predicted and used to update the model parameters. We conduct multiple experiments with varying open-set rates, γ, to evaluate the model's adaptation performance. The open-set rate is defined as the ratio of open-set samples to close-set samples in the test data:

$$\gamma = |\mathcal{X}_{OOD}|/|\mathcal{X}_{ID}|, \tag{16}$$

where $\mathcal{X}_{OOD}$ is the set of open-set samples and $\mathcal{X}_{ID}$ is the set of close-set samples.

Evaluation Metrics Following UniEnt[Gao *et al.*, 2024], we selected three evaluation metrics to assess the performance of our model: ACC, AUROC, and OSCR[Dhamija *et al.*, 2018]. ACC is used to evaluate the classification performance on close-set samples, AUROC measures the model's ability to recognize open-set samples by calculating the area under the receiver operating characteristic curve, and OSCR simultaneously quantifies both the classification accuracy on close-set data and the detection accuracy on open-set data, providing a comprehensive evaluation of the model's performance in open-set classification tasks.

Baseline Methods **Source** refers to using the pre-trained model's predictions without updates. **Pred BN** [Nado *et al.*, 2020] recalculates the BN layer's statistics at the arrival of each batch of test data. **Tent** [Wang *et al.*, 2021] uses entropy minimization loss to encourage the model to output more confident predictions, and updates only the affine parameters of the BN layers to reduce time overhead. **EATA** [Niu *et al.*, 2022] builds on entropy minimization and selects reliable, non-redundant samples for model adaptation. **OWTTT** [Li *et al.*, 2023] dynamically expands prototypes to improve the separation of strong and weak open-set samples. **OSTTA** [Lee *et al.*, 2023] filters out open-set samples during model adaptation based on changes in confidence. **UniEnt** [Gao *et al.*, 2024] employs entropy minimization for close-set samples and entropy maximization for open-set samples. Note that the basic loss functions of all the methods are based on entropy minimization for a fair comparison.

Implementation Details We evaluate the results under the most severe disturbance level (level 5). Following previous research [Lee *et al.*, 2023; Li *et al.*, 2023], we use a 40-layer WideResNet [Zagoruyko and Komodakis, 2016] with an expansion factor of 2 as the source-domain pre-trained model for the Cifar10-C and Cifar100-C experiments, and a pre-trained ResNet-50 [He *et al.*, 2016] for the TinyImageNet-C experiments. During TTA, we only update the affine parameters of the BN layers and the prompt word parameter matrix. We use the Adam optimizer [Kingma and Ba, 2014] with a fixed learning rate of 0.001, a batch size of 256, a balance

Method	$\gamma = 0.5$			$\gamma = 1$			$\gamma = 1.5$		
	ACC	AUROC	OSCR	ACC	AUROC	OSCR	ACC	AUROC	OSCR
Source	81.73	79.39	57.51	81.73	79.47	69.52	81.73	79.42	68.78
Pred BN[Schneider *et al.*, 2020]	84.95	82.57	60.56	83.18	82.13	72.81	81.81	82.09	71.08
Tent[Wang *et al.*, 2021]	85.92	71.34	56.36	83.89	74.46	67.41	82.72	75.75	67.01
EATA[Niu *et al.*, 2022]	86.72	82.46	61.23	85.32	83.00	74.67	84.17	82.29	72.64
OWTTT[Li *et al.*, 2023]	85.87	85.88	62.24	84.32	85.04	75.83	83.64	84.57	74.24
OSTTA[Lee *et al.*, 2023]	85.49	71.50	56.25	83.08	73.20	66.11	81.44	74.45	65.53
UniEnt[Gao *et al.*, 2024]	86.15	86.51	62.59	84.57	85.41	76.08	83.35	84.80	74.06
DM-POSA	**87.47**	**92.29**	**65.29**	**86.49**	**90.16**	**80.74**	**85.54**	**88.01**	**77.96**

Table 1: Results on Cifar10-C Dataset under Different Open Set Rates

factor $\lambda_1 = 0.1$ for the loss function, a feature masking rate $\rho = 0.6$, and an update momentum $\alpha = 0.8$.

4.2 Effect Analysis

Tables 1 to 3 report the performance of the proposed method across three datasets under different open-set rates γ.

On Cifar10-C, the proposed DM-POSA consistently outperforms other methods across all metrics under different open-set rates, especially in terms of the AUROC, where it shows a significant advantage. This demonstrates that the proposed method is more effective at identifying unknown class samples in open-set tasks, offering stronger robustness.

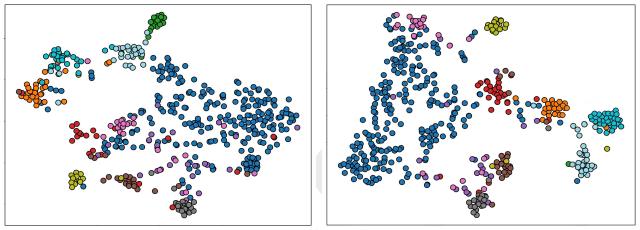
On Cifar100-C, the difficulty of both open-set recognition and close-set classification increases. However, DM-POSA remains the top performer. At open-set rates $\gamma = 0.5$ and $\gamma = 1$, the proposed method significantly outperforms others across all three metrics. At $\gamma = 1.5$, the accuracy of the proposed method is slightly lower than that of OWTTT, but it still leads in terms of AUROC and OSCR. In contrast, methods like OSTTA, although performing well at $\gamma = 0.5$, experience a rapid drop in accuracy as the open-set rate increases, indicating weaker adaptation to unknown class samples.

The results on the TinyImageNet-C dataset further validate the superiority of the proposed method. Especially in terms of AUROC, the proposed method significantly outperforms others, indicating that it is better at distinguishing open-set samples in more complex open-set TTA scenarios. In comparison, other methods show significantly weaker performance in this stage. For example, OWTTT's performance on the TinyImageNet-C dataset is notably lower than on the previous two datasets, showing a performance bottleneck on more challenging tasks.

As the open-set rate γ increases, all methods experience a degree of performance decline. However, the proposed method exhibits strong robustness and adaptability to open-set tasks across all datasets, consistently leading other methods. On the Cifar10-C and Cifar100-C datasets, the proposed method significantly improves classification accuracy and performs excellently in handling open-set samples. The results on the TinyImageNet-C dataset further demonstrate the generalization ability and effectiveness of the proposed method on larger datasets.

4.3 Feature Visualization

We visualize the features using t-SNE [van der Maaten and Hinton, 2008] and compare the visualization results of the final layer representations of the model when performing adaptation on Cifar10-C, between the proposed DM-POSA method and several other open-set adaptation methods. As shown in Figure 5, the feature distinguishment for different classes of other methods is generally lower than that of the proposed method, particularly for class 3. It can be observed that in the other three methods, the representations of class 3 samples are more scattered and are mixed with the representations of open-set samples, making it difficult to distinguish them. In contrast, the representations of class 3 samples extracted by DM-POSA are more clustered and located on the edge of the open-set sample representations. This validates that DM-POSA adapts to a more robust representation space, thereby enhancing open-set sample distinguishment.



(a) OWTTT (b) OSTTA

(c) UniEnt (d) DM-POSA

Figure 5: t-SNE Feature Visualization on Cifar10-C

4.4 Ablation Study

We conduct a comprehensive ablation study on each component of DM-POSA. As shown in Tables 4 and 5, when

Method	$\gamma = 0.5$			$\gamma = 1$			$\gamma = 1.5$		
	ACC	AUROC	OSCR	ACC	AUROC	OSCR	ACC	AUROC	OSCR
Source	53.25	67.91	37.75	53.25	67.80	42.65	53.25	67.83	42.33
Pred BN[Schneider *et al.*, 2020]	58.71	77.60	43.14	55.30	77.54	47.75	52.84	76.76	44.69
Tent[Wang *et al.*, 2021]	61.16	75.32	44.25	57.87	76.36	49.46	55.30	75.70	46.54
EATA[Niu *et al.*, 2022]	61.71	84.31	46.86	59.63	85.85	55.18	57.56	85.24	52.58
OWTTT[Li *et al.*, 2023]	61.84	82.72	46.58	59.76	83.86	54.51	**58.33**	83.12	52.33
OSTTA[Lee *et al.*, 2023]	60.85	76.04	44.27	57.50	77.37	49.68	55.19	76.87	46.99
UniEnt[Gao *et al.*, 2024]	61.25	83.57	46.46	58.51	84.42	53.76	56.41	83.76	51.08
DM-POSA	**62.37**	**87.84**	**47.91**	**60.14**	**87.36**	**56.13**	58.29	**85.71**	**53.35**

Table 2: Results on Cifar100-C Dataset under Different Open Set Rates

Method	$\gamma = 0.5$			$\gamma = 1$			$\gamma = 1.5$		
	ACC	AUROC	OSCR	ACC	AUROC	OSCR	ACC	AUROC	OSCR
Source	21.50	43.15	14.17	21.51	43.23	14.02	21.51	43.21	14.03
Pred BN[Schneider *et al.*, 2020]	34.74	49.64	23.71	32.47	47.03	22.12	30.83	45.80	20.71
Tent[Wang *et al.*, 2021]	37.17	51.22	25.12	34.66	48.29	23.24	32.74	46.42	21.24
EATA[Niu *et al.*, 2022]	36.39	50.36	25.04	34.39	48.31	23.90	32.43	47.13	22.19
OWTTT[Li *et al.*, 2023]	32.31	53.18	22.01	29.56	50.62	20.26	26.38	51.12	17.69
OSTTA[Lee *et al.*, 2023]	37.15	50.32	25.07	34.64	47.65	23.10	32.83	45.72	21.27
UniEnt[Gao *et al.*, 2024]	36.46	55.87	25.83	34.46	53.90	25.09	32.88	53.09	23.56
DM-POSA	**37.76**	**63.68**	**27.15**	**35.64**	**59.56**	**26.51**	**33.86**	**57.22**	**24.20**

Table 3: Results on TinyImageNet-C Dataset under Different Open Set Rates

only entropy minimization is used for updates, the model performs poorly and is unable to effectively distinguish between open-set and close-set samples. After incorporating model parameter update matching (MM), the model's performance improves, enabling it to better adapt to changes in close-set samples. Further introducing feature space variation matching (FM) enhances the model's ability to discriminate open-set samples. Finally, after combining the paired difference loss (PDL) under the prompt word design, the model performs best across all metrics, validating its key role in improving open-set recognition capabilities. At the same time, it can be observed that the improvement in close-set accuracy is relatively small, whereas there is a significant increase in the open-set metric AUROC. Particularly, after adding the paired difference loss, the model's open-set performance improves dramatically, indicating that the paired difference loss based on the prompt word design is a crucial component for enhancing the model's open-set discrimination ability.

MM	FM	PDL	ACC	AUROC	OSCR
-	-	-	57.87	76.36	49.46
✓	-	-	59.18	83.11	53.64
✓	✓	-	59.32	85.58	54.77
✓	✓	✓	**60.14**	**87.16**	**56.13**

Table 4: Ablation Experiment on Cifar100-C (Perturbation Level 5)

MM	FM	PDL	ACC	AUROC	OSCR
-	-	-	83.89	74.46	67.41
✓	-	-	83.14	75.50	67.64
✓	✓	-	83.72	77.63	69.50
✓	✓	✓	**86.49**	**90.16**	**80.74**

Table 5: Ablation Experiment on Cifar10-C (Perturbation Level 5)

5 Conclusion

When the test data contains open-set samples, traditional TTA methods suffer from performance degradation. To timely identify and correctly handle these open-set samples during the testing phase, and prevent them from interfering with the model's adaptation process, we propose the DM-POSA method. It includes an online open-set recognition strategy based on dual-mode matching and a pairwise divergence loss based on visual prompts. The dual-mode matching strategy enables the quick and effective identification of open-set samples upon the arrival of each data batch, while the pairwise divergence loss based on visual prompts further enhances the model's ability to distinguish difficult open-set samples, increasing the distribution gap between known and unknown categories, thus effectively mitigating the negative impact of open-set samples on model performance. Experimental results show that DM-POSA significantly improves performance in open-set test scenarios across multiple TTA datasets. Compared to existing open-set TTA methods, it can more accurately identify and handle open-set samples while maintaining classification accuracy for close-set categories.

Acknowledgments

This work is supported by National Science and Technology Major Project(2022ZD0114801), Natural Science Foundation of China (62472224), Fundamental Research Funds for the Central Universities(NS2024059), Open Project Funds for the Joint Laboratory of Spatial intelligent Perception and Large Model Application (SIPLMA-2024-YB-05).

References

[Boudiaf *et al.*, 2022] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8344–8353, June 2022.

[Chen *et al.*, 2022] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[Dhamija *et al.*, 2018] Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult. Reducing network agnostophobia. In *the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9175–9186, Red Hook, NY, USA, 2018. Curran Associates Inc.

[Döbler *et al.*, 2023] Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7704–7714, 2023.

[Feng *et al.*, 2023] Chun-Mei Feng, Kai Yu, Yong Liu, Salman A. Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2704–2714, 2023.

[Gao *et al.*, 2024] Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23975–23984, 2024.

[Hakim *et al.*, 2024] Gustavo Adolfo Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Clipartt: Adaptation of clip to new domains at test time, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hendrycks and Dietterich, 2019] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 12 2014.

[Lee *et al.*, 2023] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16334–16334, 2023.

[Li *et al.*, 2023] 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11802–11812, Los Alamitos, CA, USA, October 2023. IEEE Computer Society.

[Liu *et al.*, 2021] 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? In *Advances in Neural Information Processing Systems*, volume 34, pages 21808–21820. Curran Associates, Inc, 2021.

[Nado *et al.*, 2020] 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*, abs/2006.10963, 2020.

[Niu *et al.*, 2022] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The Internetional Conference on Machine Learning*, 2022.

[Niu *et al.*, 2023] 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 *Internetional Conference on Learning Representations*, 2023.

[Osowiechi *et al.*, 2024] David Osowiechi, Gustavo A. Vargas Hakim, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Nc-ttt: A noise constrastive approach for test-time training. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6078–6086, 2024.

[Robbins and Monro, 1951] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[Schneider *et al.*, 2020] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551, 2020.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[Su *et al.*, 2024] Yongyi Su, Xun Xu, Tianrui Li, and Kui Jia. Revisiting Realistic Test-Time Training: Sequential Infer-

ence and Adaptation by Anchored Clustering Regularized Self-Training . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(08):5524–5540, August 2024.

[Sun *et al.*, 2020] 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 *the 37th International Conference on Machine Learning*, volume 119, pages 9229–9248, 13–18 Jul 2020.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[Wang *et al.*, 2021] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[Wang *et al.*, 2022] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2022.

[Yuan *et al.*, 2024] Yige Yuan, Bingbing Xu, Liang Hou, Fei Sun, Huawei Shen, and Xueqi Cheng. Tea: Test-time energy adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23901–23911, 2024.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[Zhang *et al.*, 2022] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: test time robustness via adaptation and augmentation. In *the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[Zhang *et al.*, 2023] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *the 40th International Conference on Machine Learning*, volume 202, pages 41647–41676, 23–29 Jul 2023.

[Zhao *et al.*, 2024] Shiji Zhao, Shao-Yuan Li, and Sheng-Jun Huang. Nanoadapt: Mitigating negative transfer in test time adaptation with extremely small batch sizes. In *International Joint Conference on Artificial Intelligence*, 2024.